Analytics-Toolkit.com

Issues with Current Bayesian Approaches to A/B Testing in Conversion Rate Optimization

Georgi Georgiev

Analytics-Toolkit.com

May 18, 2017

ABSTRACT

This paper attempts to cover the more significant issues with current approaches to statistical design and statistical analysis of A/B testing experiments, mostly as applied in fields of Conversion Rate Optimization (CRO) and Landing Page Optimization (LPO).

The paper argues that there are some fundamental issues with applying Bayesian inferential procedures to the case of A/B testing, such as the problem of non-informative priors, the difficulties in interpreting results from Bayesian statistical analyses without proper understanding of the priors involved, as well a general misunderstanding of the effect of stopping rules on the posterior distributions.

It also covers to some extent the approaches of industry leaders VWO and Optimizely, pointing out potential issues based on analyses of the vendor's technical papers. The control of statistical power is examined, as well as the related concept of early termination of experiments due to futility.

This analyses is by no means without limitations, especially when it comes to the particular tools discussed, due to lacking or insufficient details about those

implementations to facilitate an exhaustive evaluation of their performance. This by itself is a non-negligible issue when it comes to Bayesian approaches, as is argued in the paper.

CONTENTS

1. INTRODUCTION

In the field of Conversion Rate Optimization (CRO) practitioners are often interested in testing one or more alternatives to a current text, interface, process, etc. in order to determine which one performs better for a given business objective. Examples of objectives one might consider performing an A/B test for are adding a product to cart, purchasing a product, providing contact details, etc. Practitioners often use empirical data from AB or multivariate tests with actual users of a website or application and employ statistical procedures to control the amount of error in the data to a level they can tolerate with regards to business and practical considerations.

However, for many years the industry as a whole did not have statistical tools that align well with the business case where business owners and executives want to implement the perceived winner or get rid of the perceived loser as quickly as possible. It should be noted that they do so for very good reasons – no one wants to lose users, leads or money.

The above is in stark contrast with the fundamental assumptions of the very basic statistical tests, recommended in many articles and implemented in many A/B testing tools and statistical calculators. Failure to abide to these assumptions quickly leads to highly inflated error rates. In addition to that many practitioners failed to design adequately powered tests since many tools offer poor to no control or measure of statistical power. Finally, there were no rigorous rules for stopping a test early for futility, that is when the chance to detect an effect of a given minimum size with a desired level of certainty. In other words, there was no statistically justified way to stop a test early when the results are highly unpromising.

In an attempt to address some of these issues, two of the leading tool providers launched improved statistical engines that rely on Bayesian approaches to design and analysis of A/B tests.

While these efforts are certainly admirable and they likely led to some improvement in the practice of online experimentation they do appear to be suffering from some significant issues. Below it is argued that part of them are a consequence of the fundamental nature of the Bayesian approach to statistical inference while showing

others are due to specifics in the implementations, including lack of information and lack of control of key design parameters.

2. ISSUES WITH CURRENT BAYESIAN APPROACHES

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

Based on statements by lead statisticians from these vendors 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 patently false, as demonstrated below [2.1], so it cannot be a good justification. The second one is more interesting as Bayesian inference can indeed deliver such answers, but it is neither easy, nor as straightforward as one 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. Most times practitioners start AB tests 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 an interesting time when starting with zero

knowledge as there appears to be 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 "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 hypothesis 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 a practitioner would interpret the resulting posterior probabilities. Since a posterior is an effect of the prior, 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 the 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 experiment at hand and how much it is 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, which can be seen as a significant drawback in itself.

In the sub-chapters to follow several key points are discussed in more detail: 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, the lack of user control on the statistical power of the test or the total lack of such control, and lack of rules for stopping for futility.

The above issues are argued to lead to potential unaccounted errors and/or sub-optimal efficiency in A/B testing, with all the consequences from that. Each point is discussed in significant detail below.

2.1. Choice of Prior Distribution and Interpretation of Results

In Bayesian inference, unlike frequentist inference, one gets as an output the (posterior) probability of a particular hypothesis, given a prior probability and some observed data. Expanding on the description, we can say that the posterior probability is a result of combining prior belief or knowledge, the data generating procedure, model assumptions and the observed data and resulting statistics. These are formally encapsulated into a prior probability and a "likelihood function" derived from a statistical model for the observed data.

The result of applying the Bayes rule is that if the evidence from a particular experiment does not match up with a hypothesis, guidance to reject the hypothesis is produced. However, if the hypothesis was extremely unlikely in the first place (a priori), again guidance to reject it is given, even if the evidence from that particular experiment offers support for the hypothesis. It means that the result from a Bayesian statistic is not just an objective account of the observed data, but a mix of some prior information and the data.

It is exactly in the choice of prior probability distribution that difficulties arise. The choice of a prior is supposed to encapsulate prior knowledge and information about the hypothesis at hand, or in some approaches – subjective beliefs about it.

The issue is that if one's priors are strong enough, even very large quantities of observed data, supporting an opposite conclusion, may not be enough to overcome the prior and reject the null. If the prior information is correct, then this is as expected, but it is not, then it is leading us to wrong decisions. The opposite is also true – one will fail to reject a hypothesis even with minimal data import given a strong enough prior. In both cases, a pertinent question emerges about the resulting posteriors: are they are mainly an effect of the prior speaking, or of the data at hand?

What should then the prior distribution be in the case of an A/B testing setting, is it possible to have an objective enough prior that we will not doubt the conclusions?

In Stucchio (2015) [13], in writing about **VWO's Smart Stats engine** we see the following claim: "It is important to emphasize that there is usually no scientific basis for choosing a prior. The prior is completely subjective…". This raises the question of how a subjective opinion or belief is helpful in measuring the real world of users and transactions, or at least how is it more helpful to use it compared to using frequentist methods which do not mix subjective opinions with the data. Introducing subjective bias in either direction at the input stage might make some test run more quickly, but it is arguably not the proper thing to do if the evidential input of the data and it's error probabilities are of interest.

Further in the same paper, contrary to the claim of subjectivity of priors, there is an attempt to use a so-called objective prior, otherwise known as "non-informative" or "minimally-informative" prior. It is of the family of conjugate Beta priors and it is a uniform Beta(1;1) prior.

Having such a prior is equivalent to adding 2 observations to a test in its beginning – of one success and one failure, that is of one converting user and one non-converting user, or in this case – to one converting user to both the control and the variant tested. Even from the example above one can see that it is not the least influential prior of the Beta family and can in fact be quite informative, especially with small number of observations or with a small proportion of successes (conversions). The fact that this prior is not the least informative one for a Binomial distribution is demonstrated in a very accessible manner in Zhu & Lu (2004) [14].

There are other issues with priors of the conjugate Beta family. Such priors are non-robust in the sense that if the prior and data conflict, the prior has a potentially unbounded influence on the posterior distribution:

"The conjugate binomial and Poisson models are commonly used for estimating proportions or rates. However, it is not well known that the conventional non-informative conjugate priors tend to shrink the posterior quantiles toward the boundary or toward the middle of the parameter space, making them thus appear excessively informative. The shrinkage is always largest when the number of observed events is small. This behavior persists for all sample sizes and exposures. The effect of the prior is therefore most conspicuous and potentially controversial when analyzing rare events."

The above quote was from Kerman (2011) [5].

We read similarly in Fuquene P., Cook, & Pericchi (2009) [2]:

"Even though conjugate analysis is perceived to be simpler computationally […], the price to be paid is high: such analysis is not robust with respect to the prior, i.e. changing the prior may affect the conclusions without bound. Furthermore, conjugate Bayesian analysis is blind with respect to the potential conflict between the prior and the data."

Even though the effect of such a prior diminishes with increasing numbers of users and conversions, it is not rare in practice that tests are called with less than 100 or even less than 50 conversions per variant with sample sizes in the thousands or tens of thousands per arm, and in these cases such a prior could be skewing the posterior in significant way due to the small proportion of conversions.

Given that the reported credible intervals and other statistics in an A/B testing platform are a product of the usage of such priors, it is unclear how a practitioner is to interpret them properly, since that depends on understanding both the observed data and the prior mixed with it.

In reviewing the work of Pekelis et al. (2015) [8] about **Optimizely's "New Statistical Engine"** one finds no attempt to use "objective" or "non-informative" priors - informative priors are used instead: "We chose our prior […] as the result of extensive analysis of historical experiments run on Optimizely's platform.". However, there is

very little actual information about the used prior. In reading the paper one can only conclude that it is of the normal distribution family, with no specifics given.

An end user would be left to wonder: what prior exactly is used in the calculations? Does it concentrate probability mass around a certain point? How informative exactly is it and what weight does it have over the observed data from a particular test? How robust with regards to the data and the resulting posterior is it? Without answers to these and other questions an end user might have a hard time interpreting results from the Optimizely platform in a consistent manner.

No matter how informative the prior, the ground on which historical data is taken as representative of future tests and test outcomes is interesting to examine. Is that data from a controlled subset of tests or from all tests on the platform? How did Pekelis et. al. knew if the experiments used to construct the priors were actually good enough evidence for the observed difference (θ), so as to be taken at face value or were there adjustments made to compensate for bias introduced by user actions and what were they?

Even if one makes the significant assumption of perfect representativeness of the data averages from historical tests, this approach still appears questionable when applied to individual tests, especially if the prior is not robust. Aside from consequences on power and running time (discussed below in 2.3), a non-robust prior means that the analysis will be dogmatic instead of promoting "self-criticism" since prior and sample information are not on equal footing. Arguably, such practice would amount to mixing generic prior information with an unknown relation to the user's particular A/B test statistics, potentially resulting in conclusions contrary to what the data warrants.

A consequence of that might be that the statistical software would be working counter to what the user intent usually is: to know what data X is needed in order to have proper evidence in support of or against a given inference, or ultimately, decision.

Based on the above analysis of the approaches of the vendors VWO and Optimizely it appears that in both cases proper error control and accurate interpretation of results by an end user might be impaired.

The lack of information of the priors and their effect on the posterior has further implications for knowledge sharing and comparisons of results of different tests, ran

on different platforms. Basically, a practitioner who wants to do that will find himself in a situation where it cannot really be done, since a test ran on one platform and ended with a given value of a statistic of interest cannot be compared to another test with the same value of a statistic of interest ran on another platform, due to the different priors involved. This makes sharing of knowledge between practitioners of such platforms significantly more difficult, if not impossible since the priors might not be known to the user.

2.2. Control for the Error, Introduced by Outcome-Based Optional Stopping

In a review of both vendor's software there were no controls provided in the interface for specifying the number, frequency or analyses times for interim analyses. This means that if the vendors are to properly control type I error with optional stopping on the side of the end user, procedures that make assumptions about the number, frequency and timing of interim analyses need to be deployed. These will inherently be sub-optimal, when compared to procedures that do take the actual information into account, instead of assuming it in one form or another.

Before we go into more detail about the actual approaches, a brief overview of the issue of stopping based on interim results might be beneficial.

Armitage et al. (1969) [1] estimates the increase in error probability from outcome-based optional stopping practices and it is measured in the multiples of the nominal error probability. For example, peeking twice using an outcome-based stopping rule results in an actual error probability that is more than twice the nominal one. Peeking 5 times results in ~3.2 times larger than nominal error probability, while if we peaked 10 times it's ~5 times larger.

An example of an often-employed outcome-based stopping rule is checking for an observed naive p-value, confidence interval or an equivalent statistic and basing the decision to stop or continue the test based on that observation. The result from using an outcome-based stopping rule without compensating for it is usually that a winner is declared when there is severely less input from the data at hand, than is perceived.

There are those among the practitioners of Bayesian methods that claim that even when a stopping rule based on the data is employed, it is irrelevant to the resulting Bayes factors. In effect they declare themselves immune from the effect of outcome-based stopping rules.

However, the above is not the case. There are multiple papers [4,6,7,10,12,15] and articles [3,9,11] detailing why not including the stopping rule in an analysis can have just as disastrous effect on the error-rate control (or Bayes factors) for a Bayesian design as they do in a frequentist one. As Uri Simonsohn (2014) [11] puts it while discussing optional stopping "When a researcher p-hacks, she also Bayes-factor-hacks.".

The reason why not accounting for the stopping rule leads to bias and much higher than actual error rates, is that in doing so crucial information is omitted from the statistical model about the procedure that generates the data. This information, if properly taken into account, would shift the sample space significantly and result in very different conclusions, even if the same number of users and conversion rates were observed at the end of a test. Following Bayesian practices does not justify not taking into account a crucial part of the observed data.

Some quotes from papers and articles that support and expand the above explanation are provided below. The reader is, of course, encouraged to read the full papers/posts.

Lindsey (1997) [6] gives examples of designs which result in the same reported likelihood (or nominal likelihood) for the final stopped experiment while in fact they have different underlying likelihood functions. This means that when an outcome-based stopping rule is not taken into account the Bayesian inference suffers from the issue of reporting significantly higher likelihood than the one actually warranted by the data.

This is further supported by Gelman, A., Carlin, J.B et al. (2003) [4] in their highly influential book where they state: "A naïve student of Bayesian inference might claim that because all inference is conditional on the observed data, it makes no difference how those data were collected, […] the essential flaw in the argument is that a complete definition of "the observed data" should include information on how the observed values arose […]".

Prominent Bayesian statistician Prof. Andrew Gelman explains how and when the stopping rule should be accounted for in Gelman (2014) [3]: "the stopping rule enters Bayesian data analysis in two places: inference and model checking:

1. For inference, the key is that the stopping rule is only ignorable if time is included in the model. To put it another way, treatment effects (or whatever it is that you are measuring) can vary over time, and that possibility should be allowed for in your model, if you're using a data-dependent stopping rule. To put it yet another way, if you use a data-dependent stopping rule and do not allow for possible time trends in your outcome, then your analysis will not be robust to failures with that assumption.

2. For model checking, the key is that if you're comparing observed data to hypothetical replications under the model (for example, using a p-value), these hypothetical replications depend on the design of your data collection. If you use a data-dependent stopping rule, this should be included in your data model, otherwise your p-value isn't what it claims to be."

Data scientist David Robinson puts it a bit differently [9], noting the differences in what Bayesian inference offers as compared to the frequentist error-control type of inference we are most used to:

"Bayesian methods do not claim to control type I error rate. They instead set a goal about the expected loss. In a sense, this means we haven't solved the problem of peeking described in "How Not to Run an A/B Test", we've just decided to stop keeping score!" and also: "Bayesian reasoning depends on your priors being appropriate." and the choice of a prior is a point of contention among statisticians of all stripes and colors.

Mayo and Kruse (2001) [7] from the philosophy of science point of view are even more critical to the Bayesian approach to inference as a whole and to optional stopping and its relation to the likelihood principle in particular: "Embracing the LP is at odds with the goal of distinguishing the import of data on grounds of the error statistical characteristics of the procedure that generated them.". They take their argument to great lengths so it is a recommended read for anyone interested in foundational epistemological issues.

Let us again review the vendor's technical papers to see what can be learned about their specific approaches. In Pekelis et al. (2015) [8], writing about **Optimizely's "New Statistical Engine"** we see that there is an attempt to correct for optional stopping, but in what is arguably, at best, a sub-optimal way. What Optimizely appears to be doing is it treats each time a test is loaded in the interface by a user as a "look" or "peek" and then it adjusts the stats accordingly using the False Discovery Rate control method, proposed by Benjamini and Hochberg and improved by Benjamini and Yekutieli.

While this statistical procedure has very interesting applications and has in fact been popularized for use in A/B testing by the author of this paper, applying it for optional stopping does not seem to be warranted. Optional stopping, if viewed from the perspective Optimizely approaches it, would be a case for controlling Family-Wise Error Rates (FWER), not False Discovery Rates (FDR).

FDR control is less stringent than FWER control, since FDR controls the proportion of type I errors, while FWER controls the probability of at least 1 type I error. As Pekelis et al. themselves state about FDR: "It is defined as the expected proportion of false detections among all **detections** made" (emphasis mine). This makes FDR more powerful, however that comes at the cost of a less stringent type I error control compared to FWER. Since in A/B testing one is interested in the probability of erroneously stopping the test after any given data observation, control that is relevant to making a single error is required. FDR is not able to provide such control, since by design it offers no such guarantees.

In fact, as the number of looks or peeks increases, so does the number of errors permitted by the BH-FDR or BHY-FDR procedures and as consequence – the probability that we would stop when we should not. The more stringent control of FWER is required, instead.

The performance of the approach would be sub-optimal in terms of the sample size required to test at a given significance level and power, even if proper FWER control was in place. That happens as consequence of the fact that Optimizely apparently treats each view of a test in their interface as data observation with the intention to stop the test, if a threshold was crossed: "The act of viewing an A/B test must mean that it is a candidate in an experimenter's selection procedure."

A seasoned practitioner can easily think of at least several common cases where that will not be the true: one or more tabs are loaded when a browser is started or restarted; one or more tabs remain active for extended periods of time (in the background)l; the user is looking at the data, but just to make sure the test is technically OK (users are gathering at the expected rate, users are assigned to all variations, etc.); the user is looking at the data to compile a report, but not necessarily with intent to react to the data… Due to this, the optional stopping adjustment will be less powerful than optimal, requiring longer test times to preserve all other design attributes.

In Stucchio (2015) [13] when writing about **VWO's Smart Stats engine** it seems that there is no attempt to adjust error probabilities accounting for optional stopping at all, hence the approach as described there appears to be subject to issues of underestimating the uncertainty in the reported "credible intervals". The paper doesn't really cover that topic, so assessments of VWO's approach was difficult and external sources were sought.

In an online discussion under Robinson (2015) [9] Stucchio admits verbatim that optional stopping inflates the error rates, but adds that the loss function is adjusted for optional stopping in an unspecified manner:

"Also, the point is well noted that peeking does affect the error rates, though the loss does remain below the threshold."

However, even though it is also mentioned that "If there is a cost to switching, then the ideal way to handle this is to build it directly into the loss function." it remains unclear if and how a user has an input on that loss function calculation.

Furthermore, the loss function as described in both VWO's paper and the comment cited above appears to rely on currently observed values, treating them as given. If one takes the currently observed CTR, CR or whatever metric as given true values, this is sure to introduce significant bias in the loss function calculations, as it is equivalent to optional stopping.

Examining the paper reveals no adjustments or corrections described as relating to these values or the loss function as a whole. Given the non-informative prior, it is also clear that optional stopping is not accounted for in the prior.

In conclusion, it appears that both vendor approaches above fail to offer satisfactory error control in the case of optional stopping, which inevitably results in an unknown but potentially significant disconnect between the actual and reported statistics.

2.3. Lack of Control for Statistical Power

A major drawback of many statistical guidelines for A/B testing has been the lack of consideration of statistical power in the design of the experiment. Statistical power, also called "sensitivity" of a test, is the probability that the experiment will detect a genuine effect of a given minimal value at a given minimum statistical significance threshold. Underpowered tests lead to missed opportunities to improve, or worse – to unwarranted conclusions that the baseline is better than the control, while overpowered tests lead to wasted time and users, resulting in slower implementation of winners and overall slower A/B testing cycles.

Here it is examined if and how statistical power is accounted for in the same two vendors' approaches and what kind of control does an end user have over the power of the tests they perform.

Pekelis et al. (2015) [8] mention that they try to maximize power based on historical data, however, the user is seemingly left with no control over it, which is a significant issue. Historical data might be a good average predictor, but might fail miserably in any particular experiment and with no option to specify details, related to power calculation the end user usually has no way around that.

The procedure that tries to maximize power in the Optimizely engine relies on subjective, undisclosed priors: "The prior lets us focus on the effect sizes we most anticipate, and the test is optimized for fast detection on θ where the prior is large, i.e., for those effects we are likely to observe. […] We chose our prior […] as the result of extensive analysis of historical experiments run on Optimizely's platform. It should be noted that without this personalization, sequential testing did not give results quickly enough to be a viable for use in an industry platform."

To illustrate why that is an issue for the efficiency of a test, let us review a simple example. For the purpose of illustration, let us assume that the chosen priors

concentrate probability mass around θ = 1%. This sounds like a reasonable average lift for experiments that test low-impact changes far up the funnel and measure against things like number of purchases. However, if a given practitioner tests disruptive changes to the checkout flow that can easily result in double digit lift versus the control, said practitioner will have to gather significantly more data before the evidential input overturns the bad prior, resulting in a very inefficient test due to the fact that the prior maximizes power for a difference of 1%.

If the practitioner is instead able to choose what kind of minimum effect and what power ("sensitivity") level makes business sense to them, the test would have been more efficient, even for small differences and regardless of the outcome, given that a proper stopping rule is employed, as is argued further in the paper.

On the other side, we have the Stucchio (2015) [13] paper, which does not even contain the word "power" or "type II error" and one is left to wonder how the question of test sensitivity and thus the issue of falsely declaring that there is no evidence for a discrepancy is addressed in the VWO testing framework.

Neither software is known to provide interface control for setting or adjusting the power level of their experiments, at the moment of writing.

The practical effect of both approaches is that users are likely to find themselves running quick, underpowered tests that result in no winners when in fact there would have been a discovery (with the specified effect size and certainty) if the test was just given a fair chance. Not giving the test variations a chance translates into wasted resources for CRO planning and deployment and missed opportunities to capitalize on real improvements.

Sometimes significantly overpowered tests will be conducted as well, such as in the illustrative case provided regarding the approach in Pekelis et. al. – if the test is concluded successfully, resulting in wasted opportunity to implement a winner as early as possible, as well as logistical and other resources for running the test.

2.4. Lack of Rules for Early Stopping for Futility

The lack of rules for stopping tests for futility – "early" stopping when the test is highly unlikely to yield a positive result, as defined in the test design, can be considered as another possible drawback in existing solutions, not limited to the two vendors previously discussed. There is nothing in the two technical papers reviewed that was identified as describing procedures or stopping rules from the point of futility.

Having a rule for early stopping for futility means that a test can be terminated before its maximum scheduled time, if results in any given point fall below a minimum threshold. That is, if early results are negative enough that continuing the tests is unlikely to result in a positive discovery, the test can be terminated with high certainty that the baseline is better than the tested variant. "High" here depends on the level and standards chosen for statistical power.

In many cases, especially in later-stage optimization efforts when the baseline variant has been through many optimizations and refinements, lack of a futility stopping rule can lead to unnecessarily prolonged tests where gains of a given minimum detectable effect size are less likely than a chosen threshold of the power of the test at the specified significance threshold.

Failure to stop for futility is likely to result in revenue losses as well as resources and time spent on clearly inferior tests, instead of being redirected to testing other, more promising hypotheses.

3. LIMITATIONS

The current paper only discusses two of the most prominent and popular A/B testing statistical solutions based on Bayesian inference. Other solutions not covered in the paper might have better or, in fact, worse approaches to the issues discussed or may suffer from issues not discussed in this paper.

The two solutions are analyzed based on publicly available information from the vendors and their representatives. This should make the analysis more applicable to the average A/B testing practitioner, but it also makes it incomplete, since details about some of the inner-workings of the statistical engines are not readily available and cannot be deduced.

The latter is, arguably, both a limitation and an argument against the adoption of Bayesian solutions where information about the priors is not available to the practitioner.

4. DISCUSSION

In this paper, we discussed some persistent issues in AB testing for CRO when Bayesian statistical analysis is employed. Fundamental issues and controversies, such as the ones surrounding the choice of prior, the interpretation of results with none or minimal information about the prior, and consideration for outcome-based stopping rules were presented and their implications on the ability to properly conduct and analyze an A/B test were examined. Vendor-specific approaches of leading software providers with relation to the control of statistical power, outcome-based stopping rules and futility stopping were critically examined.

The fundamental issues present in Bayesian inference when applied to situations of no prior knowledge were found to be important, especially in cases of low sample sizes or low proportions of the tested effect, while the choice of prior remains controversial both in theory and in practice, as evidenced by the completely opposite approaches the two vendors discussed took to constructing their priors.

The lack of information about the prior in use was identified as a potential issue for software where the priors are not disclosed, as it inhibits proper understanding of the reported posterior distributions and resulting statistics.

Consideration for the stopping rule is another case where implementations and approaches differ drastically between the leading vendors. Practitioners should make themselves familiar with the particular solution in order to avoid a discrepancy between nominal and actual error rates.

Control of statistical power is where both reviewed vendors scored lowest, not least by offering little to no control to the end users. A related concept - futility stopping, does not seem to be employed in any way, potentially limiting the efficiency of the tests designed.

Some of the issues described above can be overcome with better practical implementations. For example, control over the power of the test can be provided to end users and rules for stopping for futility can be introduced, which should result in both better statistical design and more efficient tests. A fair amount of education may need to accompany these changes so that users are aware of what these new options give them control over.

Taking into consideration the effect of stopping rules on the resulting statistics can also be baked into existing solutions or new Bayesian solutions, although how this would be done precisely will determine its effectiveness in producing better control for type I and type II errors.

However, other issues appear to be harder to solve and will probably need a resolution within the broader Bayesian practitioner base, before these make it into the A/B testing arena. The choice of prior and the communication of the characteristics of the chosen prior will likely remain a debated topic, given that it is an issue right at the heart of Bayesian inference as a whole. This is tied in with the interpretation of results, however it is not certain that this issue has a proper solution, even if A/B testing platforms open up their priors to the public, as the general practitioner will still require a lot of effort before he or she can compare the results from one platform to another.

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] Fuquene P., J.A., Cook, J.D., Pericchi, L.R. (2009) "A Case for Robust Bayesian Priors With Applications to Binary Clinical Trials", *Bayesian Analysis* Volume 4, Number 4, 817-846.

[3] Gelman, A. (2014) "Stopping Rules and Bayesian Analysis" http://andrewgelman.com/2014/02/13/stopping-rules-bayesian-analysis/

[4] Gelman, A., Carlin, J.B, Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2003) "Bayesian Data Analysis" (2nd edition), p.203

[5] Kerman, J. (2011) "Neutral Noninformative and Informative Conjugate Beta and Gamma Prior Distributions", *Electronic Journal of Statistics*, 5:1450-1470

[6] Lindsey, J.K. (1997) "Stopping rules and the likelihood function", *Journal of Statistical Planning and Inference* 59:167-177

[7] Mayo, D.G., Kruse M. (2001) "Principles of Inference and Their Consequences", *Foundations of Bayesianism* (vol. 24 of the Applied Logic Series) pp 381-403

[8] Pekelis, L., Walsh, D., Johari, R. (2015) "The New Stats Engine" (Optimizely)

[9] Robinson, D. (2015) "Is Bayesian A/B Testing Immune to Peeking? Not Exactly" http://varianceexplained.org/r/bayesian-ab-testing/

[10] Sanborn, A.N., Hills, T.T. (2014) "The frequentist implications of optional stopping on Bayesian hypothesis tests", *Psychonomic Bulletin & Review* 21 Issue 2, pp 283-300

[11] Simonsohn, U. (2014) "Posterior Hacking" http://datacolada.org/13

[12] Steel, D. (2003) "A Bayesian Way to Make Stopping Rules Matter", *Erkenntnis* 58:213-227

[13] Stucchio, C. (2015) "Bayesian A/B Testing at VWO" https://cdn2.hubspot.net/hubfs/310840/VWO_SmartStats_technical_whitepaper.pdf

[14] Zhu, M., Lu, Arthur (2004) “The Counter-Intuitive Non-Informative Prior for the Bernoulli Family”, *Journal of Statistics Education* Vol 12, Issue 2

[15] Yu, E.C., Sprenger A.M., Thomas, R.P., and Dougherty, M.R. (2014) “When Decision Heuristics and Science Collide”, *Psychonomic Bulletin & Review* 21 Issue 2, p268