Analytics-Toolkit.com

Non-Inferiority Designs in A/B Testing

Georgi Z. Georgiev

Analytics-Toolkit.com

Sep 12, 2017

ABSTRACT

Most, if not all the current statistical literature on online randomized controlled experiments, (commonly referred to as "A/B Tests"), focuses on superiority designs. That is, the error of the first kind is formulated as incorrectly rejecting a composite null hypothesis of the treatment having no effect or having a negative effect. It is then controlled via a statistical significance threshold or confidence intervals, or posterior probabilities, and credible intervals in Bayesian approaches.

However, there is no reason to limit all A/B testing practice to tests for superiority. The current paper argues that there are many cases where testing for non-inferiority is both more appropriate and more powerful in the statistical sense, resulting in better decision-making, and in some cases: significantly faster tests. Non-inferiority tests are appropriate when one cares about the treatment being at least as good as the current solution, with "as good as" being defined by a specified noninferiority margin (sometimes referred to as "equivalence margin"). Certain non-inferiority designs can result in faster testing compared to a similar superiority test.

The paper introduces two separate approaches for designing non-inferiority A/B tests: tests planned for a true difference of zero or more and tests planned for a

positive true difference. It provides several examples of applying both approaches to cases from conversion rate optimization. Sample size calculations are provided for both approaches and a comparison is made between them and between non-inferiority and superiority tests.

Finally, drawbacks specific to non-inferiority tests are discussed, with guidance on how to limit or control them in practice.

TABLE OF CONTENTS

1. ERRORS OF THE FIRST KIND AND TESTING FOR NON-INFERIORITY

Practitioners in the field of Conversion Rate Optimization (CRO), Landing Page Optimization (LRO) and User Experience (UX) Testing are usually interested in determining whether a proposed alternative or alternatives to a current text, interface, or process, performs better than the existing one. After all, most of the time the job of such a specialist is to design and deploy an improved variant that leads to increased purchases, leads, or engagement metrics, ultimately improving the business bottom-line.

Establishing the uncertainty associated to measurements of the achieved improvement is instrumental both in making sure the overall return on investment (ROI) is positive, and in proving the value of the work performed. In many cases the primary error one can commit is to reject a current solution that is just as good as, or better than the new solutions tested. This is the case when there is an ongoing cost in maintaining the new solution, when it is hard or impossible to reverse it, or when re-testing at a future date is hard.

In these cases, the null hypothesis is chosen such that the error of the first kind – type I error, reflects that primary concern. The null is usually a composite hypothesis covering the possibility that the current solution (control) is better or equal to the proposed improved variant(s). Statistical significance tests or confidence intervals are used to control the error.

The possibility that a truly better variant will fail to pass the statistical significance threshold is deemed a secondary concern and is controlled by the statistical power of the test, often called test sensitivity. With high enough power, there is good certainty that a true improvement of a given magnitude will be detected with a desired statistical significance.

However, there are many cases when the above is simply not true and it no longer makes sense to define the null hypothesis in the same way, since “Every experiment may be said to exist only in order to give the facts a chance of disproving the null

hypothesis." [1] Of course, in a Neyman-Pearson approach where there are also type II errors, "only" needs to be replaced with "primarily".

In null hypothesis statistical tests, a statistical null hypothesis is defined such that the error of primary concern is rejecting it incorrectly. The Type I Error is thus defined by the choice of null and controlled by statistical significance. This means that the Type I Error doesn't need to measure the likelihood that the tested variant is better than the control, it can measure the likelihood that the tested variant is not doing worse than, or being inferior, to a certain margin.

In non-inferiority tests the primary concern is that an existing solution which is performing worse than or about equal to a new one for a given Key Performance Indicator (KPI) will remain in place, despite the new solution having advantages over the existing one.

2. THE NEED FOR NON-INFERIORITY DESIGNS IN A/B TESTING

There are two major cases where non-inferiority testing can prove invaluable, or at least without an alternative.

Firstly, a variant may be tested which is easy to implement, costs nothing to maintain, is easily reversible and re-testing it is cheap. Examples include trivial changes such as color or text changes on a Call to Action (CTA) element, many copy or image changes, the removal or addition of some elements of the site or third-party functionality such as trust signals, live chat, etc. I call these "easy decision" cases.

In all the easy decision cases, the primary concern is no longer that one will implement a new variant that is about equal or worse to the existing one. Instead, it is that one may fail to register many easy to implement variants that may lead to small improvements, adding up to a significant overall lift. Or, the concern can be that these improvements will be registered too late due to the large sample sizes required to reliably detect small lifts. This latest concern is alleviated by using sequential testing approaches, such as the AGILE statistical method proposed by me recently, but it is still not an ideal in the above cases.

Secondly, the test may involve a new solution that has benefits not measurable in the test. I call these "side benefits" cases, since they involve considerations outside the primary KPI a test is being measured against.

A solution having such benefits which performs equivalently to the existing solution, or even slightly worse, could still be the preferred solution, for example due to lower maintenance costs, or better brand integrity, etc.

Some concrete A/B testing examples: removing 360-degree shots of products can result in significant savings for an online merchant, and they might even tolerate a bit lower conversion rate; removing a free trial period that requires one or two additional customer support personnel can be great if the conversion rate to a paid account remains about the same; removing several payment methods may significantly simplify payment processing and invoicing, so if it only affects conversions a little bit, it might well be worth doing it.

In both types of cases above the error of primary concern is no longer that the implemented new variant is worse than or about equal to the current solution, but that one may fail to implement a variant that is about equal or slightly better (in the "easy decision" cases) or about equal or slightly worse (in the "side benefits" cases).

Thus, it makes no sense to design superiority tests in such cases. Non-inferiority designs are the natural consequence of the question asked and the error of greatest concern the practitioner needs to ward against.

3. HOW NON-INFERIORITY TESTING WORKS

As with any other null hypothesis statistical test, the first step is defining the statistical design parameters and estimating a sample size. Unlike superiority tests, a decision about a noninferiority margin should be made. A noninferiority margin is such a negative difference, or relative difference, that can be tolerated while still considering the performance of the control and variant about the same. Similar to minimum effect of interest, the smaller the margin, the larger the sample size required in order to get the same statistical significance and have the same power.

The noninferiority margin should be chosen in such a way, that even if the proposed new solution is worse than the tested control by that margin, one would still be happy to implement the new solution. I elaborate on strategies for choosing a noninferiority margin and powering test in point 4.

In a fixed sample design, an A/B or A/B/n test is evaluated once it is over by using a one-sided confidence interval or, equivalently, a two-sided confidence interval for half of the desired type I error [2,4]. If the confidence interval contains the noninferiority margin, one can't reject the null hypothesis. If the lower bound of the interval is higher than the noninferiority margin, one rejects the null hypothesis of inferior performance.

The same can be done by using other statistics like z-value and the corresponding p-value.

Figure 1 below illustrates with confidence intervals successful (statistically significant) tests of three types: superiority, non-inferiority testing and equivalence. The last one is added to help avoid the common mistake of thinking about non-inferiority tests as equivalence tests.

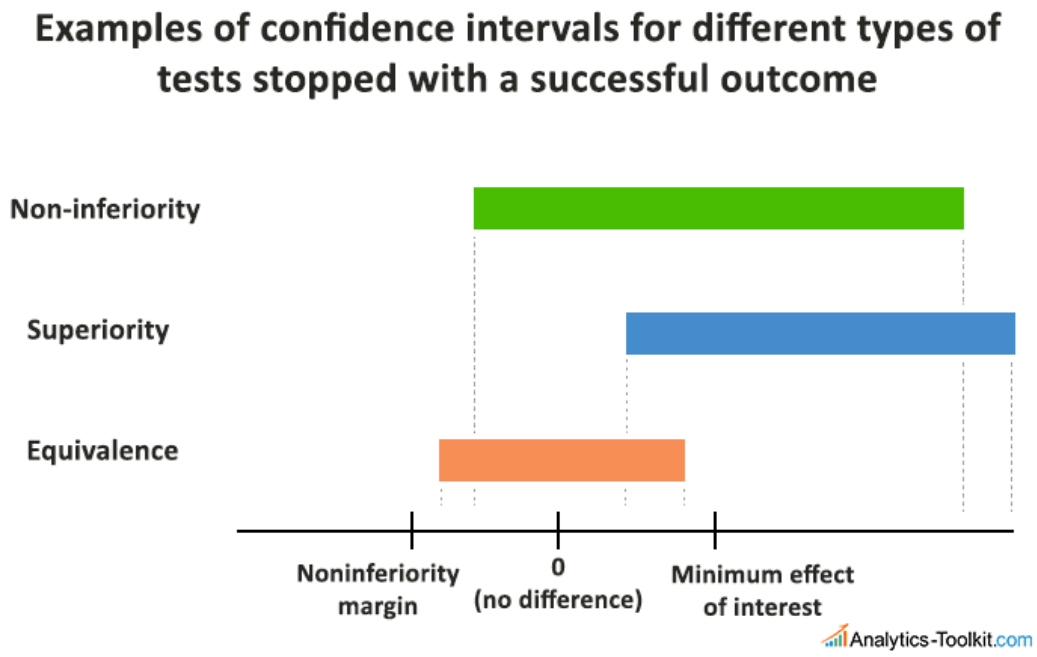

***Figure 1**: Examples of confidence intervals for different types of tests stopped with a successful outcome. The minimum effect of interest is equal in magnitude to the noninferiority margin, but this need not be the case.*

In fact, as seen on the graph, equivalence tests are a subset of non-inferiority tests. Similarly, but no in the same way, superiority tests are also a subset of non-inferiority tests. One can think about superiority tests as stricter non-inferiority tests, while equivalence tests are even stricter (in a different way) than superiority tests. The stricter the test, the more data it requires (the less wide the confidence interval, the more data it requires), so there is already a hint for where the performance gains of non-inferiority testing come from.

4. TWO TYPES OF NON-INFERIORITY A/B TEST DESIGNS

I propose two distinct approaches to choosing a noninferiority margin, depending on the type of test one is planning.

4.1. Testing for side benefits

First, when dealing with a case of "side benefits" testing, the test should be planned so even if the true difference is zero, it will be able to reject the null hypothesis that the difference is worse than the noninferiority margin. The sample size needs to be calculated assuming the two proportions will be equal and thus the noninferiority margin acts as a minimum effect of interest. By projecting the positive effect of the side benefits for several years ahead, by knowing how hard it would be reverse the decision, and other important considerations, the intersection between greater certainty and faster testing can be found.

The sample size required in this case would be almost equivalent to running a superiority test with the same parameters and a minimum effect of interest equal in magnitude to the noninferiority margin. The benefit from using a non-inferiority design instead of a superiority one, is that the A/B test can be a success even if the control and variant are equivalent, or about equivalent. If it was a superiority test,

and there is barely any difference between control and variant, the tested variant should be abandoned and the control would stay in place.

In cases of fixed-sample tests, even if the test was designed as a superiority test, one can still analyze the data as a non-inferiority test after the test is completed. However, there should be care to not introduce bias in the choice of noninferiority margin, as it can heavily skew the test result. Switching the design once a test has started is highly discouraged in sequential tests, where the test design has direct impact on the stopping probabilities at different stages of the test.

4.2. Testing easy decisions

Second, when dealing with "easy decision" type of A/B tests, one is hoping for an improvement, but due to the nature of the test is willing to implement a roughly equivalent variant. In this case the minimum effect of interest can be used in a somewhat unorthodox way. Instead of powering a test based on a worst case of no difference in the proportions, it can be powered based on the expected minimum effect of interest, effectively designing a test that has a desired power to reject an inferior null, assuming the variant performance is better than the control performance by a margin set by the minimum effect of interest.

By powering the test for a worst case of a difference with the magnitude of the minimum effect of interest a vast reduction of the required sample size is achieved (Figure 3), without compromising the ability to inform a decision. The trade-off involved compared to the case of "side benefits" tests is that there is lower power to reject the null if our version is about equal to the control. However, having high power to reject the null in such a case is not something a practitioner should be interested in with an "easy decision" situation at hand.

5. EXAMPLES AND SAMPLE SIZE CALCULATIONS

5.1. Side benefits example and sample size calculations

There is an online merchant who is currently doing 360-degree shots of all his products, probably, but not necessarily based on prior tests showing that this

improves conversion rate by some extent. The merchant is considering stopping adding new 360-degree shots due to the technical and personnel costs involved in making, uploading and hosting them. Still, it would only make sense to do that if this doesn't result in a decline of the e-commerce conversion rate of over 4%, relative to the baseline. This is a clear case for a non-inferiority design, since one would be perfectly happy if the new variant performs about equal to the control, with a 4% noninferiority margin.

Assuming a baseline user-based conversion rate of 1.5%, and about 170,000 unique users per month, one needs to design an acceptable non-inferiority test. Figure 2 contains several sample size calculations [3] that can be used to guide the decision about the test plan.

Significance Level	Statistical Power	Users required	Number of weeks Required
99%	90%	1,068,474	27
98%	90%	913,114	23
95%	90%	702,950	18
90%	90%	539,247	14
99%	80%	823,791	21
98%	80%	688,119	18
95%	80%	507,485	13
90%	80%	370,021	10
85%	80%	289,516	8

***Figure 2**: Sample size requirements for a side benefits non-inferiority A/B test.*

If looking for maximum speed, one can go with 80% or 90% statistical significance and 80% power and be done in 2 and a half months. If waiting more for higher uncertainty is preferred, options that require up to half a year are available.

Comparing the numbers above to a superiority test with a minimum effect of interest of the same magnitude, we see the numbers are practically the same. For 95% significance and 80% power, we get 517,474, which is about 10,000 worse. For 95% significance and 90% power we get 716,786 which is about 14,000 worse.

5.2. Easy decision example and sample size calculations

There is an online SaaS website for which we want to test a simple change of a button text. Currently, the button says "Free Trial" and the test is whether adding an action-inducing word to it will change things, so the variant tested is "Start Free Trial". The current free trial conversion rate is 9% and one would be able to reliably detect an improvement of 5% or more. It is also acceptable to change the text quicker, even if it means it might be equivalent to, or up to 2% worse than, the current text.

Figure 3 shows a comparison between the sample size required by a non-inferiority design with both a minimum effect of interest and an noninferiority margin, compared to a classic non-inferiority design as the one used in 5.1. Since a classic non-inferiority design is about equivalent to a superiority design with the same margin in the positive direction, the table can be used as a comparison to a superiority design as well.

Significance Level	Statistical Power	Users required (easy decision)	Users required (classic)
99%	90%	62,488	748,474
98%	90%	53,402	639,644
95%	90%	41,111	492,422
90%	90%	31,537	377,747
99%	80%	48,179	577,072
98%	80%	40,244	482,033
95%	80%	29,680	355,498
90%	80%	21,641	259,203
85%	80%	16,932	202,808

***Figure 3**: Comparison between the sample size requirements of a classic non-inferiority design versus a design with a minimum effect of interest.*

It easy to see that the easy decision design requires only 8.35% of what a classic non-inferiority design would require. Of course, this number depends heavily on both the noninferiority margin and the minimum effect of interest chosen. Changing the minimum effect of interest from 5% to 2% means an easy decision design will now require 22.23% of a classical non-inferiority design, and about the same proportion of a superiority design with the same minimum effect of interest.

6. RISKS AND DRAWBACKS

Non-inferiority tests have significant benefits in many situations, but they come with two drawbacks that are of no concern when doing superiority tests only.

6.1. Cascading Loses

The cascading loses issue refers to loses that can accumulate due to the noninferiority margin allowed. In a worst-case scenario, one can do everything properly and still end up accumulating loses over the course of several tests.

To illustrate: say one runs 5 consecutive tests, each with a non-inferiority margin of 2%. If each A/B test ends up with a winning variant that is 2% worse than the control, one will end up with a -10% lift.

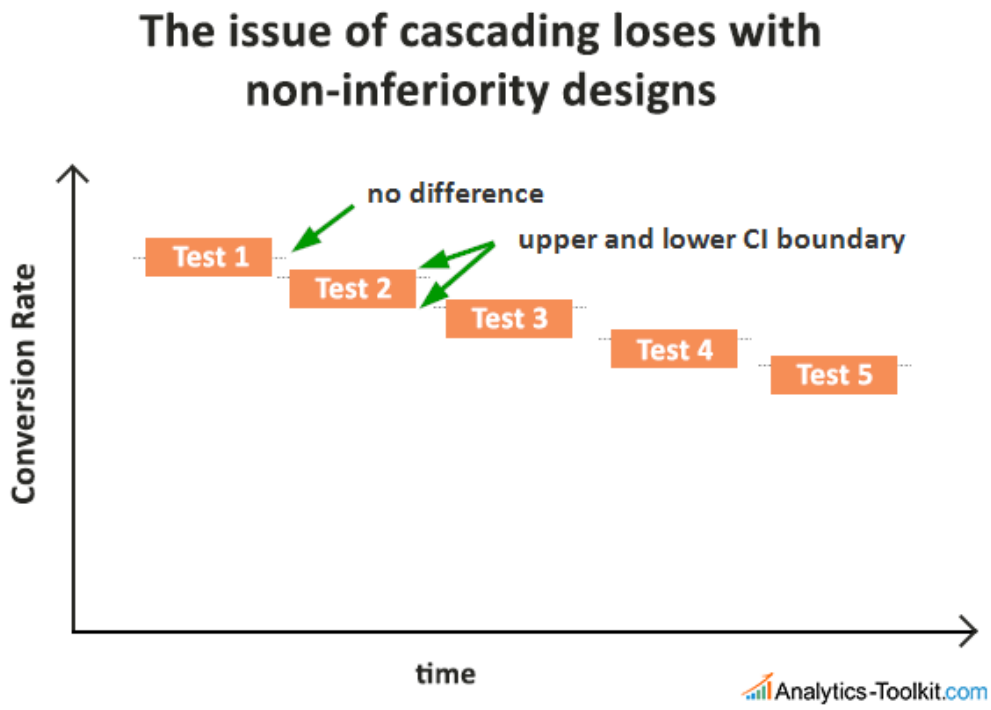

***Figure 4**: Cascading (accumulating) loses as a worst-case scenario of non-inferiority A/B tests.*

The risk is real and grows bigger with the number of A/B tests performed where the observed confidence interval includes negative difference values. How acceptable it is will vary from one circumstance to another.

The risk of accumulating loses can be alleviated in several ways. First, by using best practices and prior A/B testing experience to guide the development of new tests, a practitioner is less likely to test inferior variants. Second, by doing quality user research when designing tests, a practitioner will be more likely to come up with improvements, rather than variants that will perform about equally to the control. Automated testing and "testing for the sake of testing" are best avoided as well.

A way to detect and in fact control the presence of cascading loses is to periodically run superiority or non-inferiority A/B tests where the control is a version of the element/page/process from a few tests ago, while the tested variant is the winner of the latest test on the same element/page or process. Such tests can combine the outcomes of many A/B tests affecting different parts of a website or ad campaign, in both variant and control, though careful consideration for risks of running into the Simpson's paradox and other segmenting issues is advised when doing so.

6.2. Demonstrating Improvement

Another drawback, specific to the "easy decision" type of non-inferiority tests is that proving the improvements achieved through A/B testing might be harder when such tests end up with confidence intervals that cover "no difference". Justifying the continuing optimization efforts might be a challenge in such a case.

The same approach as the one used to control cascading loses can be employed to demonstrate the improvement achieved by a series of non-inferiority A/B tests. Tests where an old version is compared to the latest winner from a series of tests can be used to demonstrate the cumulative improvement for any given set of tests.

While not without challenges and costs, running such cumulative tests for a series of "easy decision" tests should require low effort in practice, due to the nature of the things tested. In addition, since non-inferiority tests are often much faster than superiority tests, running 5 non-inferiority tests and a cumulative superiority test to estimate the total efficiency can still be significantly better than doing 5 superiority tests. Of course, the above assumes one is using non-inferiority tests where they are suitable.

7. DISCUSSION

There are cases where a non-inferiority statistical design is the appropriate approach due to our concern with variables outside of the outcome variable measured in a test. In such cases, while a non-inferiority design is not more efficient than a superiority design in terms of testing speed, it should still be preferred, as it allows data-driven decisions to be made even when a variant is equal, to a given margin, to the tested control, unlike a superiority design.

In the case of A/B tests of the easy decision type, non-inferiority designs are somewhat trickier, as on top of deciding on a noninferiority margin, one needs to decide on a minimum effect of interest. Running a non-inferiority design instead of a superiority one also means the conclusion one gets is weaker, as the implemented variant might not be superior to the one it replaces. Still, this can be an acceptable trade-off due to the potentially vast improvement in testing speed and efficiency in cases where the major reasons for running a superiority test do not apply: the variant is easy to implement, costs nothing to maintain, is easily reversible and re-testing it is cheap.

There is an abundance of cases of both types in the conversion rate optimization and user texting practice, so non-inferiority should be considered by practitioners.

REFERENCES

[1] Fisher, R.A. (1935) - "The Design of Experiments", Edinburgh: Oliver & Boyd

[2] Schuirmann, D.J. (1987) – "A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability", *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680

[3] Silva, G. T. da, Logan, B. R., & Klein, J. P. (2008) – "Methods for Equivalence and Noninferiority Testing", *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation*, 15 (1 Suppl), 120–127

[4] Walker, E., & Nowacki, A. S. (2011) – "Understanding Equivalence and Noninferiority Testing", *Journal of General Internal Medicine*, 26(2), 192–196