MOST WINNING A/B TEST RESULTS ARE ILLUSORY

Martin Goodson (DPhil)
Research lead, Qubit

Produced on behalf of
Qubit.

INTRODUCTION

Marketers have begun to question the value of A/B testing, asking: ‘Where is my 20% uplift? Why doesn’t it ever seem to appear in the bottom line?’ Their A/B test reports an uplift of 20% and yet this increase never seems to translate into increased profits. So what’s going on?

In this article I’ll show that badly performed A/B tests can produce winning results which are more likely to be false than true. At best, this leads to the needless modification of websites; at worst, to modification which damages profits.

Statisticians have known for almost a hundred years how to ensure that experimenters don’t get misled by their experiments [1]. Using this methodology has resulted in pharmaceutical drugs that work, in bridges that stay up and in the knowledge that smoking causes cancer. I’ll show how these methods ensure equally robust results when applied to A/B testing.

To do this I’ll introduce three simple concepts which come as second nature to statisticians but have been forgotten by many web A/B testing specialists. The names of these concepts are ‘statistical power’, ‘multiple testing’ and ‘regression to the mean’. Armed with them, you will be able to cut through the misinformation and confusion that plague this industry.

1 STATISTICAL POWER

Statistical power is simply the probability that a statistical test will detect a difference between two values when there *truly* is an underlying difference. It is normally expressed as a percentage.

Imagine you are trying to find out whether there is a difference between the heights of men and women. If you only measured a single man and a single women you would stand a risk that you don't detect the fact that men are taller than women. Why? Because random fluctuations mean you might choose an especially tall woman or an especially short man, just by chance.

However, if you measure many people, the averages for men and women will eventually stabilize and you will detect the difference that exists between them. That's because *statistical power* increases with the size of your 'sample' (statistician-speak for 'the number of people that you measure').

We are interested in the conversion rate difference between control and variant versions of a website. Let's take the example of a free-delivery special offer. We run an A/B test by specifying a control group, who don't see the special offer, and a variant group, who do. Our hypothesis is that the special offer will increase the probability that visitors will purchase something. As with testing the heights of men and women, whether we detect this 'uplift' in conversions depends on the *statistical power* of the test. More people; more power.

So why is statistical power important? First, if you don't calculate the sample size required up-front you might not run your experiment for long enough. Even if there is an uplift you won't have enough data to be able to detect it. The experiment will likely be a waste of time.

But under-powered experiments have a much more insidious effect. It's that any winning variants you *do* see are likely to be 'false positives'. They appear to generate an uplift but will not actually generate any increase in revenue.

Let's say our A/B test will take two months to gain good statistical power. If we decide to save time by only running the test for only two weeks instead, this test will be under-powered [2]. The result? Almost two-thirds of winning tests will be completely bogus (see box 1). Don't be surprised if revenues stay flat or even go down after implementing a few tests like these.

Please ignore the well-meaning advice that is often given on the internet about A/B testing and sample size. For instance, a recent article recommended stopping a test after only 500 conversions [3]. I've even seen tests run on only 150 people [4] or after only 100 conversions [5]. This will not work. The truth is that nearer 6000 conversion events (not necessarily purchase events) are needed (see box 1).

BOX 1 CALCULATING STATISTICAL POWER

The calculation of the required sample size for a statistical test is known as a power calculation. The power is simply the probability of detecting a genuine difference between two values. A well-powered test has power of between 80% and 90%.

We know that, occasionally, a test will generate a false positive due to random chance - we can't avoid that. By convention we normally fix this probability at 5%. You might have heard this called the 'significance level'.

Data from Google suggest that a new variant of a website is generally only 10% likely to cause a true uplift [6]. 90% have no effect or might degrade the performance of the website.

Let's imagine we perform 100 tests on a website and, by running each test for 2 months, we have a large enough sample to achieve 80% power. 10 out of our 100 variants will be truly effective and we expect to detect 80%, or 8, of these true effects. If we use a p-value cutoff of 5% we also expect to see 5 false positives[1]. So, on average, we will see 8 + 5 = 13 winning results from 100 A/B tests.

This is for a well-powered test. But what happens if the test is under-powered? Let's say you are too impatient to wait for two months so you cut the test short after two weeks. The smaller sample size reduces the power of this test from a respectable 80% to less than 30%. Now you will have 3 true positives and 5 false positives: 63% of your winning tests are completely imaginary.

You can perform power calculations by using an online calculator or a statistical package like R. If time is short, a simple rule of thumb is to use 6000 conversion events in each group if you want to detect a 5% uplift. Use 1600 if you only want to detect uplifts of over 10%. These numbers will give you around 80% power to detect a true effect.

But remember: if you limit yourself to detecting uplifts of over 10% you will miss also miss negative effects of less than 10%. Can you afford to reduce your conversion rate by 5% without realizing it?

[1] 5% of 90 is 4.5, rounded up to 5

2 MULTIPLE TESTING

Most testing software uses the system of *p-values* for testing statistical significance – this is known as the ’classical method’ of testing. There is nothing inherently wrong with this method. Indeed, most clinical trials are analyzed using classical methods. However, here are two well known dangers of using *p-values*:

- Performing many tests, not necessarily concurrently, will multiply the probability of encountering a false positive.
- False positives increase if you stop a test when you see a positive result [7].

The way medical trials are conducted ensures that these dangers are avoided. Unfortunately, the way most A/B testing software is designed means that one or both of these dangers are likely to affect A/B testing on the web.

Stopping tests as soon as you see winning results will create false positives.

Some software for A/B testing is designed in such a way that it’s natural to constantly monitor the results of a test, stopping the test as soon as a significant result is achieved [8, 9, 10, 11]. Alarmingly, false positives can comprise as much as 80% of winning test results when tests are performed in this way.

We can see how false positives surge out of control by using a computer to simulate thousands of A/A tests (technical appendix). A/A tests are simply tests where the variant is the same as the control. It follows that any successful A/A test must, by design, be a false positive.

Say we check once per day on our A/A test until completion, stopping the test once we achieve a *p-value* of less than 5%. Our simulation shows that this method results in a successful test result 41% of the time. A simple calculation shows that, under this scenario, 80% of winning results are, in fact, false[1].That’s not a misprint: at least 80% of the winning results are completely worthless.

To give them credit, a vendor of A/B testing software has recently been trying to come to terms with this problem [12]. It would appear that their customers have been asking why even A/A tests seem to produce just as many winning results as A/B tests!

[1] The false positive rate is 41%. If we assume 10% of variants have a real effect (Box 1), at most 10 of 100 tests will be true positives. Out of 100 tests, on average 41% will therefore be false positives and 10%, true positives. Of 51 winning tests, over 80% will actually be false.

I should emphasize that choosing to stop the test after peeking at the the results is quite different to running an under-powered test. An under-powered test happens when you choose, *before looking at the results*, to run the test for a short length of time. It is nothing like as treacherous as ‘peeking’.

It remains to explain: why is this phenomenon remarkably unknown and rarely voiced in the industry? Upton Sinclair said it best: “It is difficult to get a man to understand something, when his salary depends on his not understanding it”.

Multiple testing

A dangerous recent fashion has been to perform many A/B tests simultaneously in a hope that something will be successful [13]. A variant of this is to perform ‘post-test segmentation’ – splitting up your sample after a test is performed until you achieve a positive result.

Simple arithmetic shows why this is a bad idea. Each test has a 5% chance of winning even if there is no real uplift (if we use a 5% p-value threshold). If you run 20 tests, or try 20 segmentations, you will on average have one winning test and if you try 40 you will have two, *even if the variants do not generate any uplift*.

So there is always likely to be a feel-good factor from a few wins if you try many tests. But this is illusory unless two things are true. One, each test had good statistical power; two, there was a good reason, or supportive data, for the test in the first place [14].

Rather than a scattergun approach, it’s best to perform a small number of focused and well-grounded tests, all of which have adequate statistical power.

3 REGRESSION TO THE MEAN

Tests which seem to be successful but then lose their uplift over time are likely false positives.

Be very skeptical of tests which seem to be successful but whose performance degrades over time or during a follow-on validation test. It has been suggested that this is due to some kind of ‘novelty effect’ [15]. My view is that this is much more likely to be simply a false positive. Of course the uplift doesn’t stand up over time – that’s because there was no uplift to begin with.

This is a well-known phenomenon, called ‘regression to the mean’ by statisticians. Again, this is common knowledge among statisticians but does not seem to be more widely known.

Here is an example from Wikipedia on regression to the mean [16]: let’s say you have a class of students who each take a 100-item true/false test on a certain subject. Suppose each student chooses randomly on all questions. Each student would achieve a random score between 0 and 100, with an average of 50.

Now take only the top scoring 10% of the class and, declaring them ‘winners’, give them a second test, on which they again choose randomly. They will score less on the second test than the first test. That’s because, no matter what they scored on the first test they will still average 50 correct answers in the second test. It will falsely appear that suddenly they have become less ‘knowledgeable’ about the subject.

The same thing happens in A/B testing. If your original winning test was a false positive then any further testing will of course show a reduction in the uplift caused by the variant (Figure 1).

Whenever you see tests which don’t seem to maintain uplift over time ask yourself: was my original test conducted properly (e.g., was it adequately powered)? If you want to be sure of the result then always perform a second validation study to check your results are robust.

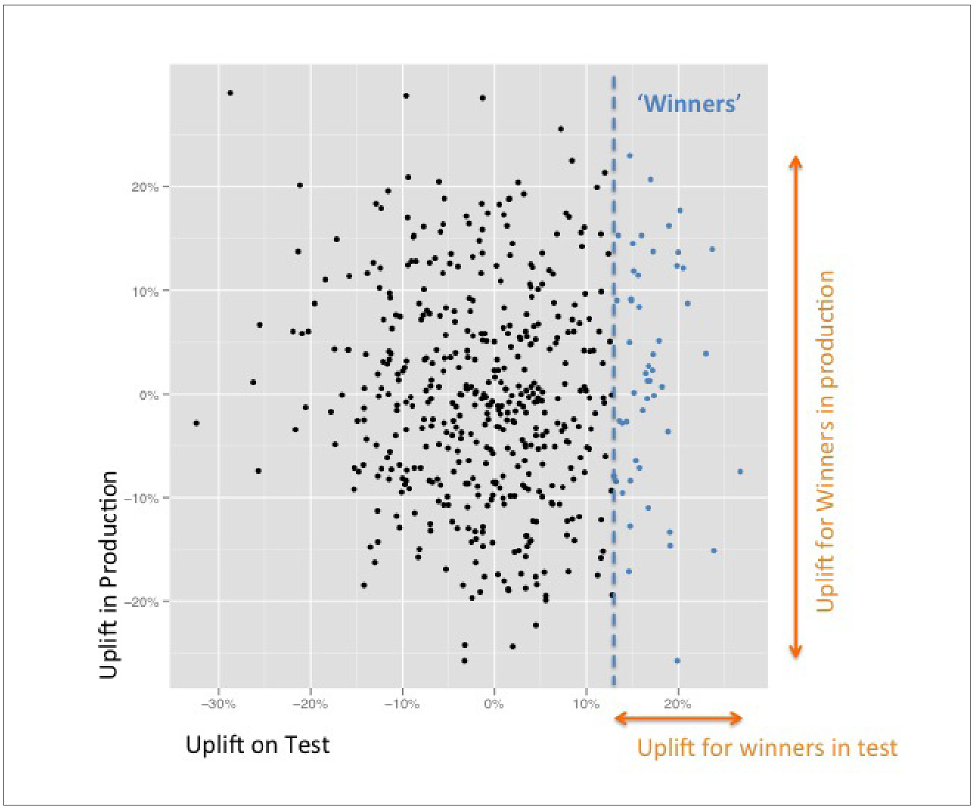

Figure 1. Regression to the mean causes an apparent 'novelty effect'. If we run 500 A/A tests and separate out the apparent 'winners', the uplift they drive in production or in follow-on validation will be lower than the uplift in the original test.

A/B tests tend to over-estimate uplift

There is a related effect which means that estimates of uplift from A/B tests tend to be *over-estimates*. This is called the 'winner's curse'. It happens because we are more likely to declare a winning test when, by chance, there is a large difference between the control and the variant conversion rates. It's obvious put like that: we overestimate the size of an effect for winning tests simply because if we *didn't* observe a large uplift we wouldn't have declared it a winner!

So, even if there is a true difference it *is* likely to be lower than the apparent uplift. The winner's curse is much stronger if the sample size is small - because small samples show bigger fluctuations [17]. So, if you are using small sample sizes you will be hit by both false positives *and* grossly inflated estimates of uplift.

SUMMARY

I’ve outlined some of the problems that beset A/B testing on the web. In most cases bad A/B testing comes from forgetting the concepts of statistical power, multiple testing and regression to the mean.

You can increase the robustness of your testing process by following this statistical standard practice:

- Use a valid hypothesis - don’t use a scattergun approach
- Do a power calculation first to estimate sample size
- Do not stop the test early if you use ‘classical methods’ of testing
- Perform a second ‘validation’ test repeating your original test to check that the effect is real

A/B TESTING GLOSSARY

Control: the original form of a website
Variant: a modified form of a website whose performance we want to test.
A/B test: an experiment to detect whether a variant version generates more conversions than a control version of a website.
False positive: an apparently winning but spurious result
True positive: an apparently winning result stemming from a genuine difference in performance between control and variant.
P-value: the probability that the measured difference in conversions would have been observed if there were no underlying difference between the control and variant.
Statistical power: the probablity that a statistical test will detect a difference between the control and variant, if there really is such a difference.
Conversion: any desired goal, often a purchase event.
Effect: a true difference between control and variant groups.

REFERENCES

[1] *http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124*

[2] *https://help.optimizely.com/hc/en-us/articles/200133789-How-long-to-run-a-test*

[3] *http://bentilly.blogspot.co.uk/2012/10/ab-testing-scale-cheat-sheet.html*

[4] *http://www.tone.co.uk/i-see-so-ab-testing-actually-works-then-conversion-rate-optimisation-case-study/*

[5] *http://blog.optimizely.com/2013/02/06/is-it-done-yet-getting-real-about-test-length-estimates-and-calling-a-test/*

[6] MANZI JIM, *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* (2012)

[7] *http://www.ncbi.nlm.nih.gov/books/NBK13920/*

[8] *http://www.bloggingtips.com/2012/09/23/increase-20-click-out-rates-through-ab-testing/*

[9] *http://blog.hubspot.com/marketing/how-to-run-an-ab-test-ht*

[10] *http://blog.kissmetrics.com/ab-tests-big-wins/*

[11] *http://www.maxymiser.com/resources/ab-testing*

[12] *https://help.optimizely.com/hc/en-us/articles/200040355-Running-and-interpreting-an-A-A-test*

[13] *http://monetate.com/2012/01/infographic-are-you-running-enough-tests-on-your-website/*

[14] *http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124*

[15] *http://monetate.com/2012/04/the-roller-coaster-ride-of-statistical-significance/*

[16] *http://en.wikipedia.org/wiki/Regression_toward_the_mean*

[17] *http://dcscience.net/ioannidis-associations-2008.pdf*

TECHNICAL APPENDIX

STATISTICAL TESTS FOR A/B TESTING

Conversion events can be modelled using a binomial distribution:

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

n = the number of visitors
p = conversion rate

To compare two conversion rates, for instance in an A/B test, a normal approximation to the binomial is commonly used for each conversion rate:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

with parameters as follows:

mean: p
variance: p(1-p)/n

The null hypothesis, H_0, is that p is equal in both variant A and variant B
The alternative hypothesis, H_1, is that p differs between variant A and variant B with difference D
The difference between the parameters, d, under the null model is modelled as a normal distribution with
variance: 2*p(1-p)/n
mean: 0

from which it follows that, under the null hypothesis, H_0:
d/sqrt(2*p(1-p)/n) ~ standard normal distribution

Under the alternative hypothesis the test statistic is distributed as, H_1:
(d - D)/sqrt(2*p(1-p)/n) ~ standard normal distribution

STATISTICAL POWER

α is defined as the probability of a type I error.
β is defined as the probability of a type II error.
Statistical power is 1-β

The cutoff for the conversion rate difference for a desired value of α (the significance probability) can be calculated using the CDF of the standard normal distribution, which is the following integral:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Numerical optimization can be used to find a value of x that satisfies:

$$1 - \Phi\left(x \times \sqrt{\frac{n}{2p(1-p)}}\right) = \alpha$$

and a value of n that satisfies:

$$\Phi\left((x - D) \times \sqrt{\frac{n}{2p(1-p)}}\right) = \beta$$

for desired α and β

This process is conveniently implemented in the R function power.prop.test.

R CODE TO SIMULATE EFFECT OF EARLY-STOPPING FOR AN A/A TEST

Experimental design:
Setup A/A test with two variants and one control all of whom have the same conversion rate (1%).
Calculate sample size to get 80% power and set experiment running for two months
Check test results every day and stop experiment when either variant beats control
Repeat 10K times and calculate the rate of false positives

```
ff3=function(n, t=1e6){
# Setup up A/A test simulating two variants against control
    # conversions in control
    a=0
    # conversions in variant 1
    b1=0
    # conversions in variant 2
    b2=0

    for (i in 1:n){
           # get new day’s results for control
           alatest=rbinom(1, as.integer(t/n), 0.01)
           a = a + alatest

           # get new day’s results for variant 1
           b1latest=rbinom(1, as.integer(t/n), 0.01)
           b1 = b1 + b1latest

           # get new day’s results for variant 2
           b2latest=rbinom(1, as.integer(t/n), 0.01)
           b2 = b2 + b2latest
 # perform tests
 p1 = prop.test(c(a, b1),
c(i*as.integer(t/n), as.integer(i*t/n)), alternative=’less’)$p.value
           p2 = prop.test(c(a, b2), c(as.integer(i*t/n), as.integer(i*t/n)),
alternative=’less’)$p.value
 if (p1<=0.05|p2<=0.05) {return (T)}
```

```
       }
       return (F)
}
n = power.prop.test(p1=0.01, p2=0.01*1.05, power=0.8, alternative='one.sided')$n
# run 10K A/A simulations at 80% power
res=sapply(1:10000, function(x)ff3(60, n))
print (mean(res))
[1]  0.4127
```

Please direct inquiries to research@qubitproducts.com