# Segmenting your traffic? You are probably doing it wrong. (https://www.chrisstucchio.com/blog/2015/ab_testing_segments_and

Mon 05 January 2015 ab testing (https://www.chrisstucchio.com/tag/ab-testing.html) / segmentation (https://www.chrisstucchio.com/tag/segmentation.html) / multiple comparisons (https://www.chrisstucchio.com/tag/multiple-comparisons.html)

Get notified of new posts

So you've jumped onboard the A/B testing bandwagon. You've just run an A/B test comparing the site redesign to the old version. Unfortunately the redesign did not differ in a statistically significant way from the old version. At this point, a variety of conversion rate experts (http://online-behavior.com/targeting/segment-or-die-214) will tell you to segment your data (http://conversionxl.com/how-to-build-a-strong-ab-testing-plan-that-gets-results/):

> An experiment that seemed to be performing poorly might have actually been successful, but only for a certain segment. For example, our experiment may have shown that a variation of our mobile landing page is not performing well. When looking into the segments though, we may see that it is performing exceptionally well for Android users, but badly for iPhone users. When not looking at segments, you can miss this detail.

This is an intuitively natural idea - everyone is different, so why isn't it possible that one version will perform worse for some groups than another? As a data geek, it also gives us the opportunity to play with Pandas or Excel and build an impressive presentation to show our boss.

**Resist the urge to do this!**

Suppose you ignore my advice. You've decided to segment your users and look for interesting results. I can virtually guarantee that if you put enough effort in, **}sy { mpār h** statistically significant correlations. Maybe your visitors using Android from Oklahoma (all 58 of them) have a significantly higher conversion rate on the redesign than on the original site.

How can I guarantee that you'll see something interesting even if there is nothing there? Because of the multiple comparisons problem (https://en.wikipedia.org/wiki/Multiple_comparisons_problem).

# Multiple comparisons - a review

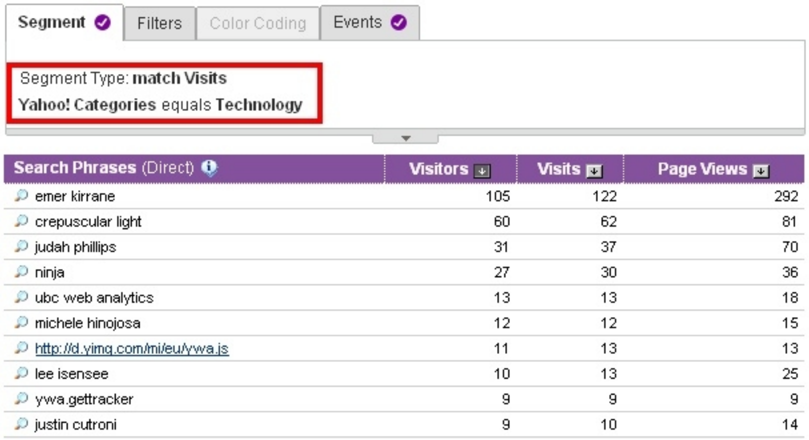
Suppose you run statistical tests with a p-value cutoff of 5%. What this means is that if you were to repeatedly run an A/A test (a test comparing the control group to itself), you would expect 5% of your tests to return a statistically significant value. In essence, the P-value cutoff is the false positive rate you've decided is acceptable.

So now lets think about how many segments you have. You've got mobile and desktop, 50 states, and perhaps 20 significant sources of referral traffic (google search, partner links, etc). All told, that's 2 x 50 x 20 = 2000 segments. Now lets assume that each segment is identical to every other segment; if you segment your data, you'll get 0.05 x 2000 = 100 **wexmwxgepp} wkr mãger x**results purely by chance. With a little luck, Android users in Kentucky referred by Google, iPhone users in Nebraska referred by Direct and Desktop users in NJ all preferred the redesign. Wow!

## It's worse than that

The calculation I've done above assumes that you have enough traffic so that you can get statistically significant data in **i egl wikqirx** In reality that assumption is likely to be false. In real life, you almost certainly have a tiny amount of data for each segment.

The folks advocating segmentation don't seem to understand this. I'm not arguing against straw men here - this is an actual picture taken from an article advocating segmentation (http://online-behavior.com/targeting/audience-segmentation):



Segment Type: **match Visits**
**Yahoo! Categories** equals **Technology**

| Search Phrases (Direct) | Visitors | Visits | Page Views |
|---|---|---|---|
| emer kirrane | 105 | 122 | 292 |
| crepuscular light | 60 | 62 | 81 |
| judah phillips | 31 | 37 | 70 |
| ninja | 27 | 30 | 36 |
| ubc web analytics | 13 | 13 | 18 |
| michele hinojosa | 12 | 12 | 15 |
| http://d.yimg.com/mi/eu/ywa.js | 11 | 13 | 13 |
| lee isensee | 10 | 13 | 25 |
| ywa.gettracker | 9 | 9 | 9 |
| justin cutroni | 9 | 10 | 14 |

The largest of those segments has 100 visitors! You simply do not have enough data to determine whether searches for "ninja" or "crepuscular light" will result in more conversions. Sorry, you are out of luck. Stop segmenting and don't try again until you've increased your traffic by 100x.

# Multiple goals - the same problem applies

A lot of people, in addition to segmentation, like to track multiple goals on their site. For example, newsletter signups, add item to shopping card, or save item for later. Congratulations - by using multiple sufficiently many goals, you'll definitely find a statistically significant result in one of them.

This effect is partially mitigated if your goals are correlated with each other. I.e., if people who sign up for the newsletter also tend to add an item to the shopping cart, then the issue of multiple goals is reduced. On the other hand, the more your goals are correlated with each other, the less useful information you actually get out of tracking multiple goals.

## An easy fix for multiple goals

The best way to handle the problem of multiple goals is to define One Key Metric (OKM). For example, you might define your OKM as:

```
OKM = 10 x purchase + 1 x newsletter_signup + 1.5 x save_item_for_later
```

Then when making decisions, you have a single number which incorporates all the factors you are interested in. You can freely run any statistical test you like on the OKM without having to worry about multiple goals.

# How to fix the problem of multiple comparisons

Ok, you are still determined to segment your traffic or use multiple goals. Now it's time for one weird trick (https://en.wikipedia.org/wiki/%C5%A0id%C3%A1k_correction) to use to avoid running into the problems I've described above. It's a simple formula you can use. Suppose you want to run a segmented test with a p-value cutoff of 0.05. You can use the following formula to compute a **ri{** cutoff that works with multiple segments:

```
new_p_cutoff = 1 - (1 - old_p_cutoff)^(1/number_of_segments)
```

According to this formula, if we have 20 segments, `new_p_cutoff=0.00256`. So suppose you've run a test with 20 segments. If you want to have a 5% chance of observing a false positive in the test, then you must declare any individual test to be statistically insignificant unless it yields a p-value smaller than 0.00256.

You can use the same formula with multiple goals as well. This formula is called the Sidak Correction (https://en.wikipedia.org/wiki/%C5%A0id%C3%A1k_correction), by the way.

It's possible your A/B testing tool has this built in, but you should not assume they do the right thing. Most do not.

# Experimenter degrees of freedom

This problem is trickier to fix. Experimenter degrees of freedom come into play when determining **{lexxs xiwx** When looking for something interesting, one might first try segmenting by browser. When that fails, one might then try segmenting by location, and if that fails by demographic. The first test involves segmenting 5 ways, so the experimenter will plug `number_of_segments=5` into the Sidak Correction above. The second test involves 50 segments, so the experimenter plugs `number_of_segments=50` into the formula.

Look kosher? It's not.

The problem is that by the time the experimenter finished segmenting by browser, **li epvieh} leh e 9) glergi sj wiimnk e jepwi tswmxmzi**. The second segmentation attempt introduced **ersxliv** 5% chance of making an error. So the data analyist now has a 10% chance, rather than a 5% chance, of seeing a false positive!

## How to reduce experimenter degrees of freedom

The only way to avoid this problem is **tvivikmwxvexmsr**. Before you look at the data, decide how many segmentation attempts you will make. Ask yourself: "Self, hypothetically, if segmenting by browser doesn't give anything interesting, what will I do next?" Once you've decided this, you must then count the number of segments **jvsq izi v} exxiqtx}sy qmklxtswwmfp} qeoi**.

So in the above example, `number_of_segments = 5 + 50 + 25` (assuming 5 browsers, 50 geographic locations and 25 demographic segments). That's the easy part.

The hard part is to **wxst** after you've done it all. At this point, you failed. No matter how you kick the data around, you'll only be obtaining statistically significant results by chance.

Andrew Gelman discusses this in a lot more detail in his article, The Garden of Forking Paths (http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf).

# But Google and Amazon make $hitton$ of money by segmenting and personalizing?!?

Google and Amazon have more traffic than you. They've also hired teams of statisticians to help them avoid making these mistakes.

# Conclusion

Lots of people on the internet are suggesting that you should segment your A/B testing data in order to understand things more deeply. Unfortunately, none of these people are telling you how to do it correctly. Segmentation is hard. Most of the time it doesn't give you anything useful. But unless you are very careful, of all the false positive will make it look like segmentation generated a big win.

## Subscribe to the mailing list

**Email Address** ▭

Subscribe

## Comments

7 Comments **Chris Stucchio's Blog** 1 Login

♡ Recommend Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

**Matthew Forr** • 4 years ago

Whoa, hold up there.

You started your article off citing two experts and their recommendations on segmentation and then pulled another bloggers article as evidence of why segmentation doesn't go well. If you go back to the first two articles they both argue for a much better segmentation strategy then Ms. Kirrane. Namely, segment based on source, behavior or outcome.

I wouldn't expect to see much of a difference in performance based on geography or location, they're inane ways of segmenting.

Also, your article does a great job of explaining why multiple comparisons can cause problems but that can be fixed by designing your A/B tests to be more directly related to the goal being tested.

Not saying you're totally wrong but it would have been nice if you explored the right way to segment traffic.

• Reply • Share ›

**stucchio** Mod → Matthew Forr • 4 years ago

The main point I'm trying to get across is that a lot of people who segment without being careful will run into far more false positives than real ones. Whether you segment by informative or uninformative characteristics doesn't change this - the point is simply that (# segments) x (p-value cutoff) = (# of false positives) unless you use the Sidak correction.

• Reply • Share ›

**Michael Chow** → stucchio • 4 years ago

That's not true. You're assuming that in all segments the null hypothesis is true. For example, if in every segment the null hypothesis is false, then you will have no false positives. But also, if the null hypothesis is not a point (E.g. Something = 0), and the true value is far away from those in the alternative hypothesis, it might not be far fetched to run many segments without any false positives.

• Reply • Share ›

**stucchio** Mod → Michael Chow • 4 years ago

Yes, I'm approaching this from a frequentist perspective - I'm computing P(positive | null hypothesis).

You are correct that P(positive && null hypothesis) might be different from this - you'd need to do a Bayesian calculation from some prior to find that out. In this more general case, the number of false positives will be approximately (p value cutoff) x (# of negative segments). So it's true that if (# of negative segments) is low, you won't get many false positives.

• Reply • Share ›

**Michael Chow** → stucchio • 4 years ago

In the second case I mentioned everything is the null hypothesis, and p(positive | null hypothesis) can still be arbitrarily low! For example, if I did a one sample t-test, and was just testing the positive tail, but the true parameter was large in the negative direction.

Sorry, I don't mean to nit pick. Your comment just reminded me of the (important) fringe cases :).

• Reply • Share ›

**stucchio** Mod → Michael Chow • 4 years ago

No worries. The main theme of this blog is nitpicking the details. The only thing that is forbidden here is platitudes and vague generalities. ;)

• Reply • Share ›

**William Högman** • 4 years ago

Thanks for raising such an important point. Loads of practitioners probably just plug their data into calculators on the web without knowing the statistics behind it, shame really.

• Reply • Share ›

✉ Subscribe Add Disqus to your siteAdd DisqusAdd Disqus' Privacy PolicyPrivacy PolicyPrivacy

🡅 Back to top