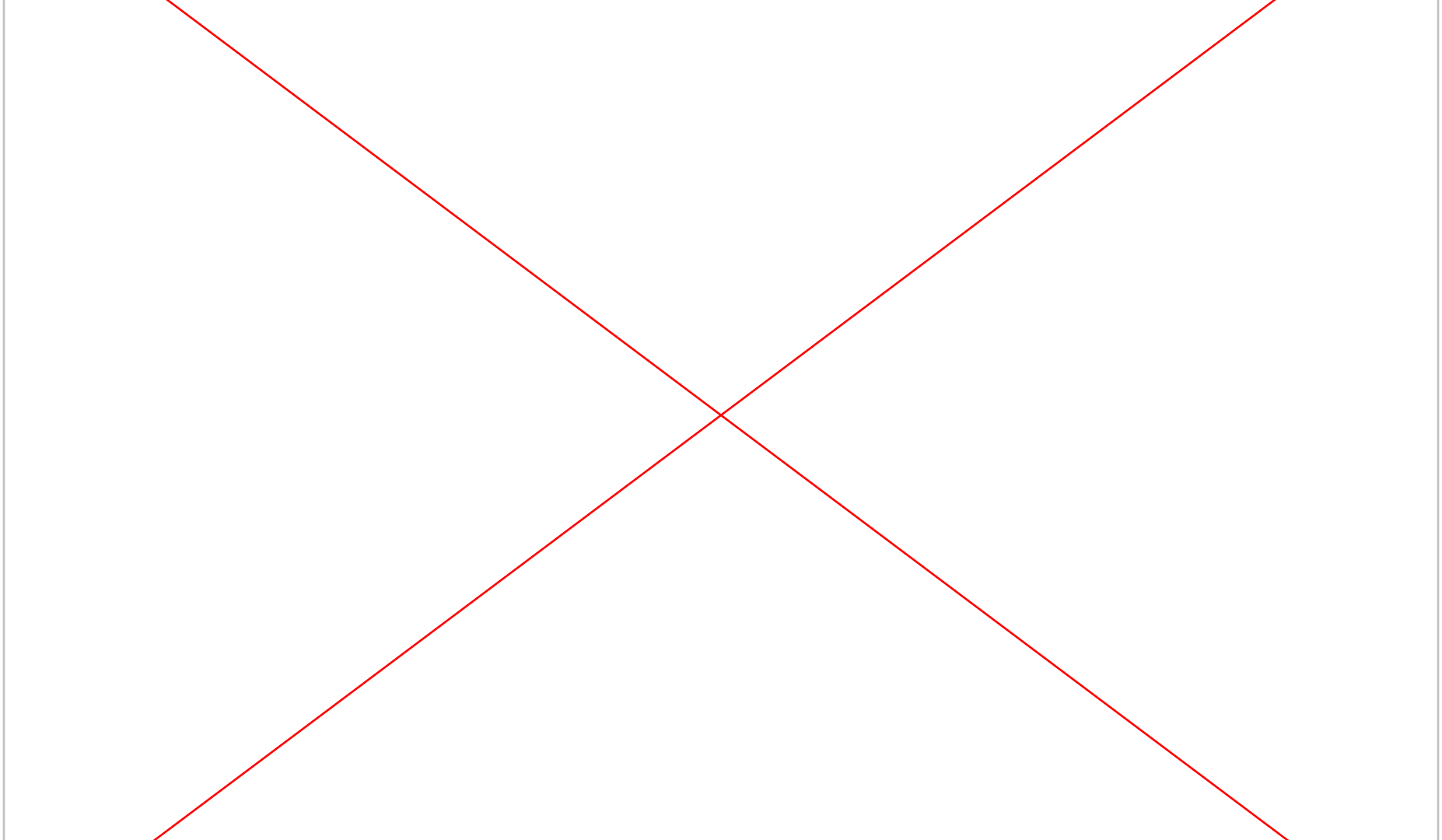
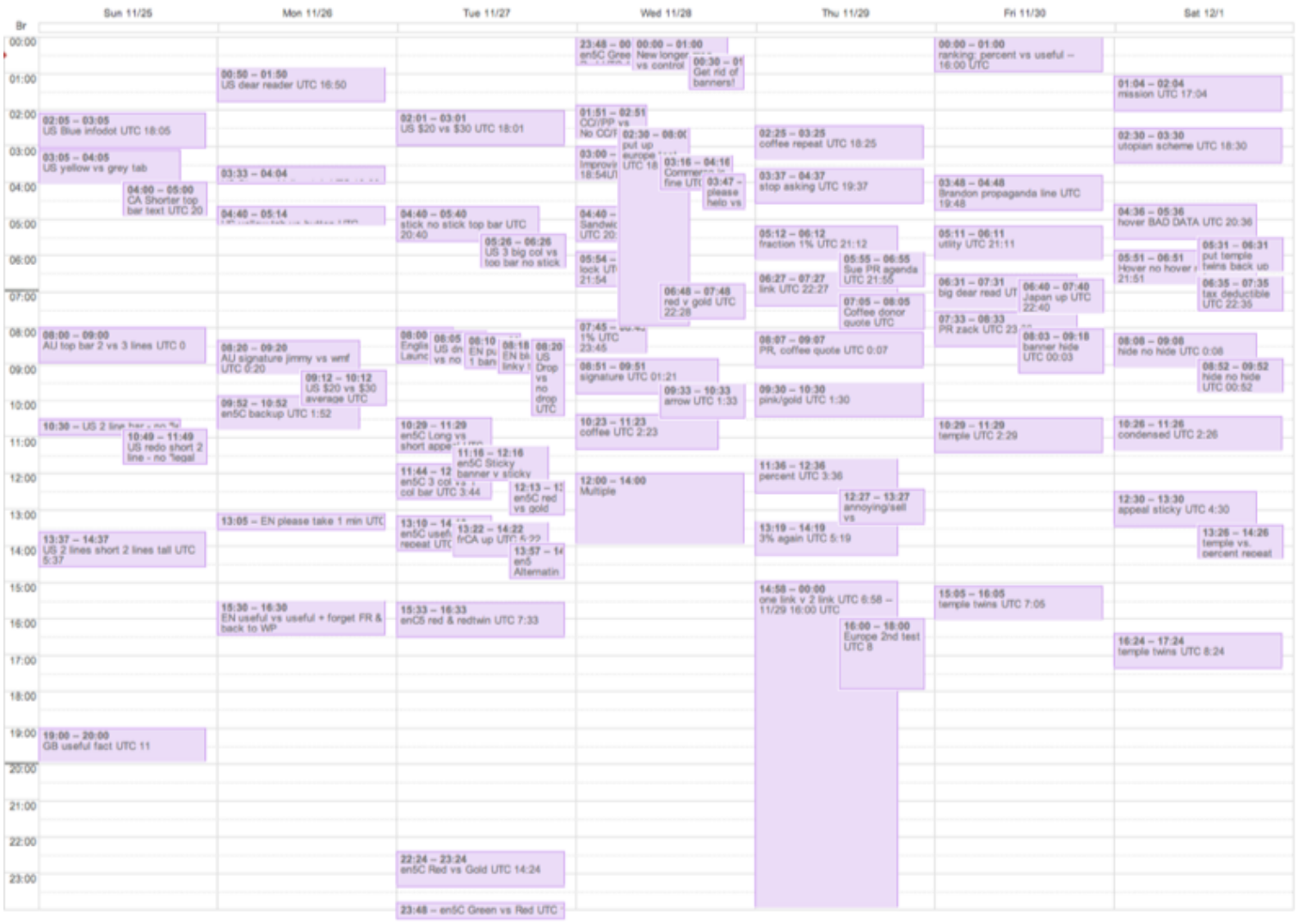
Are these your goals too?

1) To improve some metric.
2) To do as many tests as possible.
3) To find big breakthroughs…
4) ...and incremental gains.




Br
Sun 11/25
Mon 11/26
Tue 11/27
Wed 11/28
Thu 11/29
Fri 11/30
Sat 12/1
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
02:05 – 03:05
US Blue infodot UTC 18:05
03:05 – 04:05
US yellow vs grey tab
04:00 – 05:00
CA Shorter top bar text UTC 20
08:00 – 09:00
AU top bar 2 vs 3 lines UTC 0
10:30 – US 2 line
10:49 – 11:49
US redo short 2 line - no "legal
13:37 – 14:37
US 2 lines short 2 lines tall UTC 5:37
19:00 – 20:00
GB useful fact UTC 11
00:50 – 01:50
US dear reader UTC 16:50
03:33 – 04:04
04:40 – 05:14
08:20 – 09:20
AU signature jimmy vs wnf UTC 0:20
09:12 – 10:12
US $20 vs $30 average UTC
09:52 – 10:52
en5C backup UTC 1:52
13:05 – EN please take 1 min UTC
15:30 – 16:30
EN useful vs useful + forget FR & back to WP
02:01 – 03:01
US $20 vs $30 UTC 18:01
04:40 – 05:40
stick no stick top bar UTC 20:40
05:26 – 06:26
US 3 big col vs too bar no stick
08:00
Englis
Launc
08:05
US dr
vs no
08:10
EN pu
1 barn
08:18
EN bl
linky
08:20
US
Drop
vs
no
drop
UTC
10:29 – 11:29
en5C Long vs
short appe
11:16 – 12:16
en5C Sticky
banner v sticky
11:44 – 12
en5C 3 col
col bar UTC 3:44
12:13 – 1
en5C red
vs gold
13:10 – 14
en5C usef
reoeat UTC
13:22 – 14:22
frCA up UTC
13:57 – 14
en5
Alternatin
15:33 – 16:33
enC5 red & redtwin UTC 7:33
22:24 – 23:24
en5C Red vs Gold UTC 14:24
23:48 – en5C Green vs Red UTC
23:48 – 00
en5C Gree
00:00 – 01:00
New longer
vs control
00:30 – 01
Get rid of
banners!
01:51 – 02:51
OC/PP vs
No CC/
02:30 – 08:0
put up
europe
UTC 18
03:00 –
Improvi
18:54UT
03:16 – 04:16
Commer
fine UT
03:47 –
please
helo vs
04:40 –
Sandwic
UTC 20:
05:54 –
lock UT
21:54
06:48 – 07:48
red v gold UTC 22:28
07:45 –
1% UTC
23:45
08:51 – 09:51
signature UTC 01:21
09:33 – 10:33
arrow UTC 1:33
10:23 – 11:23
coffee UTC 2:23
12:00 – 14:00
Multiple
02:25 – 03:25
coffee repeat UTC 18:25
03:37 – 04:37
stop asking UTC 19:37
05:12 – 06:12
fraction 1% UTC 21:12
05:55 – 06:55
Sue PR agenda UTC 21:55
06:27 – 07:27
link UTC 22:27
07:05 – 08:05
Coffee donor quote UTC
08:07 – 09:07
PR, coffee quote UTC 0:07
09:30 – 10:30
pink/gold UTC 1:30
11:36 – 12:36
percent UTC 3:36
12:27 – 13:27
annoying/sell vs
13:19 – 14:19
3% again UTC 5:19
14:58 – 00:00
one link v 2 link UTC 6:58 – 11/29 16:00 UTC
16:00 – 18:00
Europe 2nd test UTC 8
00:00 – 01:00
ranking: percent vs useful – 16:00 UTC
03:48 – 04:48
Brandon propaganda line UTC 19:48
05:11 – 06:11
utility UTC 21:11
06:31 – 07:31
big dear read UT
06:40 – 07:40
Japan up UTC 22:40
07:33 – 08:33
PR zack UTC 23
08:03 – 09:18
banner hide UTC 00:03
10:29 – 11:29
temple UTC 2:29
15:05 – 16:05
temple twins UTC 7:05
01:04 – 02:04
mission UTC 17:04
02:30 – 03:30
utopian scheme UTC 18:30
04:36 – 05:36
hover BAD DATA UTC 20:36
05:31 – 06:31
put temple twins back up
05:51 – 06:51
Hover no hover
21:51
06:35 – 07:35
tax deductible UTC 22:35
08:08 – 09:08
hide no hide UTC 0:08
08:52 – 09:52
hide no hide UTC 00:52
10:26 – 11:26
condensed UTC 2:26
12:30 – 13:30
appeal sticky UTC 4:30
13:26 – 14:26
temple vs. percent repeat
16:24 – 17:24
temple twins UTC 8:24

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Interaction
Help
About W
Commur
Recent c
Contact
Toolbox
Print/exp
Languag
Simple E

Please read:
An urgent appeal from
Wikipedia founder Jimmy Wales

Welcome to Wikipedia,
the free encyclopedia that anyone can edit.
3,552,826 articles in English

- Arts
- Biography
- Geography
- History
- Mathematics
- Science

- Wikipedia is non-profit, but it's the #5 website in the world, serving 470 million people every month. To protect our independence, we'll never run ads.
- Google might have close to a million servers. Yahoo has something like 13,000 staff. We have 679 servers and 95 staff.
- If you rely on Wikipedia, please consider donating $5, $20, $50 or whatever you can to keep it free.

WIKIPEDIA
The Free Encyclopedia

Main Page Discussion Read View source View history Search

Welcome to Wikipedia,
the free encyclopedia that anyone can edit.
3,838,716 articles in English

- Arts
- Biography
- Geography
- History
- Mathematics
- Science

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Today's featured article

The **loggerhead sea turtle** is an oceanic turtle distributed throughout the world. An adult weighs around 135 kilograms (300 lb), with the largest specimens weighing in at more than 454 kilograms (1,000 lb). The skin ranges from yellow

In the news

- **Christopher Loeak** is elected President of the Marsh Islands.
- Azerbaijan, Guatemala, Morocco, Pakistan, and Togo

Declaration of the Rights

en.wikipedia.org/wiki/Declaration_of_Rights_of_Man?reset=1?&banner=B13_031900_bluln_Cntrl_nd_enYY

Declaration of the Rights

en.wikipedia.org/wiki/Declaration_of_Rights_of_Man?reset=1?&banner=B13_031900_bluln_Bluln_nd_enYY

Dear Wikipedia readers: **We are the small non-profit that runs the #5 website in the world. We have only 150 staff but serve 450 million users, and have costs like any other top site: servers, power, programs, and staff. Wikipedia is something special. It is like a library or a public park. It is like a temple for the mind, a place we can all go to think and learn. To protect our independence, we'll never run ads. We take no government funds. We survive on donations averaging about $30. Now is the time we ask. If everyone reading this gave $3, our fundraiser would be done within an hour. If Wikipedia is useful to you, take one minute to keep it online and ad-free another year. Please help us forget fundraising and get back to Wikipedia. *Thank you.***

Please Help

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

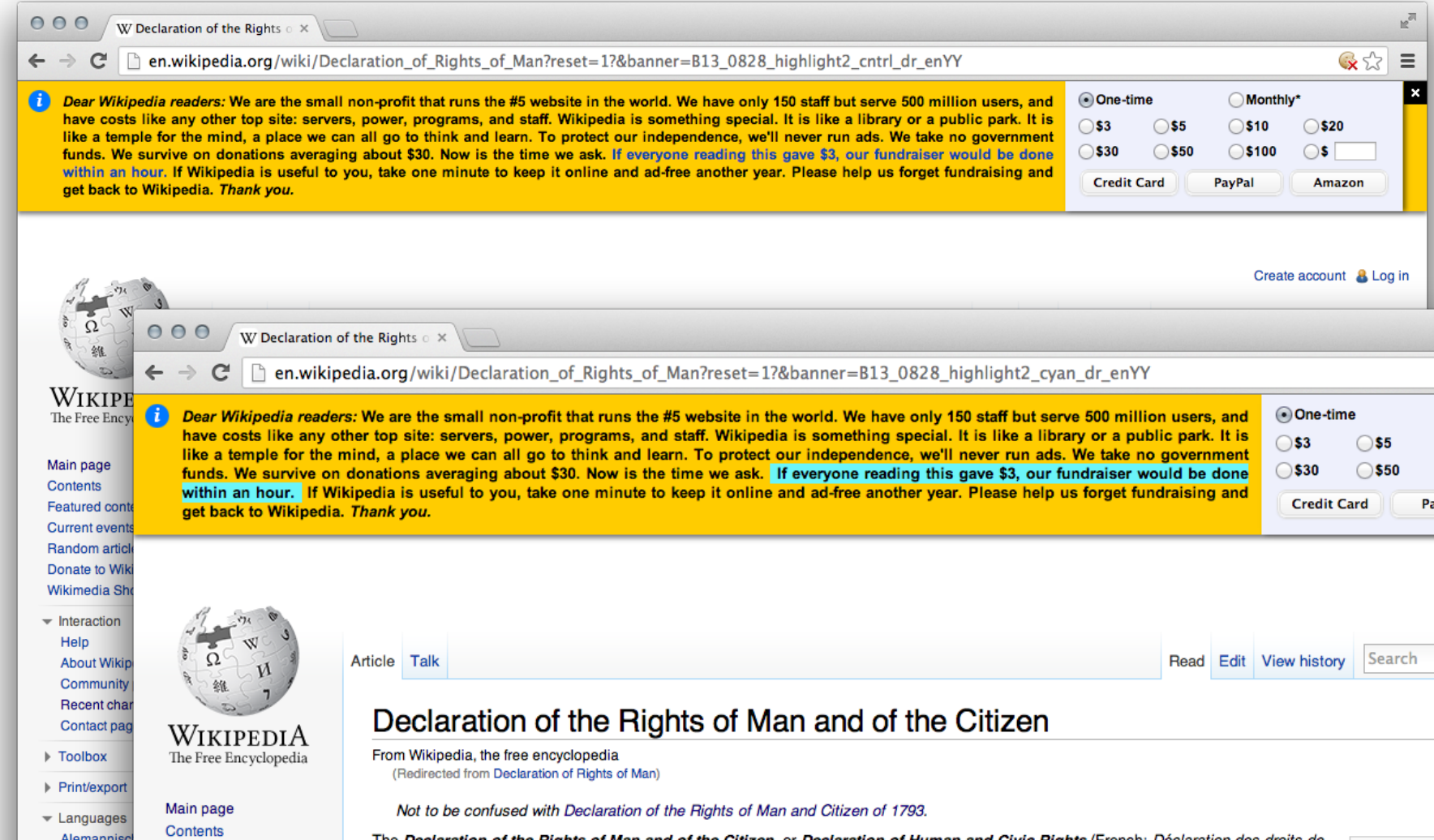
Declaration of the Rights of Man and of the Citizen

From Wikipedia, the free encyclopedia
(Redirected from Declaration of Rights of Man)

Not to be confused with Declaration of the Rights of Man and Citizen of 1793.

The ***Declaration of the Rights of Man and of the Citizen***, or ***Declaration of Human and Civic Rights*** (French: *Déclaration des droits de l'homme et du citoyen*) is a fundamental document of the French Revolution and in the history of human rights, defining the individual and collective rights of all the estates of the realm as universal. Influenced by the doctrine of "natural right", the rights of man are held to be universal: valid at all times and in every place, pertaining to human nature itself.

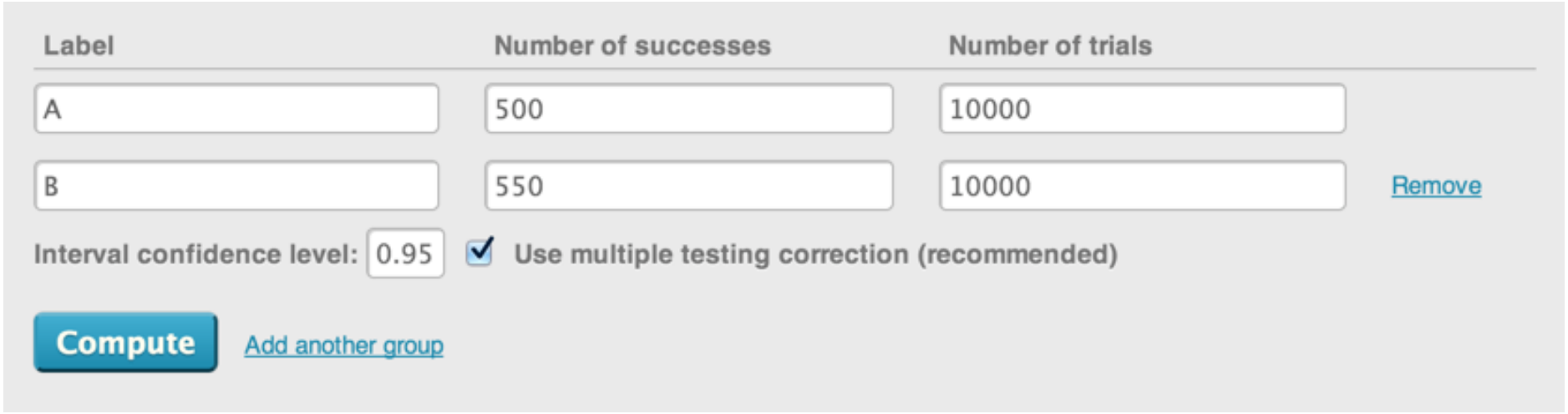
Contents [hide]


Declaration of the Rights
en.wikipedia.org/wiki/Declaration_of_Rights_of_Man?reset=1?&banner=B13_0828_highlight2_cntrl_dr_enYY
Dear Wikipedia readers: We are the small non-profit that runs the #5 website in the world. We have only 150 staff but serve 500 million users, and have costs like any other top site: servers, power, programs, and staff. Wikipedia is something special. It is like a library or a public park. It is like a temple for the mind, a place we can all go to think and learn. To protect our independence, we'll never run ads. We take no government funds. We survive on donations averaging about $30. Now is the time we ask. If everyone reading this gave $3, our fundraiser would be done within an hour. If Wikipedia is useful to you, take one minute to keep it online and ad-free another year. Please help us forget fundraising and get back to Wikipedia. Thank you.
One-time
Monthly*
$3
$5
$10
$20
$30
$50
$100
$
Credit Card
PayPal
Amazon
Create account
Log in
Declaration of the Rights
en.wikipedia.org/wiki/Declaration_of_Rights_of_Man?reset=1?&banner=B13_0828_highlight2_cyan_dr_enYY
Dear Wikipedia readers: We are the small non-profit that runs the #5 website in the world. We have only 150 staff but serve 500 million users, and have costs like any other top site: servers, power, programs, and staff. Wikipedia is something special. It is like a library or a public park. It is like a temple for the mind, a place we can all go to think and learn. To protect our independence, we'll never run ads. We take no government funds. We survive on donations averaging about $30. Now is the time we ask. If everyone reading this gave $3, our fundraiser would be done within an hour. If Wikipedia is useful to you, take one minute to keep it online and ad-free another year. Please help us forget fundraising and get back to Wikipedia. Thank you.
One-time
Monthly*
$3
$5
$10
$30
$50
$100
Credit Card
PayPal
Main page
Contents
Featured conte
Current events
Random articl
Donate to Wiki
Wikimedia Sho
Interaction
Help
About Wikip
Community
Recent char
Contact pag
Toolbox
Print/export
Languages
Create a
Article
Talk
Read
Edit
View history
Search
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Declaration of the Rights of Man and of the Citizen
From Wikipedia, the free encyclopedia
(Redirected from Declaration of Rights of Man)
Not to be confused with Declaration of the Rights of Man and Citizen of 1793.

NEED FOR Speed !!!

PLAY FREE!!

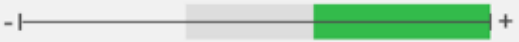
+180%



thumbtack

ABBA

A/B Test (Split Test) Calculator

	Successes	Total	Success Rate		p-value	Improvement
A -	500	10,000	4.6% – 5.4% (5%)		—	—
B -	550	10,000	5.1% – 6% (5.5%)		0.12	-2.4% – 22% (10%)

i.e.:

- B won in this sample.
- But you have a 6% chance of B actually being a loser. (And another 6% chance that B wins by a ton.)
- If you keep running this test, B will probably win by somewhere not too too far from 10%.

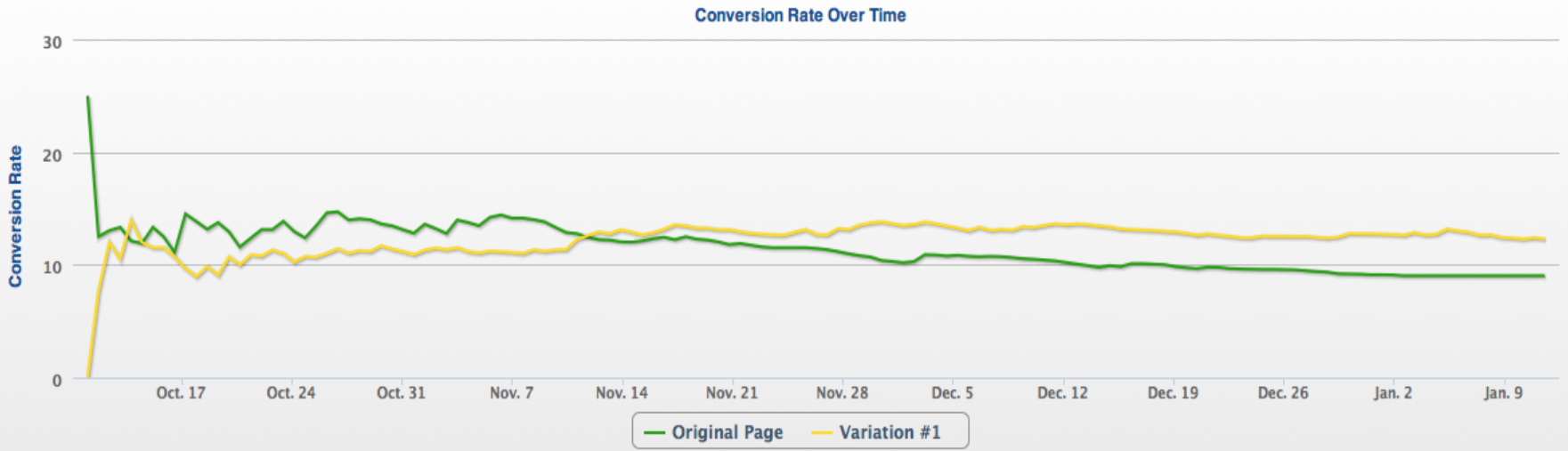
Add To Cart

Rename

The percentage of visitors who triggered Add To Cart (custom event).

Variation #1 is beating Original Page by +36.0%.

Variations	Statistics				
Experiment	Conversions / Visitors	Conversion Rate	Baseline	Chance to beat Baseline	Improvement
Variation #1	76 / 618	12.3% (±2.6%)		96.4%	+36.0%
Original Page	48 / 531	9.0% (±2.4%)	✓	—	—



Chance to beat Baseline	Improvement
✓ 96.4%	+36.0%
—	—

It is OK to peek.

!!

Evanmiller.org

How Not To Run An A/B Test

By Evan Miller

April 18, 2010

If you run A/B tests on your website and regularly check ongoing experiments for significant results, you might be falling prey to what statisticians call *repeated significance testing errors*. As a result, even though your dashboard says a result is statistically significant, there's a good chance that it's actually insignificant. This note explains why.

Background

When an A/B testing dashboard says there is a "95% chance of beating original" or "90% probability of statistical significance," it's asking the following question: Assuming there is no underlying difference between A and B, how often will we see a difference like we do in the data just by chance? The answer to that question is called the significance level, and "statistically significant results" mean that the significance level is low, e.g. 5% or 1%. Dashboards usually take the complement of this (e.g. 95% or 99%) and report it as a "chance of beating the original" or something like that.

However, the significance calculation makes a critical assumption that you have probably violated without even realizing it: *that the sample size was fixed in advance*. If instead of deciding ahead of time, "this experiment will collect exactly 1,000 observations," you say, "we'll run it until we see a significant difference," *all the reported significance levels become meaningless*. This result is completely counterintuitive and all the A/B testing packages out there ignore it, but I'll try to explain

Seven A/B testing mistakes you need to stop making in 2013

Posted in A/B Split Testing on January 4th, 2013

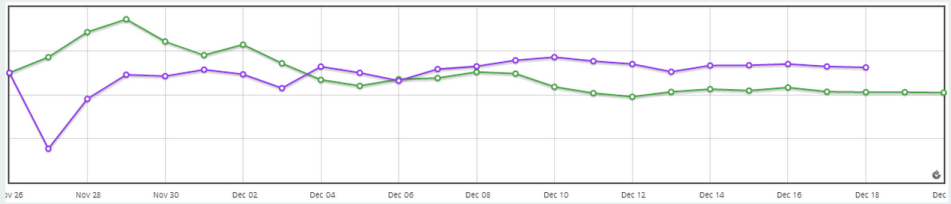
We've survived the annihilation predicted by the Mayans and made it into 2013. Ain't that absolutely awesome? What isn't so great is all those testing mistakes you've almost certainly made through 2012, mistakes (or bad practices) that were holding back your A/B testing and Conversion Rate Optimization efforts.

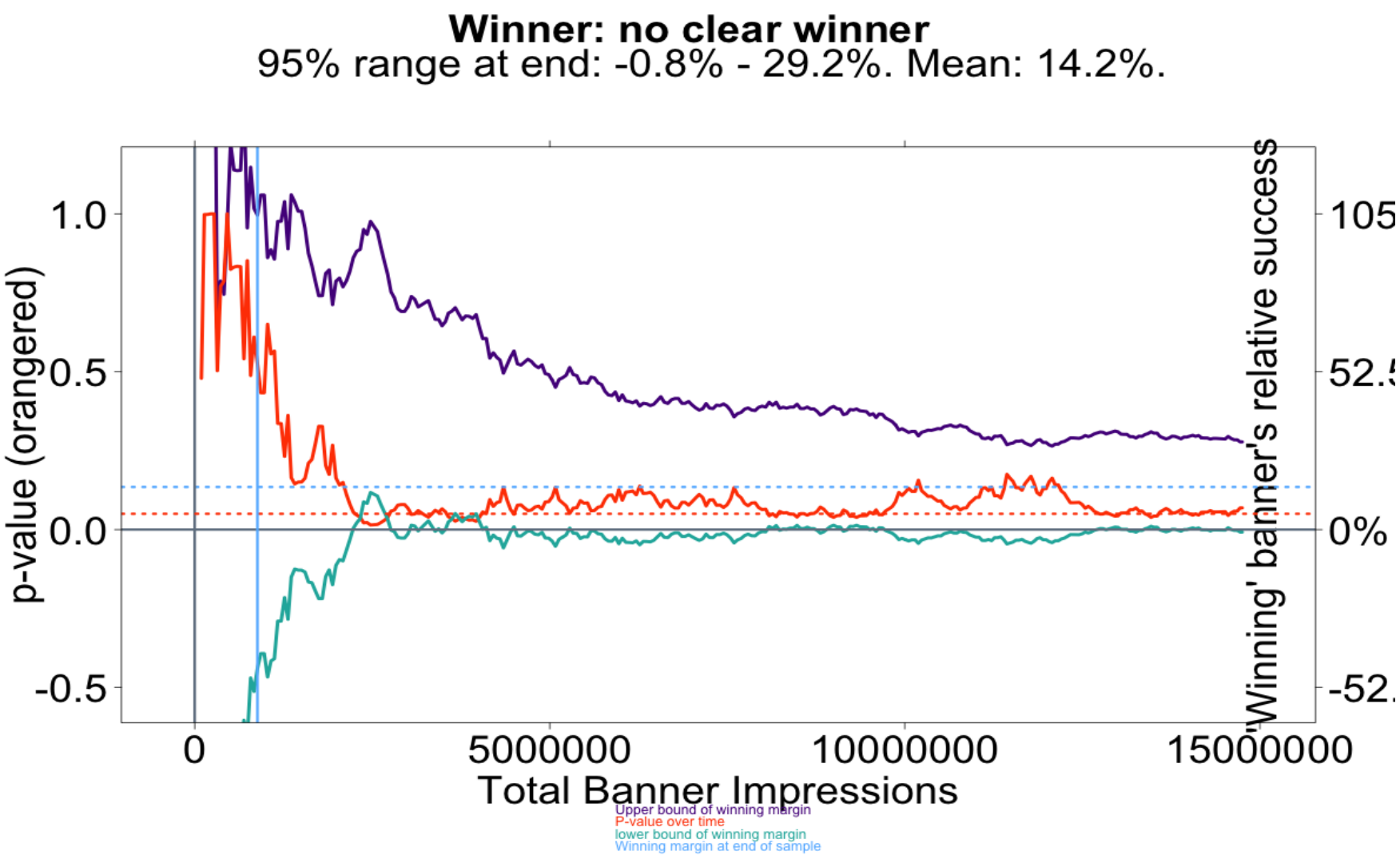
For your benefit, here's a quick recap of the sub-optimal practices you need to let go of, to truly achieve the gains promised by A/B testing.

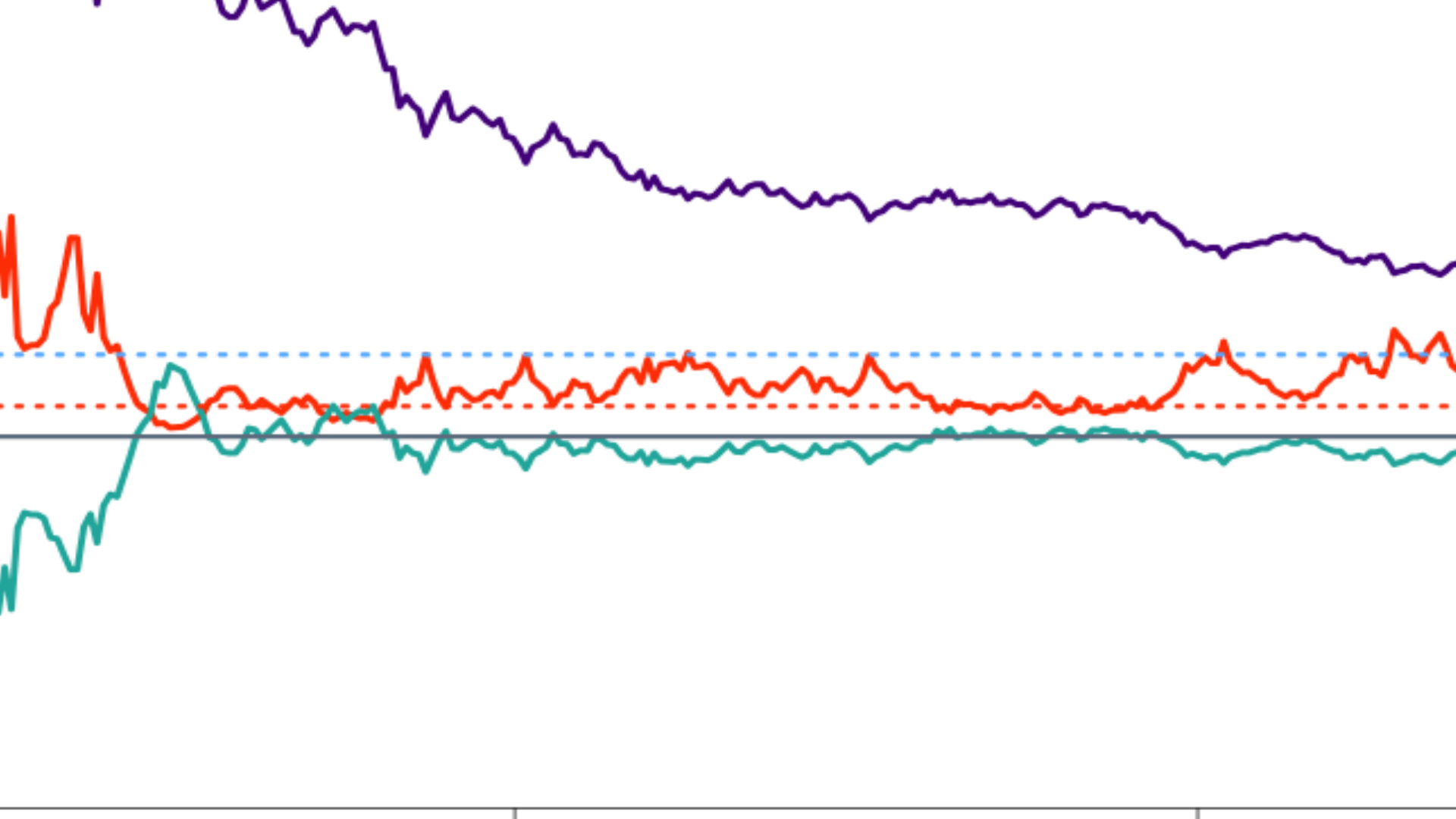
1) Not calculating your sample size before starting the test

Many marketing folks still don't calculate the number of visitors needed to run a test before starting the test. As pointed out by Evan Miller in his post "How Not to Run an A/B Test", you need to decide the required sample size before the test. This ensures that you don't get bitten by the euphoria (or depression) bug when you see your first statistically significant result and save yourself some bad decision-making.

To illustrate my point, most successful A/B test reports look like the one below.




Winner: no clear winner
95% range at end: -0.8% - 29.2%. Mean: 14.2%.
p-value (orangered)
1.0
0.5
0.0
-0.5
'Winning' banner's relative success
105
52.5
0%
-52.
0
5000000
10000000
15000000
Total Banner Impressions
Upper bound of winning margin
P-value over time
lower bound of winning margin
Winning margin at end of sample



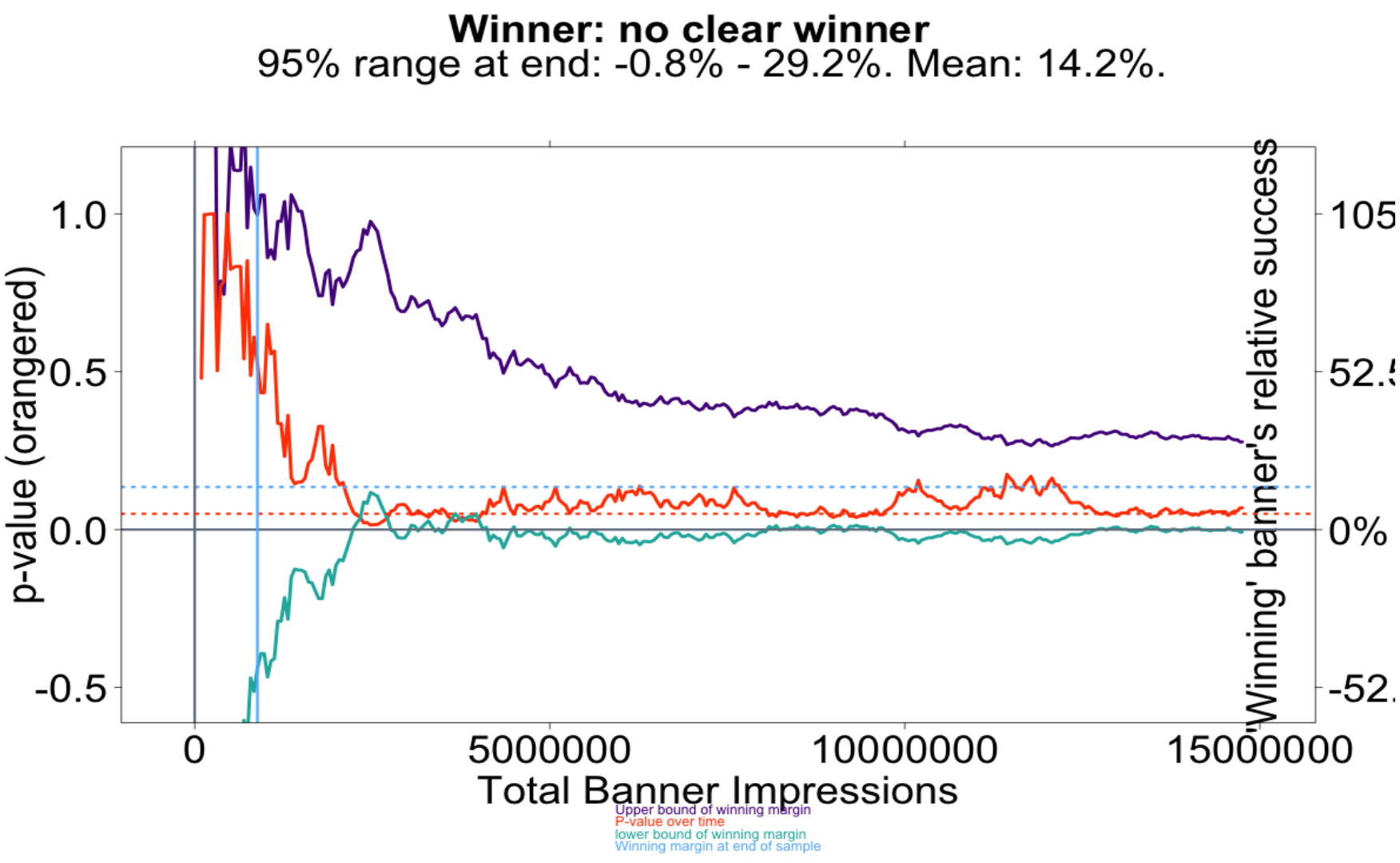
Not only is it OK to peek. You don’t even have to wait for 95% confidence!

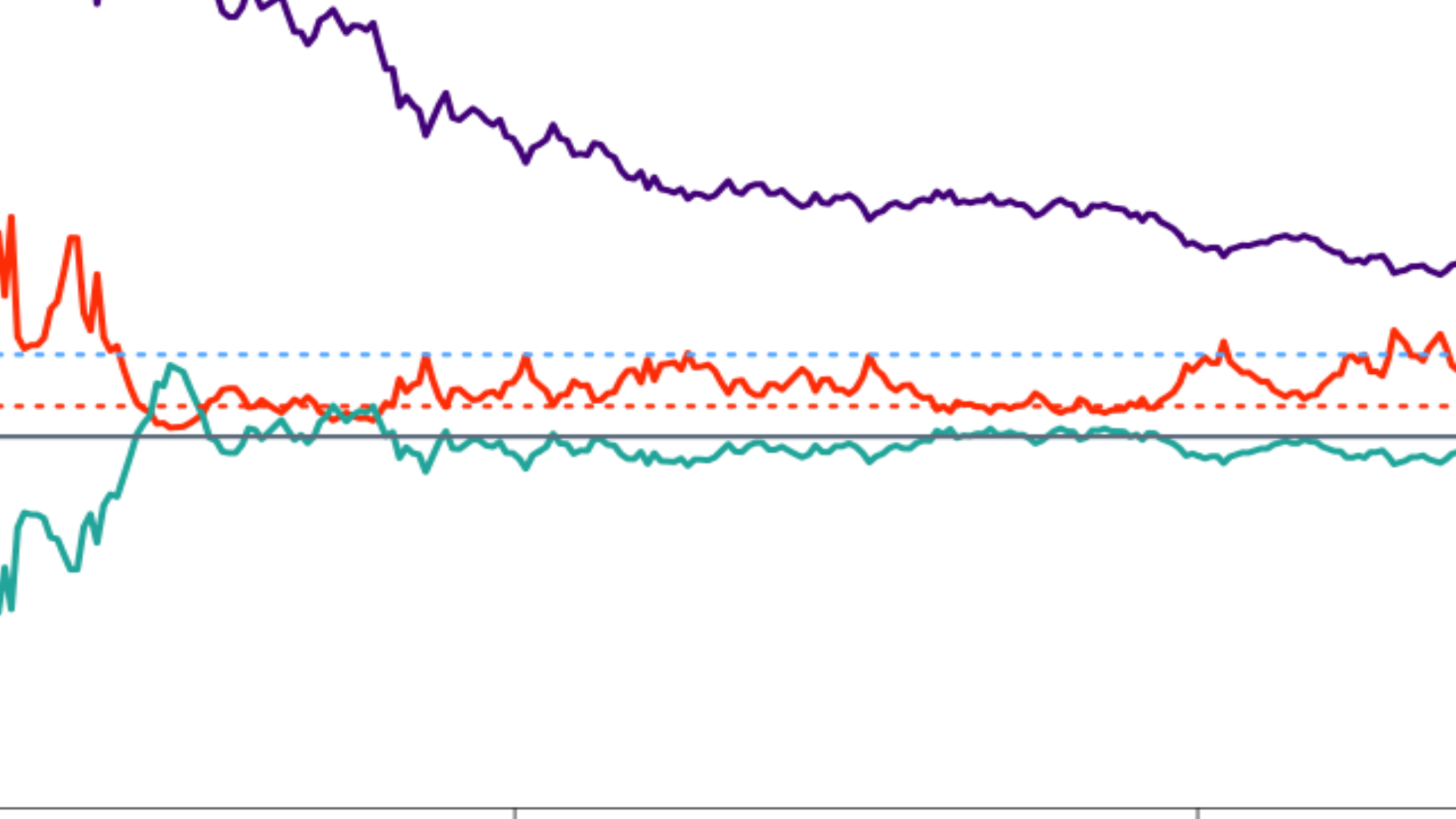
There’s no magic at p=.05 or p=.01
Every p value tells you something.

For example:

.3 = “probably a winner!”

.8 = “probably no big difference.”


Winner: no clear winner
95% range at end: -0.8% - 29.2%. Mean: 14.2%.
p-value (orangered)
1.0
0.5
0.0
-0.5
'Winning' banner's relative success
105
52.5
0%
-52.
0
5000000
10000000
15000000
Total Banner Impressions
Upper bound of winning margin
P-value over time
lower bound of winning margin
Winning margin at end of sample



OK to peek? REALLY? Yes, really. Let's think it through...

What if you peek during a moment when you've "falsely" gotten 95% confidence thanks to a handful of anomalous sales?

What if the 'true' confidence is only 90% -- i.e. if you ran the test much longer, you'd eventually get only 90% confidence.

OK, **What are you risking?**

You are mistakenly thinking that you have a 2.5% chance of picking a loser when you actually have a 5% chance of picking a loser.

BIG DEAL.

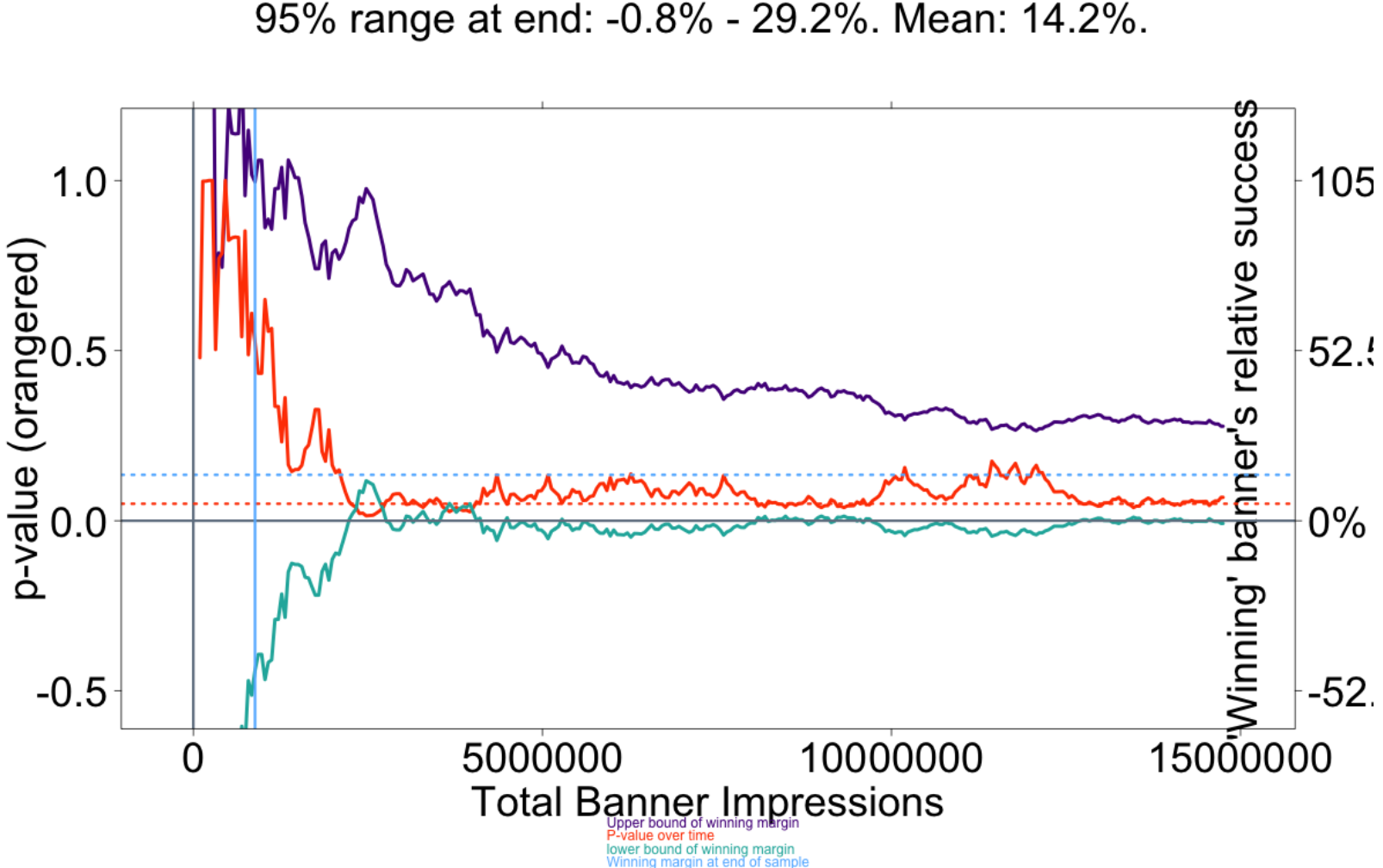
But here's what you gain:

You can move on to test something new!

Something that might make a huge difference!

So go for it!

If you're making an error, it will soon be rooted out if you're testing often enough.


Winner: no clear winner
95% range at end: -0.8% - 29.2%. Mean: 14.2%.
p-value (orangered)
1.0
0.5
0.0
-0.5
'Winning' banner's relative success
105
52.5
0%
-52.
0
5000000
10000000
15000000
Total Banner Impressions
Upper bound of winning margin
P-value over time
lower bound of winning margin
Winning margin at end of sample

OK to stop at 70% confidence? REALLY? Yes, really. Let's think it through...

That just means you're taking a 15% chance of hurting performance -- i.e. a 15% chance that you're using AB testing for **EVIL instead of GOOD!!! Oh no!**

Before you start hyperventilating: If you ARE hurting performance, chances are you're only hurting it by a percent or two. There's only a ***tiny*** chance that you're doing serious harm (to your sales...for a short time).

We're not landing someone on the moon, just playing with websites.

Out of 214 real Wikipedia tests we analyzed:

If we had stopped at the first sign of 70% confidence (after 15 donations):

We'd pick the winner : 90% of the time
We'd pick the loser: 10% of the time.

Our tests were on average 72% too long.

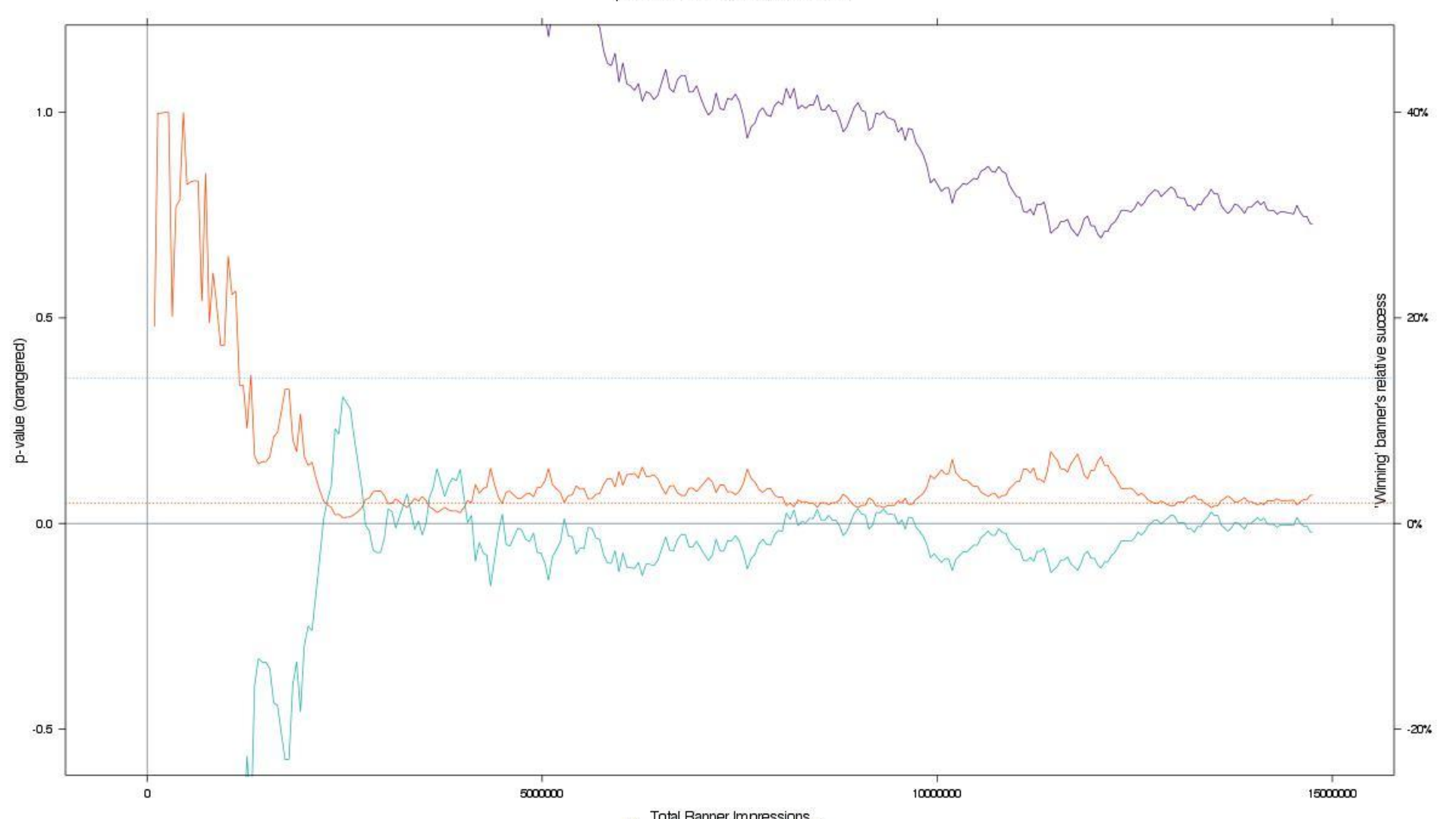
We could have done 3.6 TIMES MORE testing!

(if we were OK with that trade off, which we are!)

Hey, guess what!

When the lower bound of the confidence interval crosses zero, you have confidence!

(Now that’s something they didn’t teach you in AB testing school.)

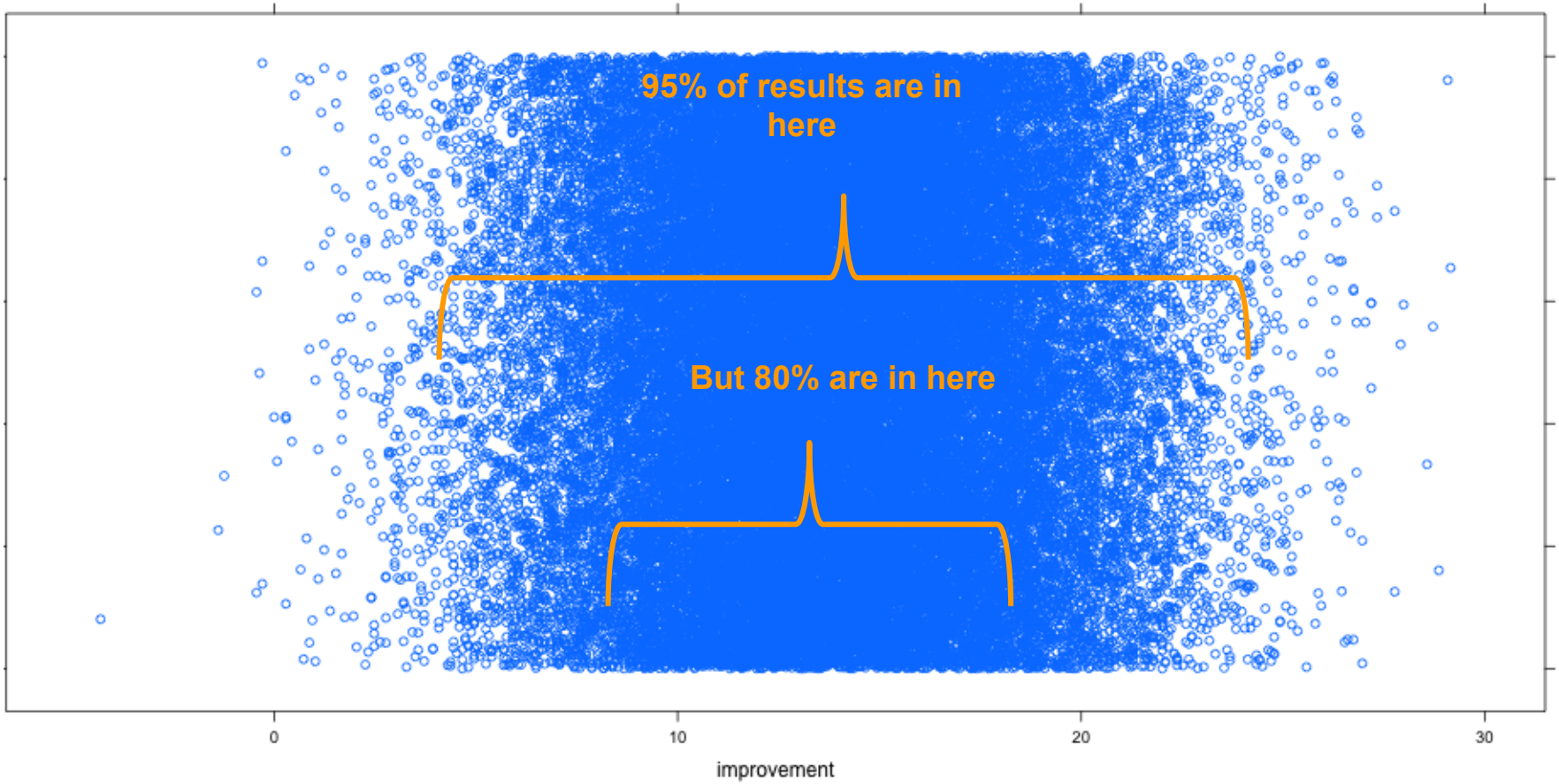

1.0
0.5
0.0
-0.5
p-value (orangered)
40%
20%
0%
-20%
'Winning' banner's relative success
0
5000000
10000000
15000000
Total Banner Impressions

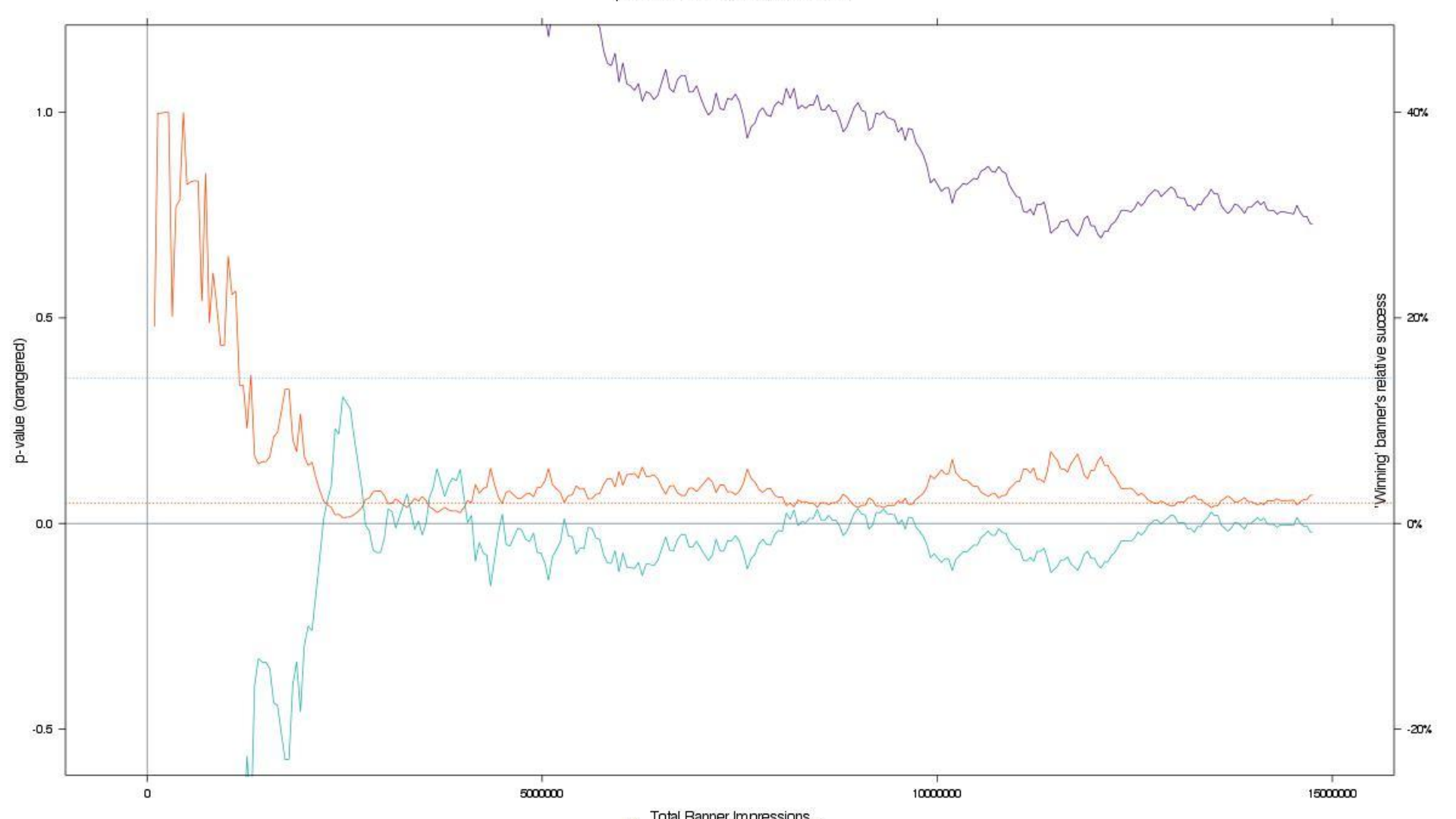
And that’s why we say….

p is nice.

But confidence interval is where it’s at.

There’s no cliff at 95% or 99% confidence.


95% of results are in here
But 80% are in here
0
10
20
30
improvement

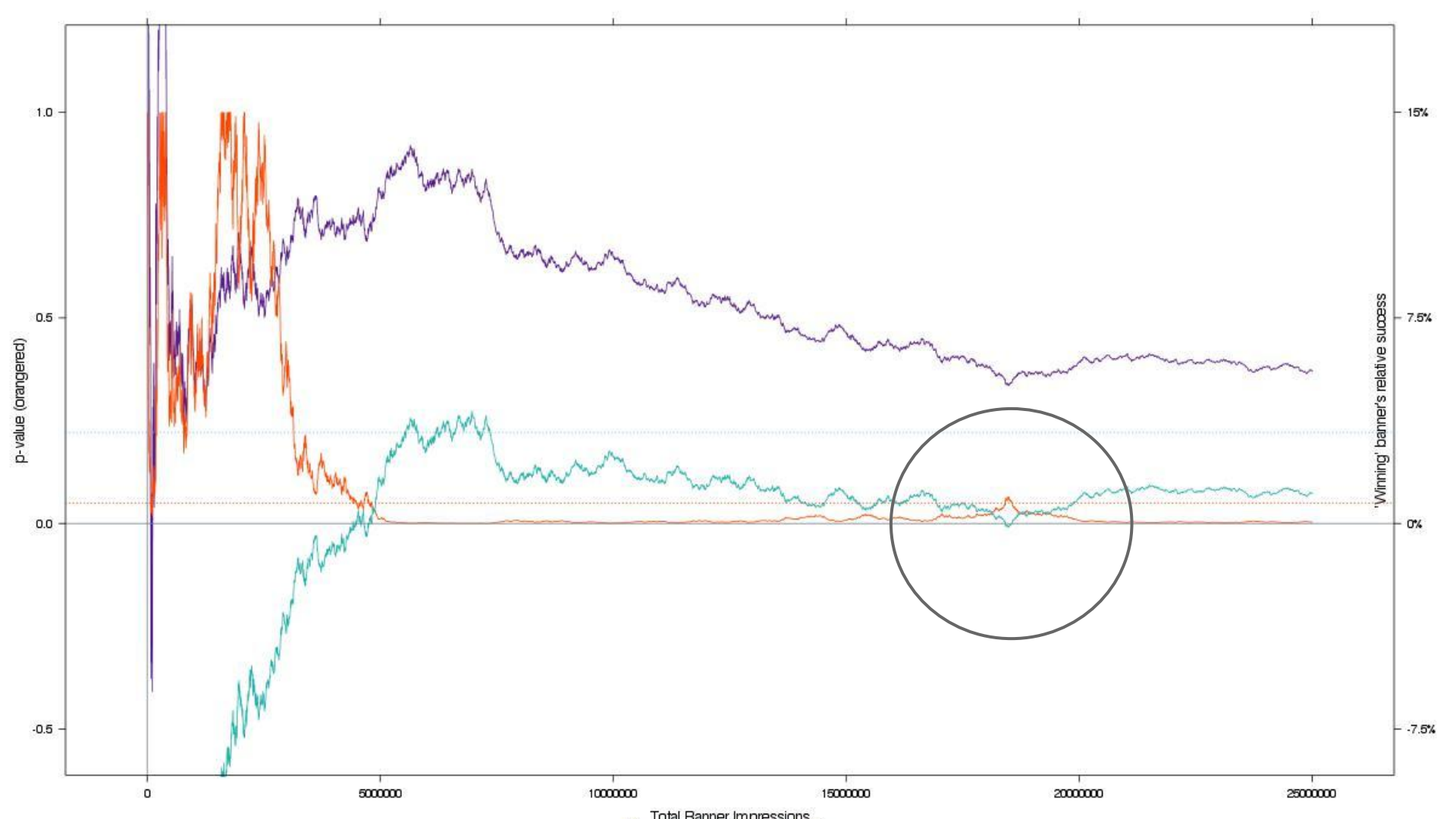

1.0
0.5
0.0
-0.5
p-value (orangered)
40%
20%
0%
-20%
'Winning' banner's relative success
0
5000000
10000000
15000000
Total Banner Impressions

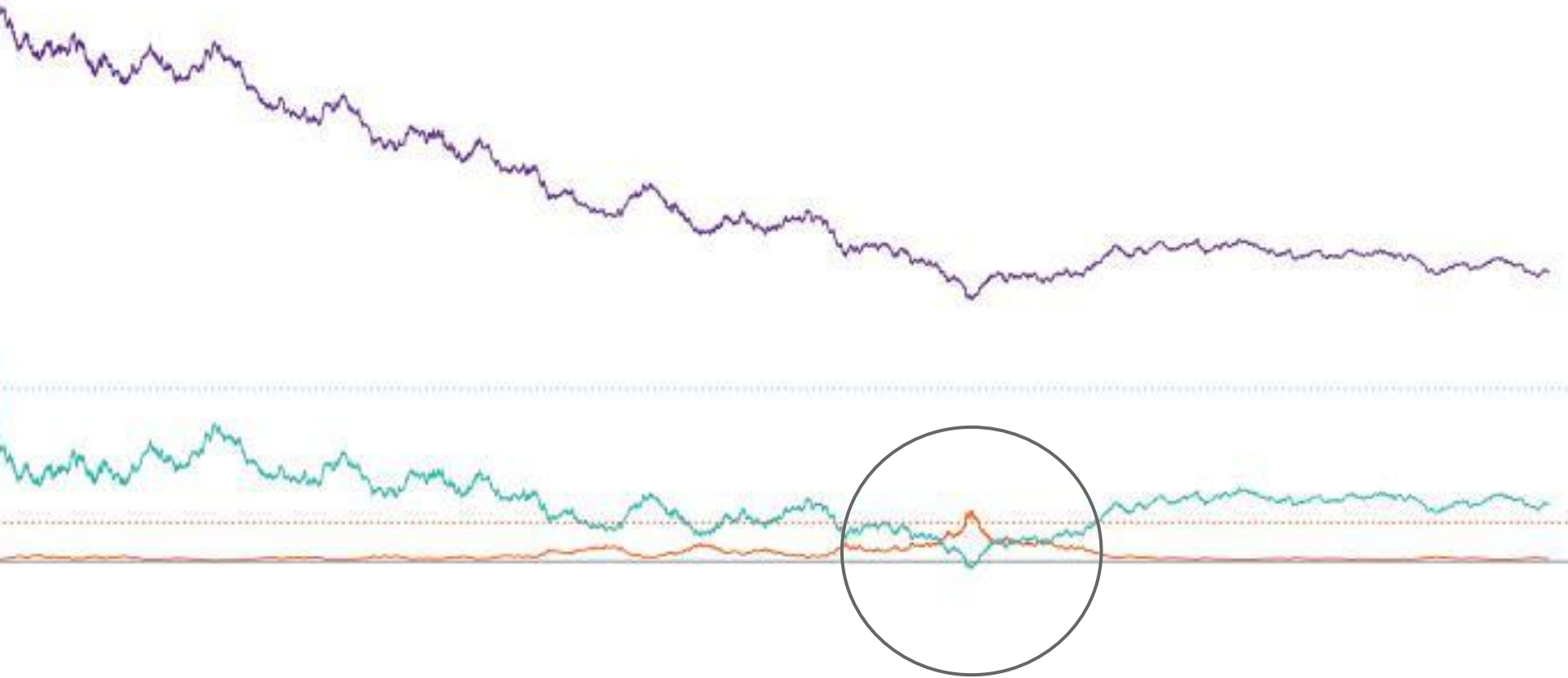
Now for some finer points and other tips.

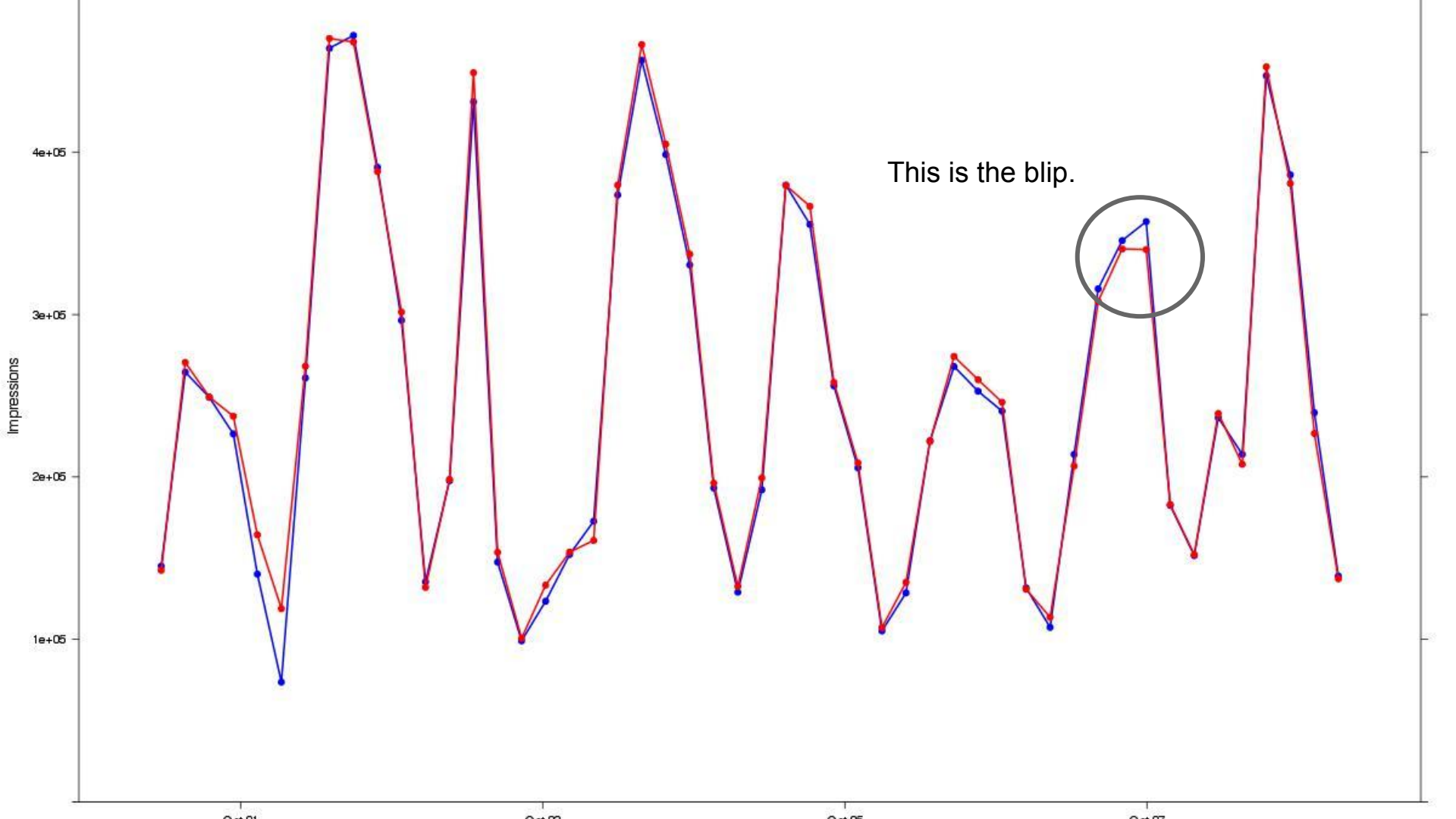
Don’t freak out when…

p shoots up for a moment.

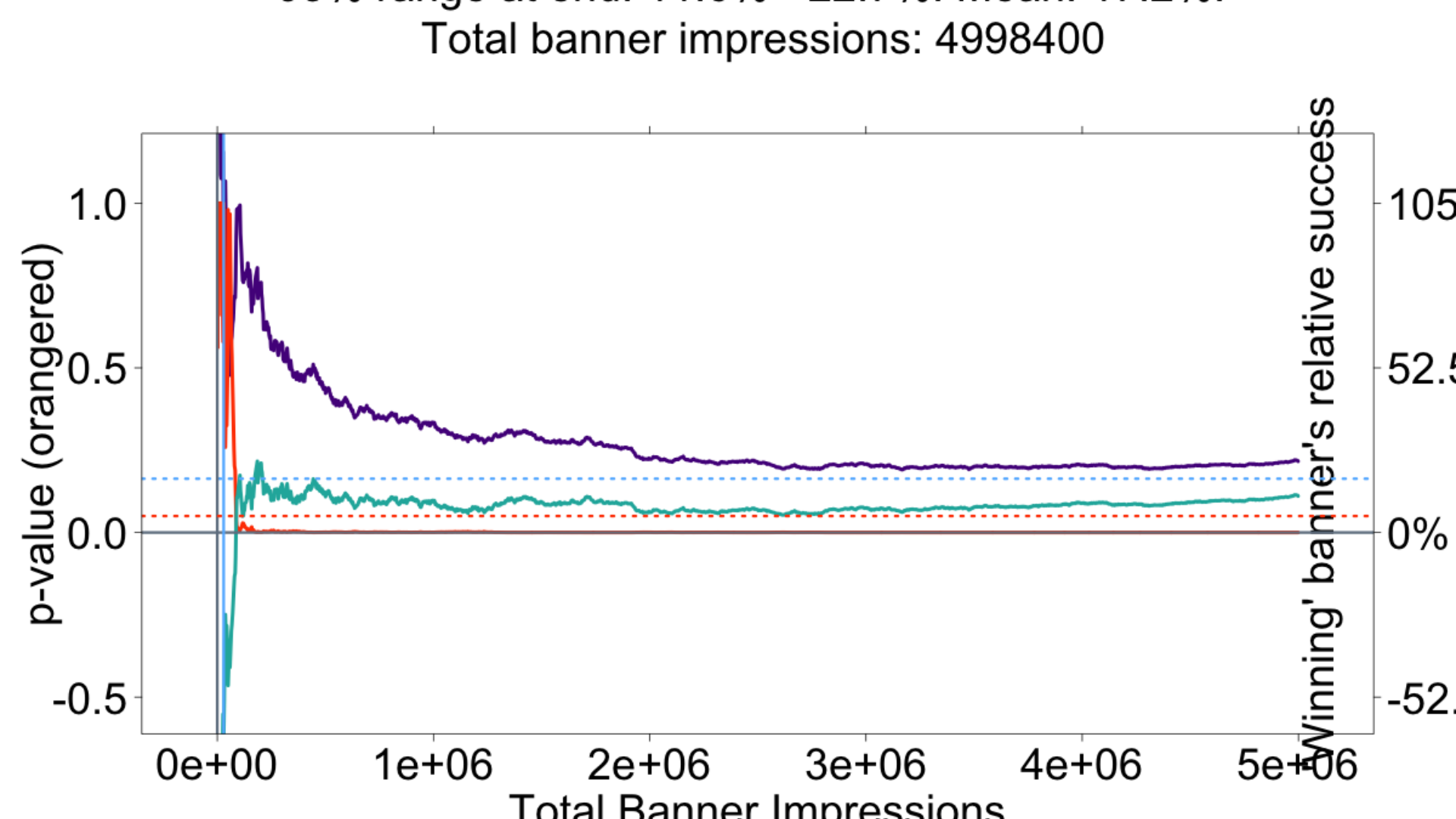
It’s just an edge case.

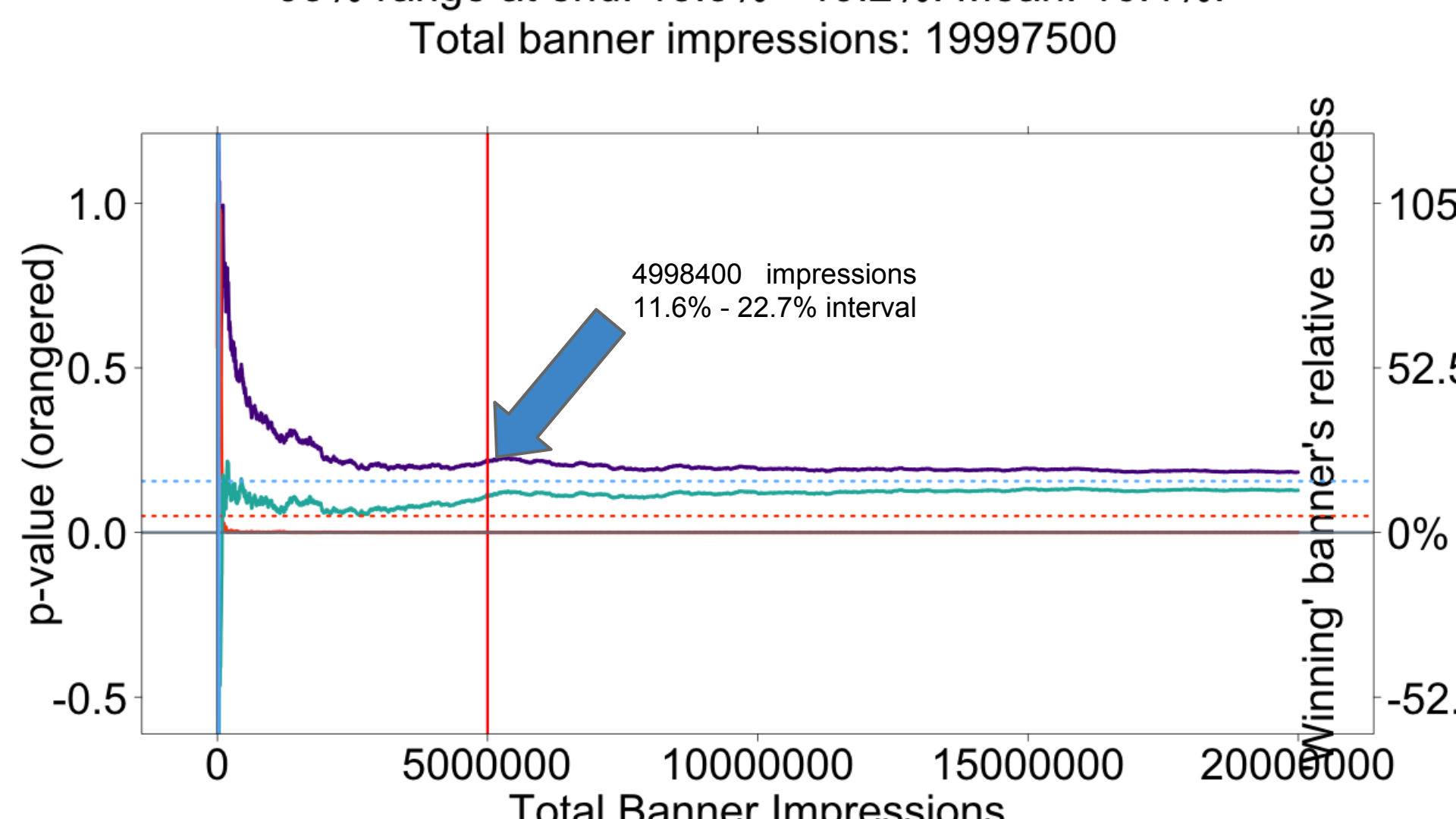

1.0
0.5
0.0
-0.5
p-value (orangered)
15%
7.5%
0%
-7.5%
'Winning' banner's relative success
0
5000000
10000000
15000000
20000000
25000000
Total Banner Impressions




This is the blip.
Impressions
4e+05
3e+05
2e+05
1e+05

To halve the confidence interval, you have to roughly **quadruple** the sample size!


Total banner impressions: 4998400
p-value (orangered)
1.0
0.5
0.0
-0.5
'Winning' banner's relative success
105
52.5
0%
-52.
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
Total Banner Impressions


Total banner impressions: 19997500
4998400 impressions
11.6% - 22.7% interval
p-value (orangered)
'Winning' banner's relative success
1.0
0.5
0.0
-0.5
105
52.5
0%
-52.
0
5000000
10000000
15000000
20000000
Total Banner Impressions

A/B Test (Split Test) Calculator

1 million

Label	Number of successes	Number of trials	
Baseline	2000	1000000	
Variation 1	2050	1000000	Remove

Interval confidence level: 0.95 ☑ Use multiple testing correction (recommended)

Compute Add another group

	Successes	Total	Success Rate	p-value	Improvement
Baseline	2,000	1,000,000	0.19% – 0.21% (0.2%)	—	—
Variation 1	2,050	1,000,000	0.2% – 0.21% (0.21%)	0.43	-3.7% – 8.7% (2.5%)

A/B Test (Split Test) Calculator

7 million!

Label	Number of successes	Number of trials	
Baseline	14000	7000000	
Variation 1	14350	7000000	Remove

Interval confidence level: 0.95 ☑ Use multiple testing correction (recommended)

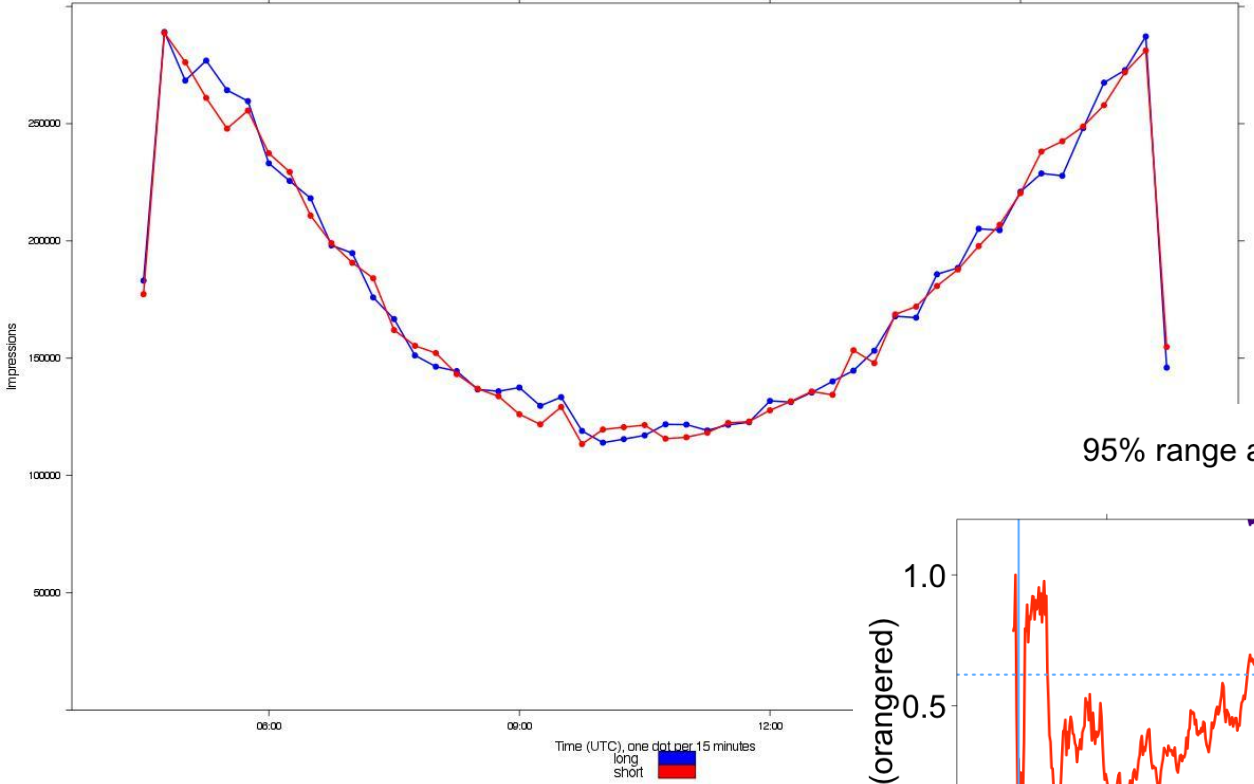
Compute | Add another group

	Successes	Total	Success Rate	p-value	Improvement
Baseline	14,000	7,000,000	0.2% – 0.2% (0.2%)	—	—
Variation 1	14,350	7,000,000	0.2% – 0.21% (0.21%)	0.037	0.15% – 4.9% (2.5%)

Another tip:

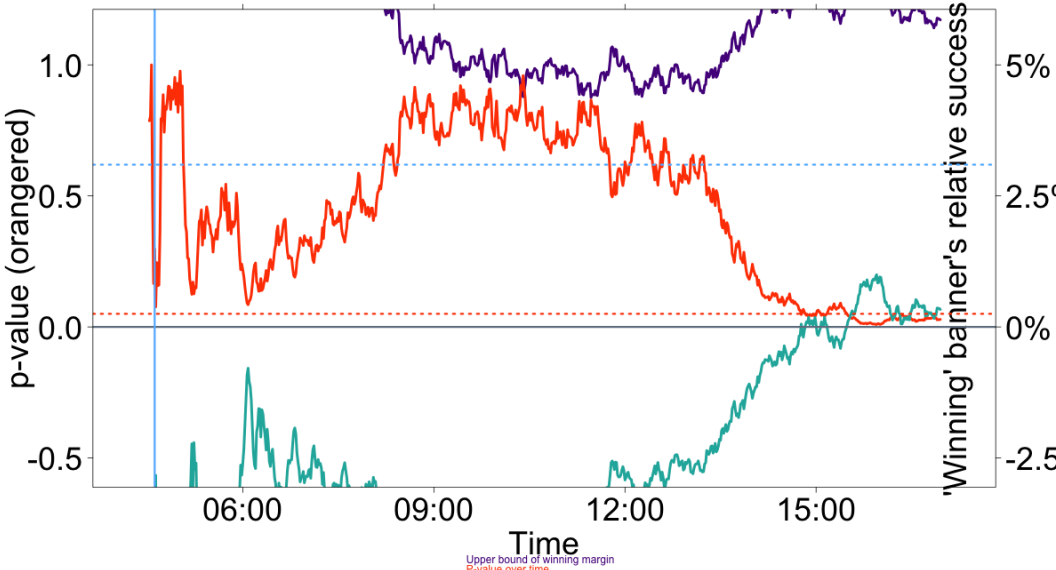
WFRs (Wildly Fluctuating Response rates) can mess you up.

Example - WMF donation rates at night are much lower than during the day, and skew our results.



Winner: A

95% range at end: 0.3% - 5.9%. Mean: 3.1%.

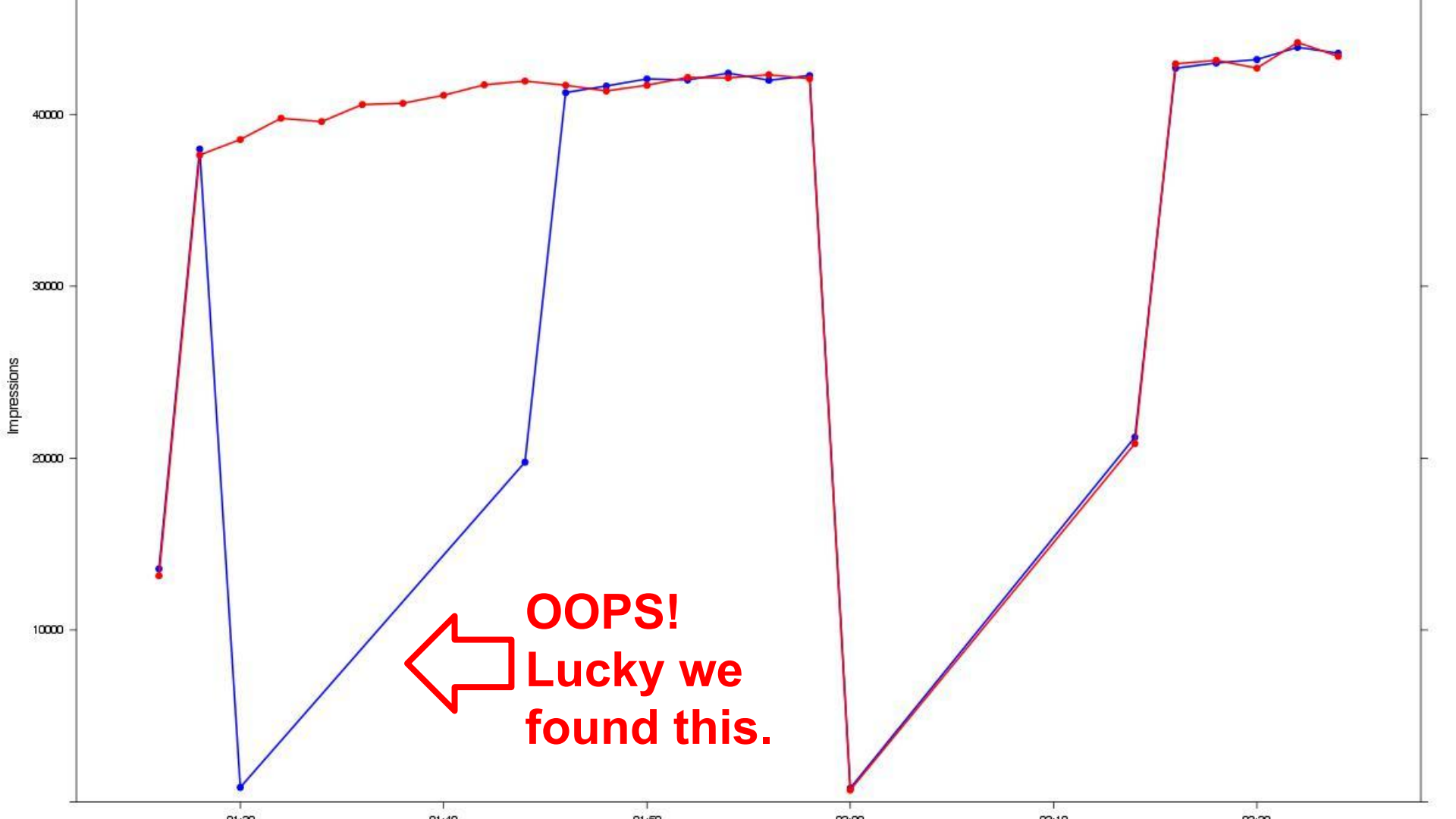


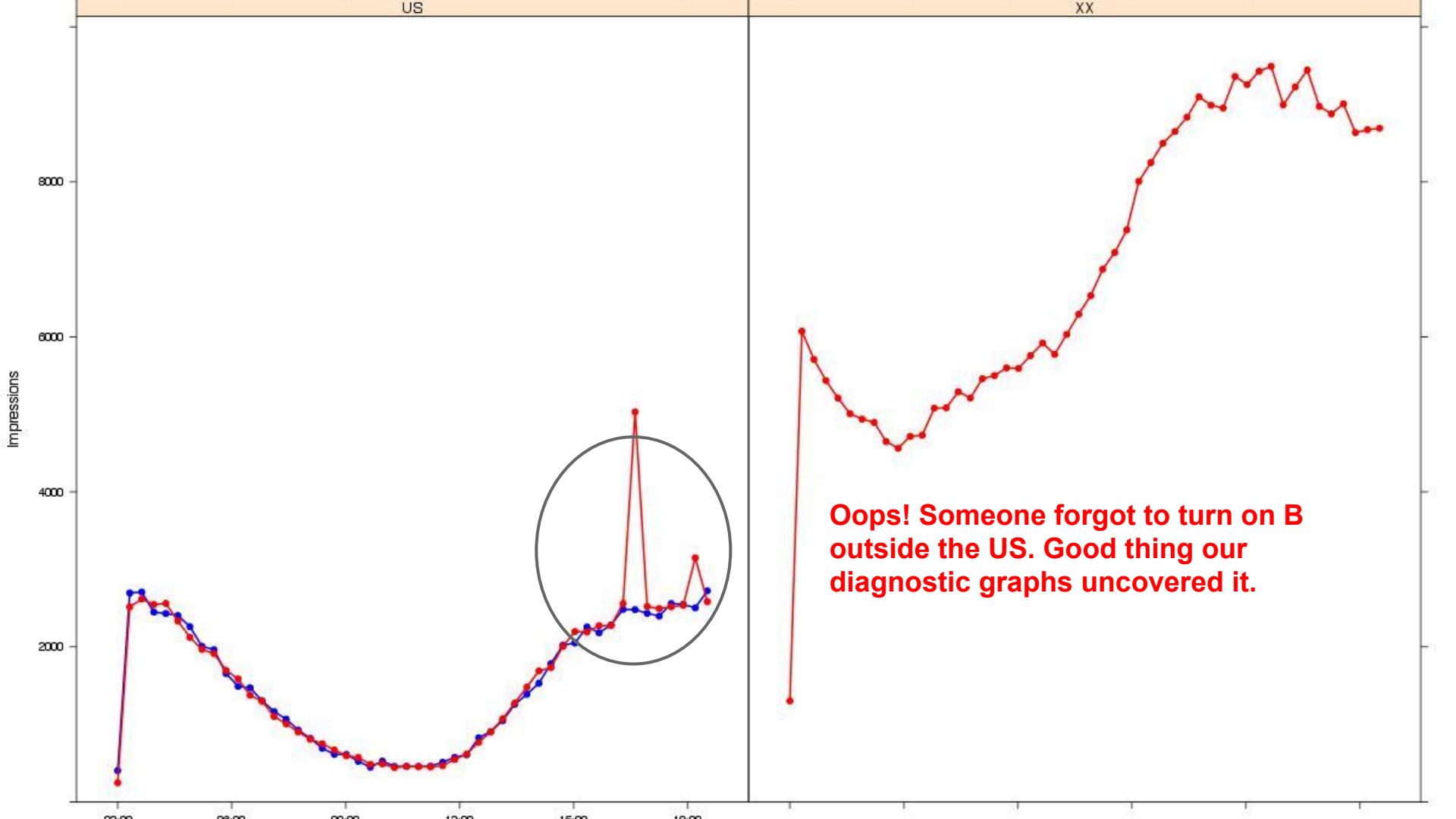
Some good news, if you’re torn between Agresti-Coull and Adjusted Wald...

Any stats test will do.

```
> binom.confint(9000, 1000000, conf.level=.95,
methods=c("agresti-coull", "asymptotic", "cloglog", "exact",
"logit", "probit", "profile", "lrt", "prop.test", "wilson"))
          method    x      n  mean        lower        upper
1  agresti-coull 9000 1e+06 0.009 0.008816767 0.009187005
2     asymptotic 9000 1e+06 0.009 0.008814900 0.009185100
3        cloglog 9000 1e+06 0.009 0.008816394 0.009186605
4          exact 9000 1e+06 0.009 0.008815823 0.009187029
5          logit 9000 1e+06 0.009 0.008816774 0.009186999
6         probit 9000 1e+06 0.009 0.008816559 0.009186775
7        profile 9000 1e+06 0.009 0.008816156 0.009186359
8            lrt 9000 1e+06 0.009 0.008839571 0.009187291
9     prop.test 9000 1e+06 0.009 0.008816282 0.009187500
10       wilson 9000 1e+06 0.009 0.008816777 0.009186995
```

Use diagnostic graphs to detect errors in your testing.

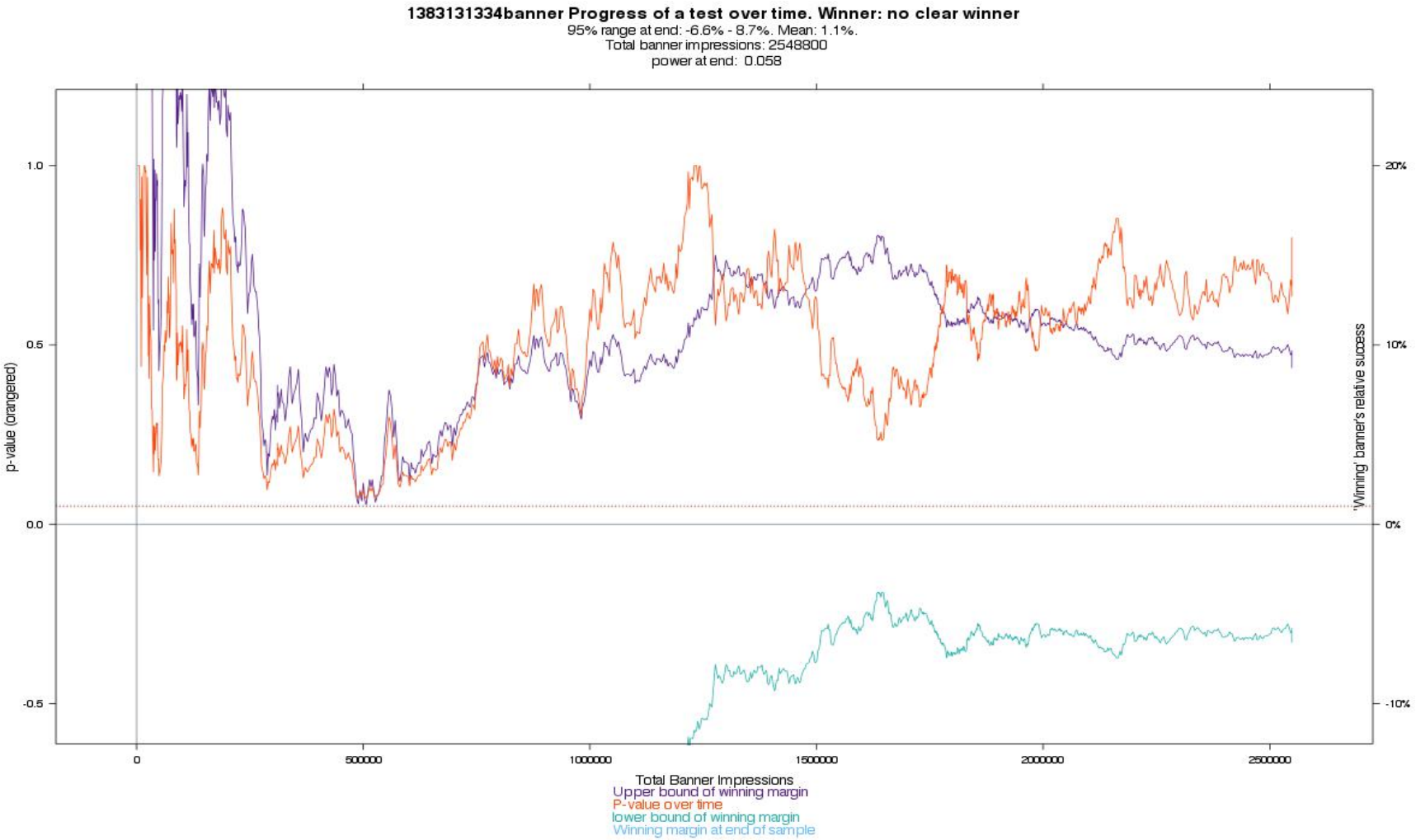

Impressions
40000
30000
20000
10000
OOPS!
Lucky we
found this.


US
XX
8000
6000
4000
2000
Impressions
Oops! Someone forgot to turn on B outside the US. Good thing our diagnostic graphs uncovered it.

Let business needs, not stats dogma decide when to stop your tests.

Is B going to be technically or politically difficult to implement permanently, but is winning by 5% to 50%? Then you need to keep running your test!

Are A and B almost the same to your business? And B is 0% to 8% better? Then stop! Time to test something else and find something with a bigger impact!


1383131334banner Progress of a test over time. Winner: no clear winner
95% range at end: -6.6% - 8.7%. Mean: 1.1%.
Total banner impressions: 2548800
power at end: 0.058
p-value (orangered)
1.0
0.5
0.0
-0.5
'Winning' banner's relative success
20%
10%
0%
-10%
0
500000
1000000
1500000
2000000
2500000
Total Banner Impressions
Upper bound of winning margin
P-value over time
lower bound of winning margin
Winning margin at end of sample

Announcement:

All of our code is free/libre software. We'd love collaborators.

zack@thoughtworks.com
sahar@vlwc.org

Review:

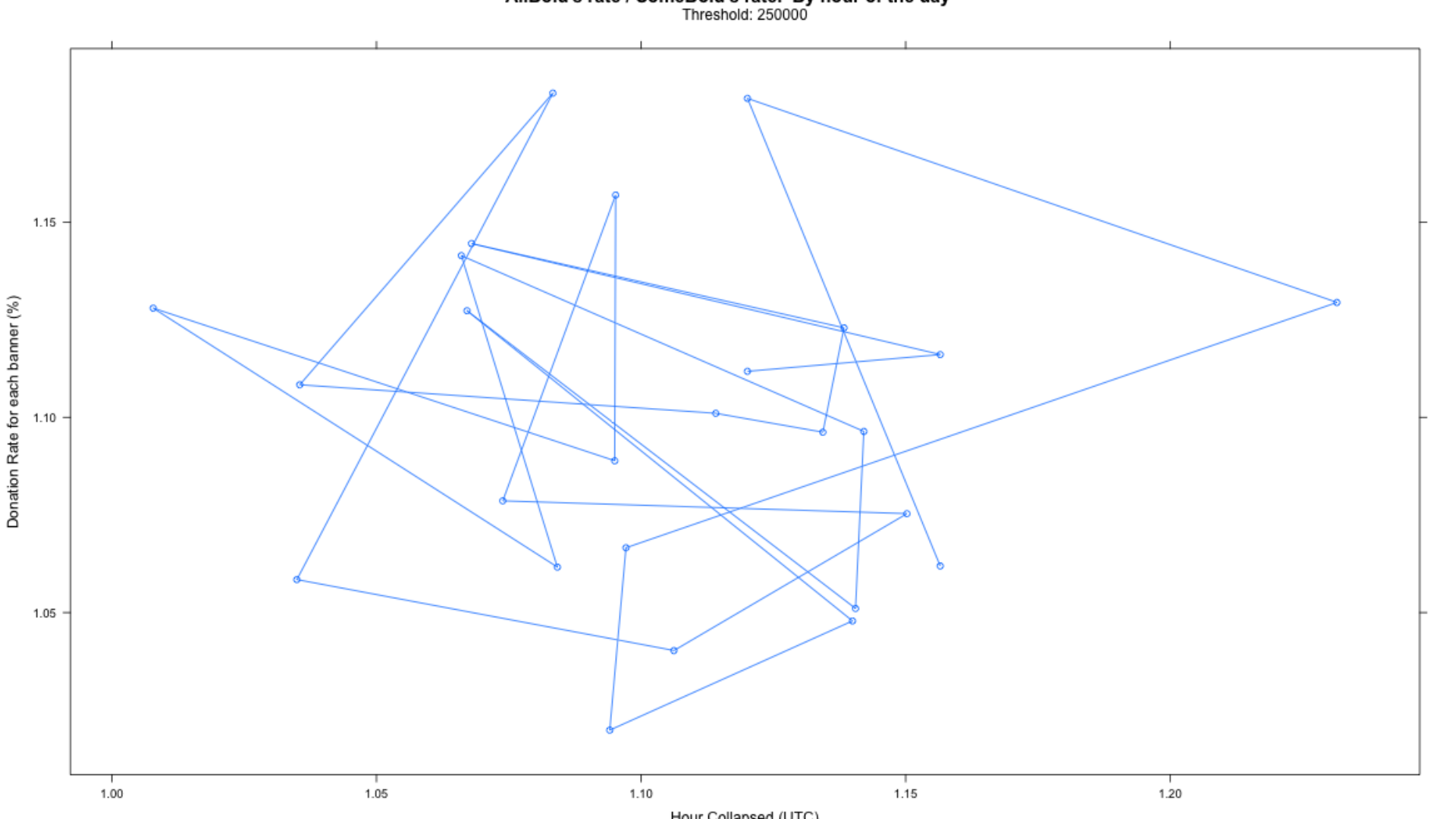
- There’s nothing magic about 95% confidence - consider using 70% or 80%.
- Decide when to end your test *dynamically,* don’t fix your sample size ahead of time. It’s totally okay to peek.
- Confidence intervals are your new best friend.
- The lower bound of your confidence interval will be > 0 when you have confidence. (When p-value is below the threshold).
- Don’t freak out if p-value spikes a bit - look at your confidence interval: is it an edge case?
- If A & B are very slightly different, you’ll need an enormous sample size to find it - it’s not worth it!
- Quadruple your sample size to halve your confidence interval.
- Wait until A & B have 15 successes each. &/or run power prop over and over.
- Beware of low response rate periods.
- Almost any statistical test for finding p/confidence is fine.
- Use diagnostic graphs to detect errors.

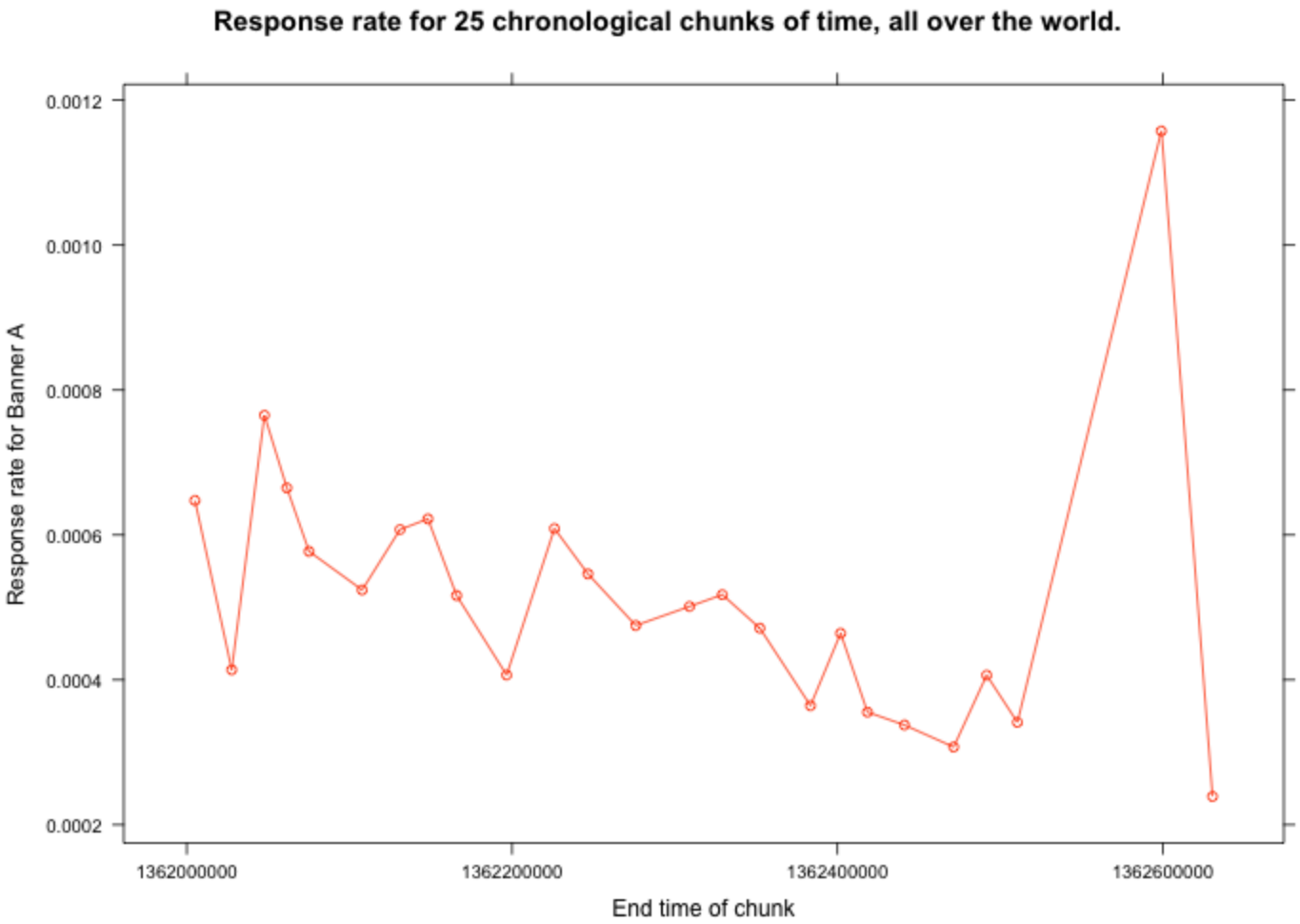
Extra slides in case we have enough time:

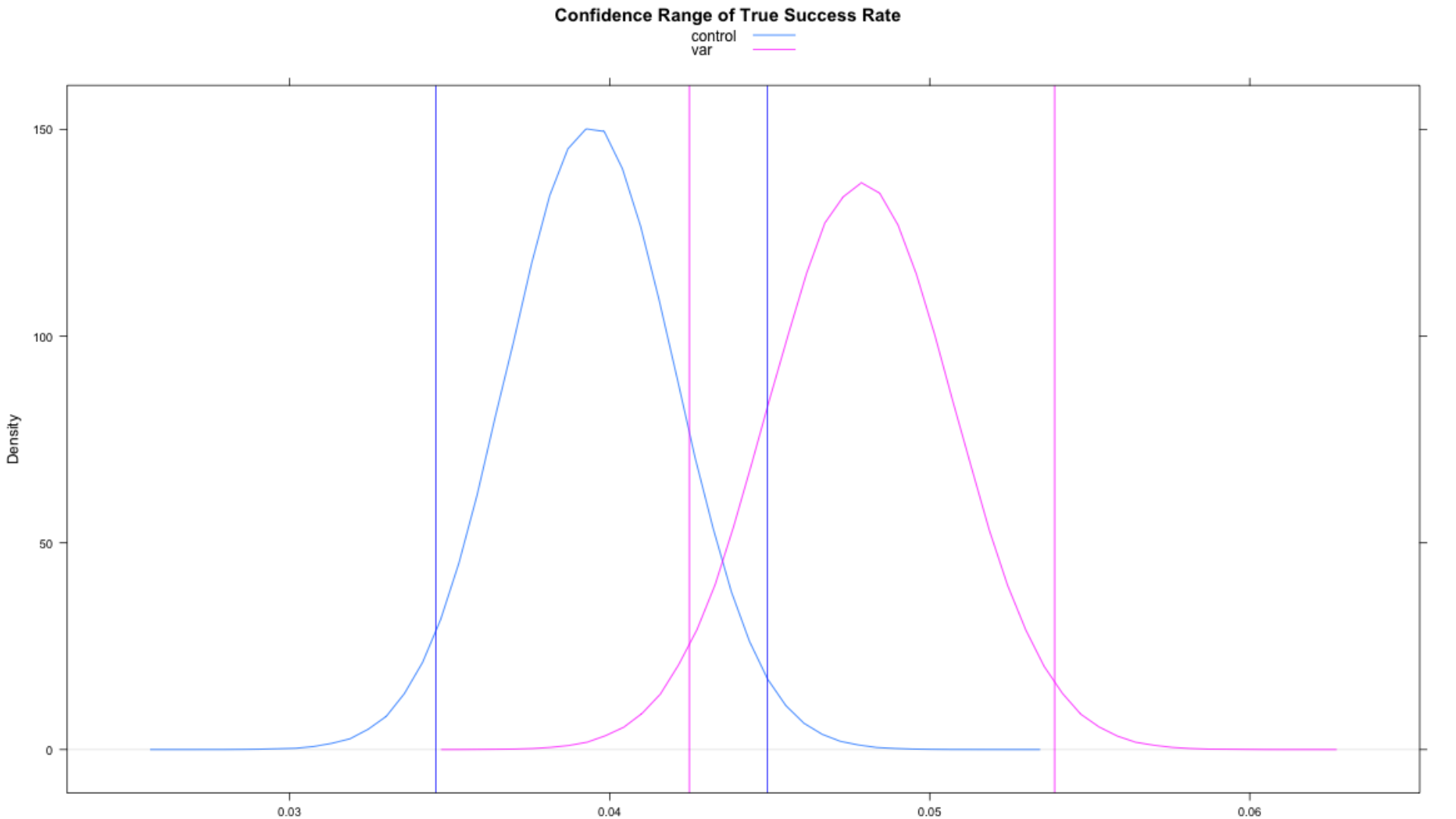
Our back up method:

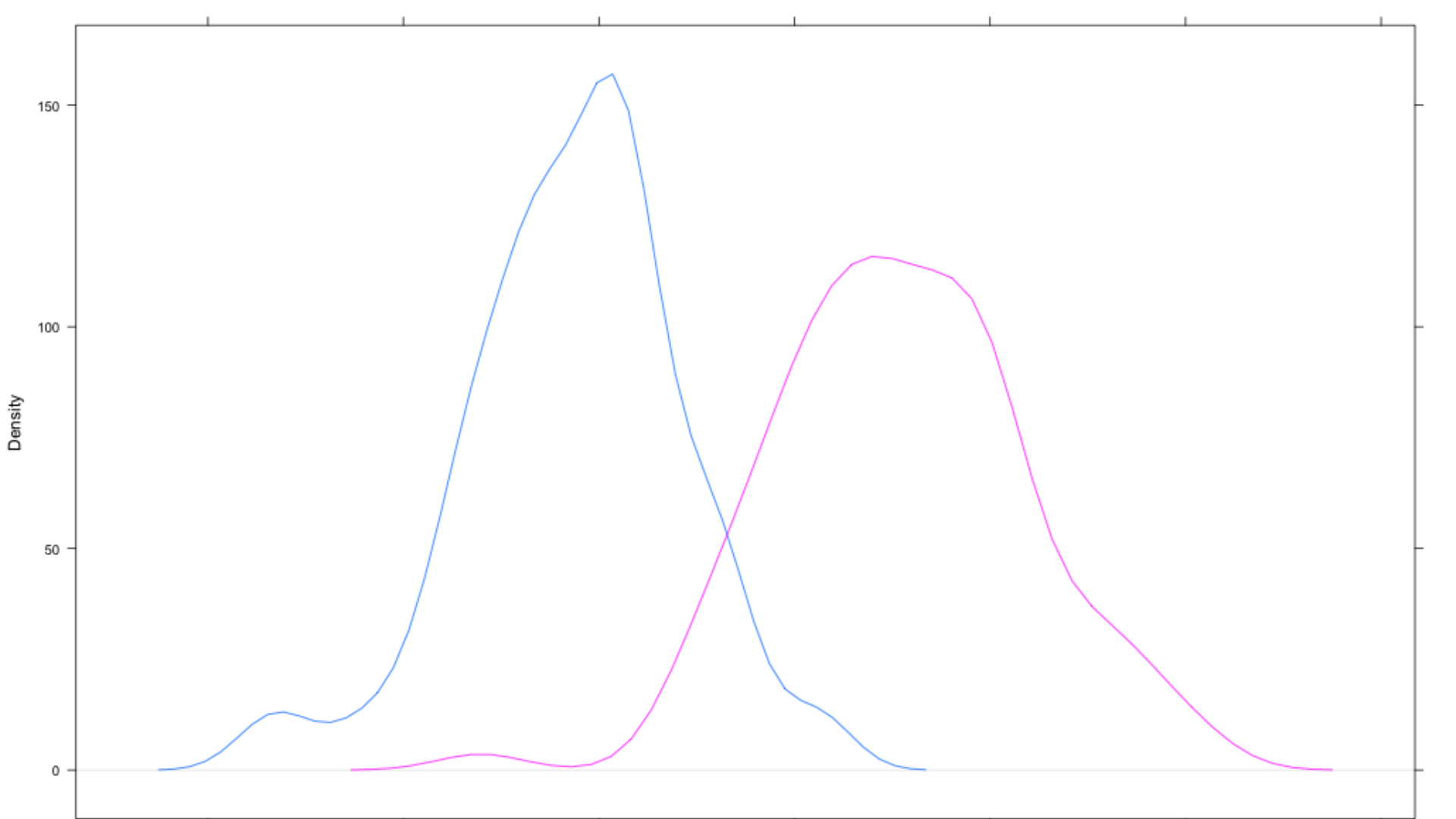
We use the power prop test in a sort of self-referential way. We continuously run power prop using the proportions we have at the moment and see if our sample is the recommended size.

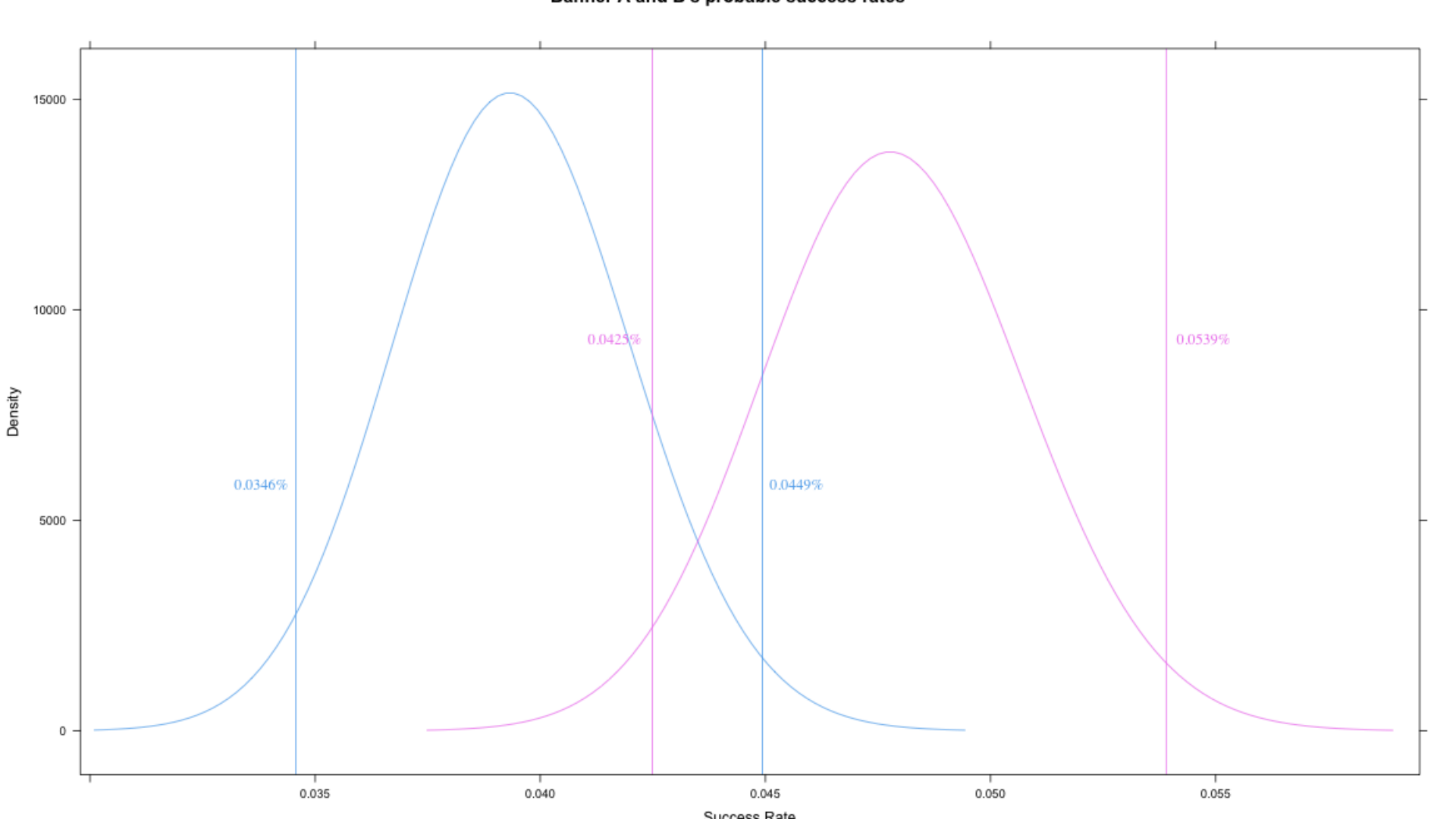
```
power.prop.test(p1=p1, p2=p2,
power=power, sig.level=alpha)$n
```

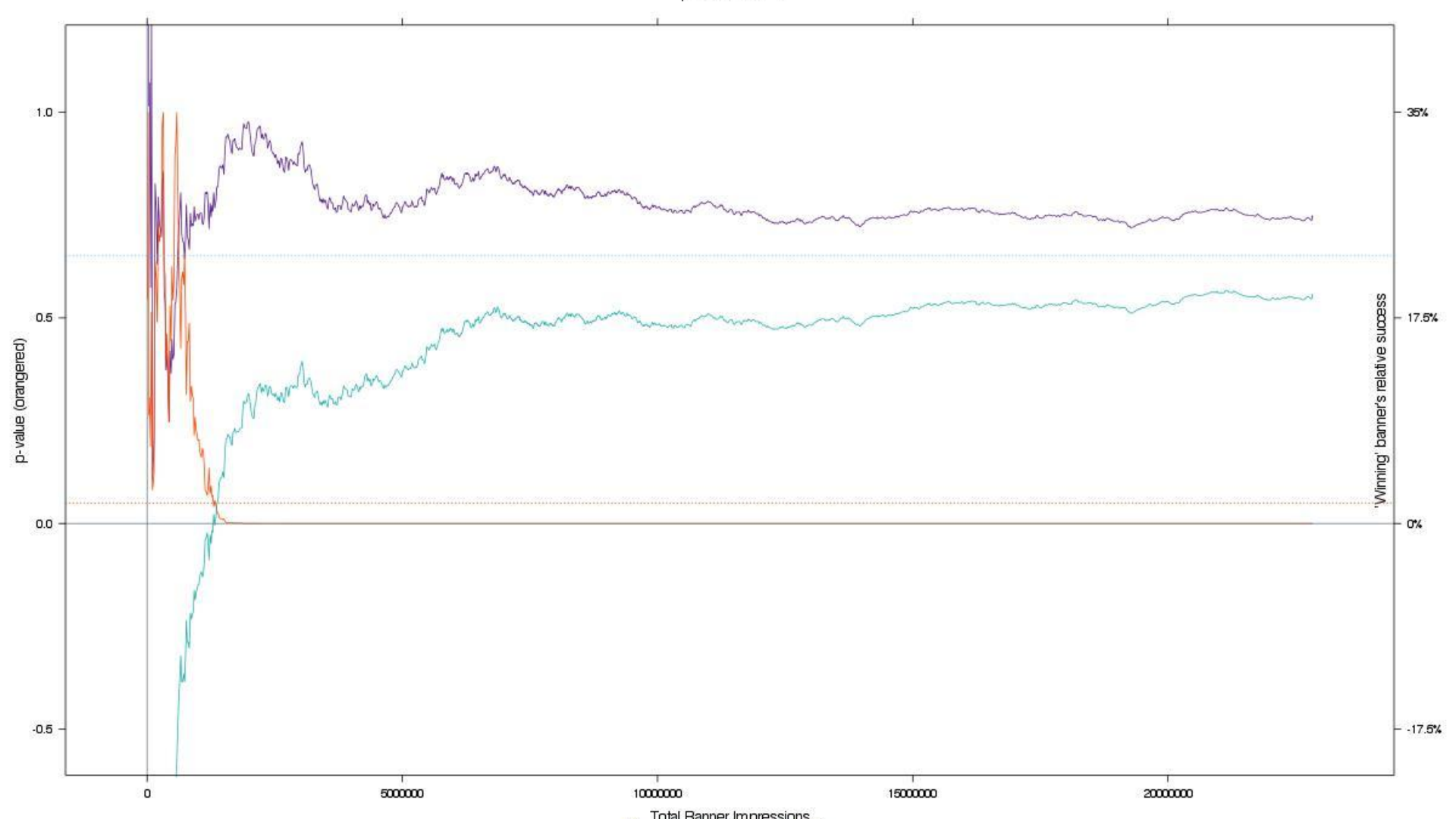

Threshold: 250000
Donation Rate for each banner (%)
1.15
1.10
1.05
1.00
1.05
1.10
1.15
1.20

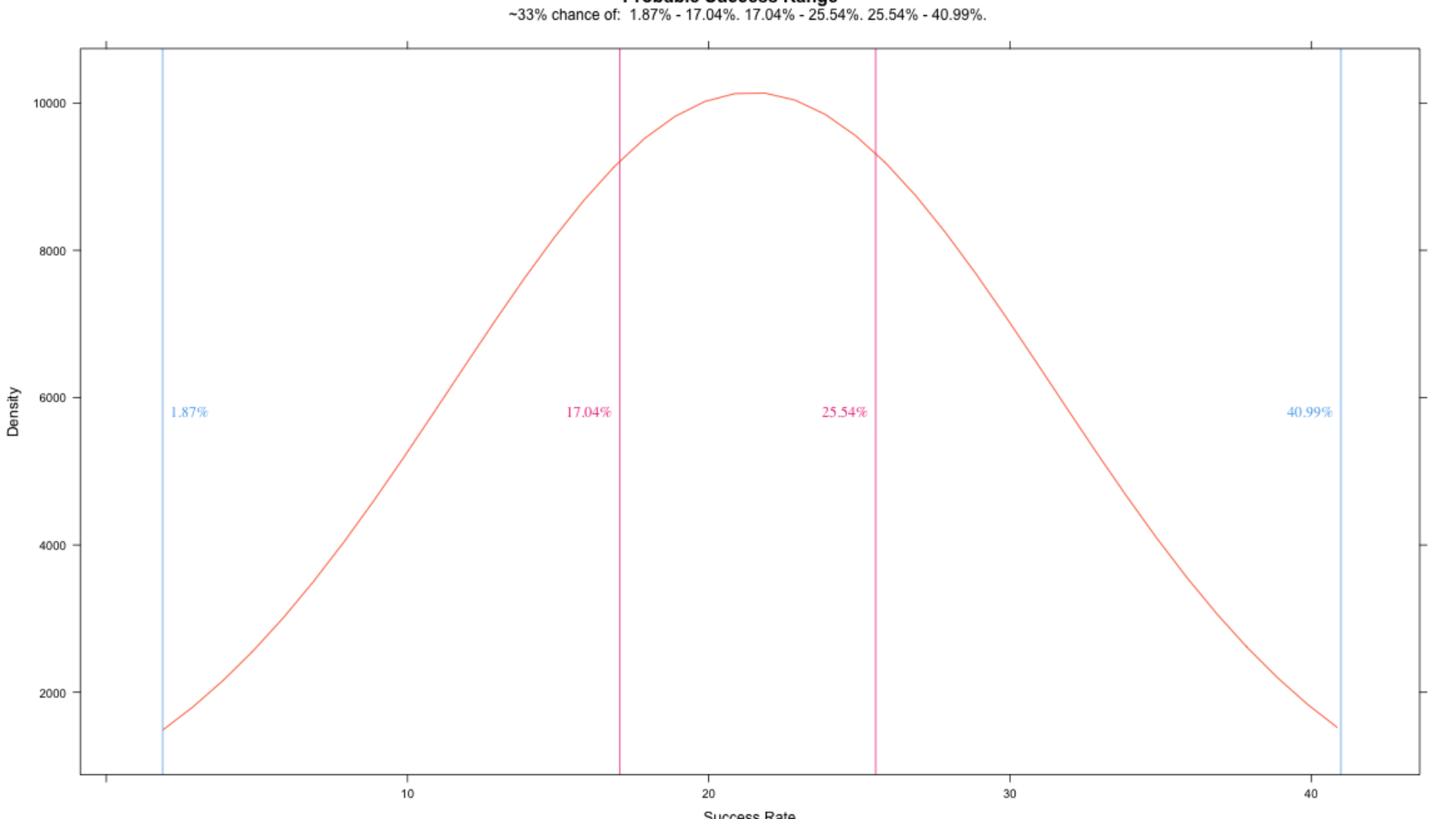

Response rate for 25 chronological chunks of time, all over the world.
Response rate for Banner A
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
1362000000
1362200000
1362400000
1362600000
End time of chunk


Confidence Range of True Success Rate
control
var
Density
150
100
50
0
0.03
0.04
0.05
0.06


150
100
50
0
Density


Banner A and B's probable success rates
15000
10000
5000
0
Density
0.0346%
0.0425%
0.0449%
0.0539%
0.035
0.040
0.045
0.050
0.055
Success Rate


1.0
0.5
0.0
-0.5
p-value (orangered)
35%
17.5%
0%
-17.5%
'Winning' banner's relative success
0
5000000
10000000
15000000
20000000
Total Banner Impressions

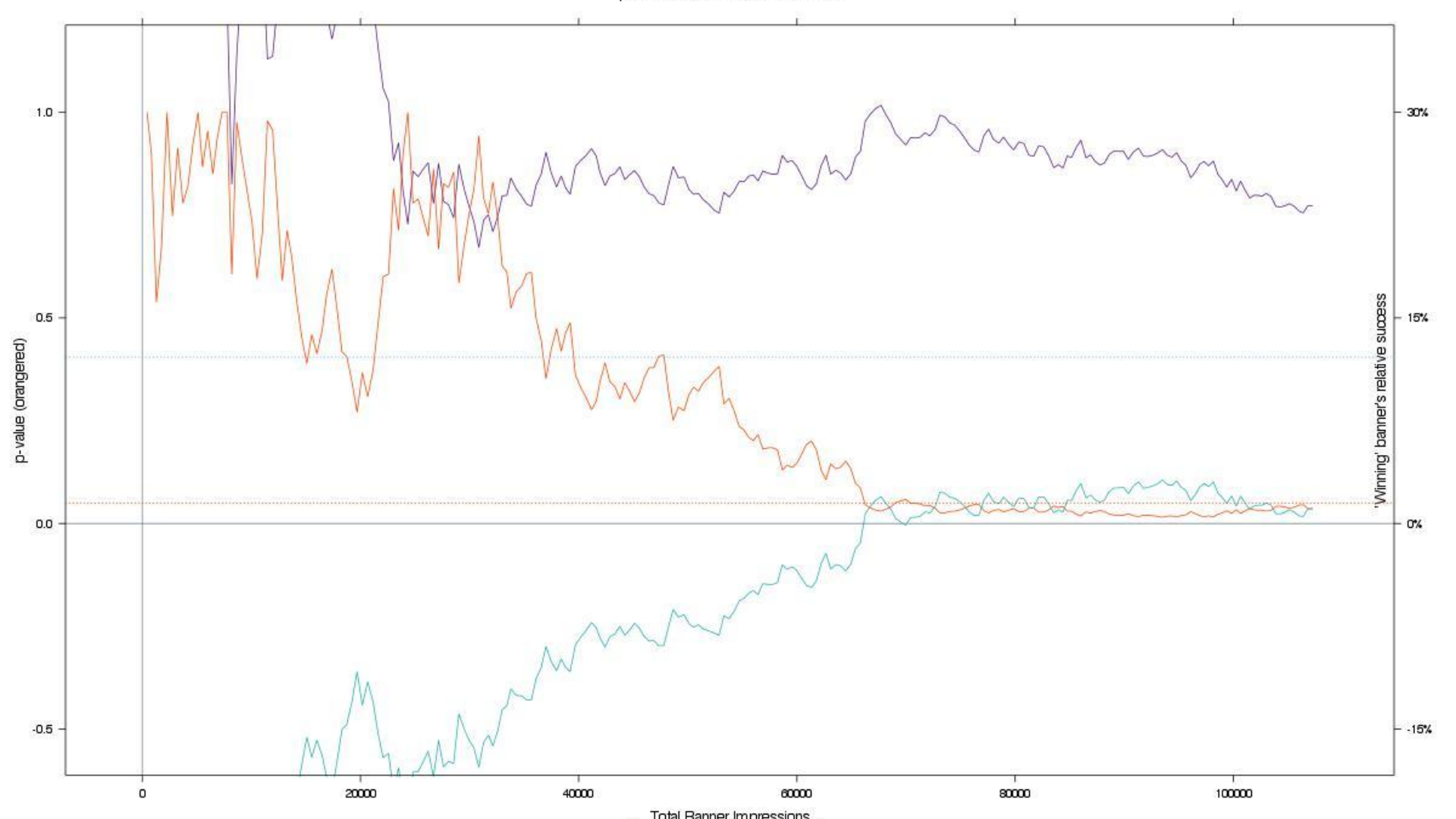

Probable Success Range
~33% chance of: 1.87% - 17.04%. 17.04% - 25.54%. 25.54% - 40.99%.
Density
10000
8000
6000
4000
2000
1.87%
17.04%
25.54%
40.99%
10
20
30
40
Success Rate

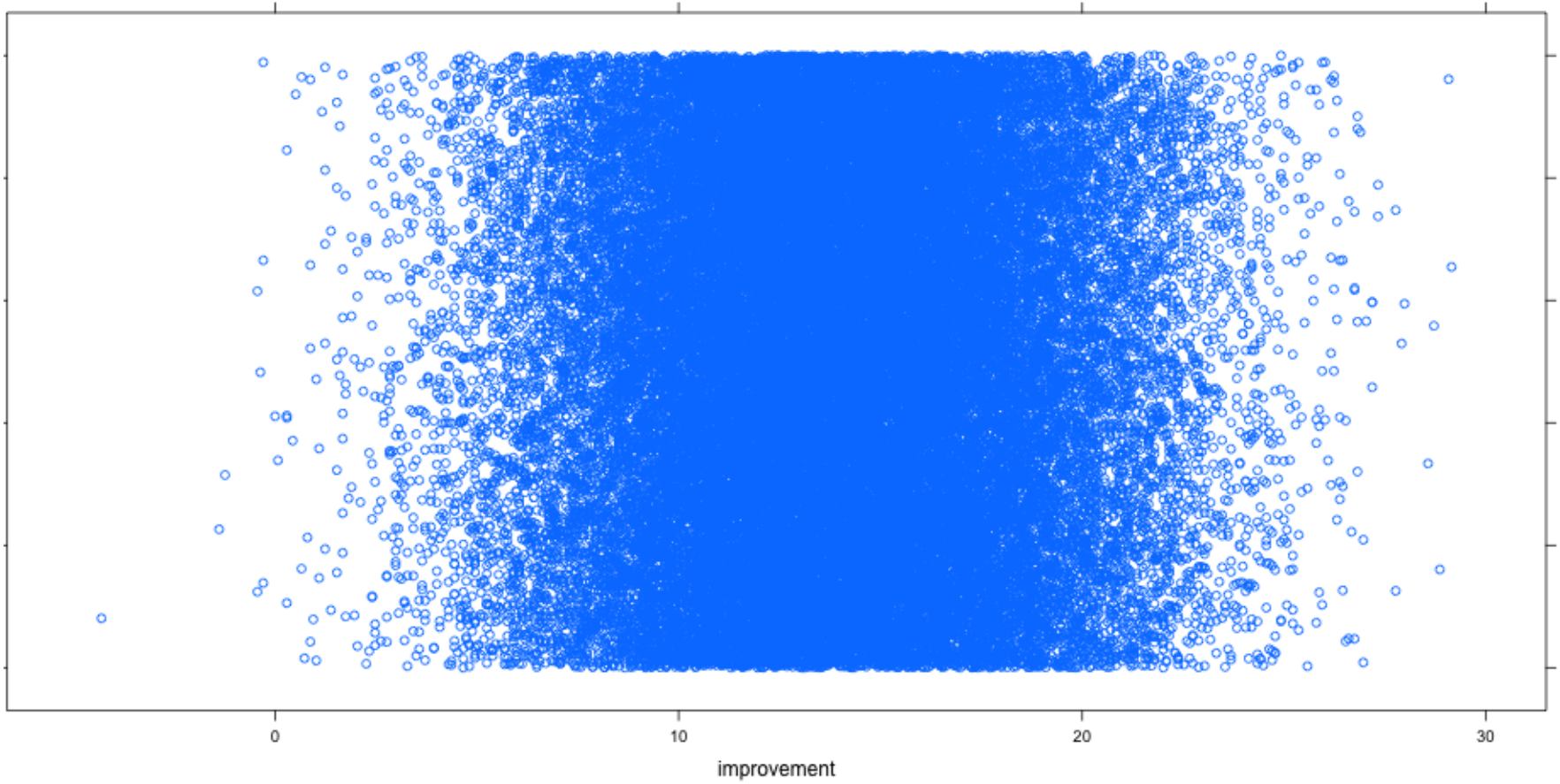
Yes, Zack, you really can trust all these standard statistical tests. They do apply to AB testing on websites too.

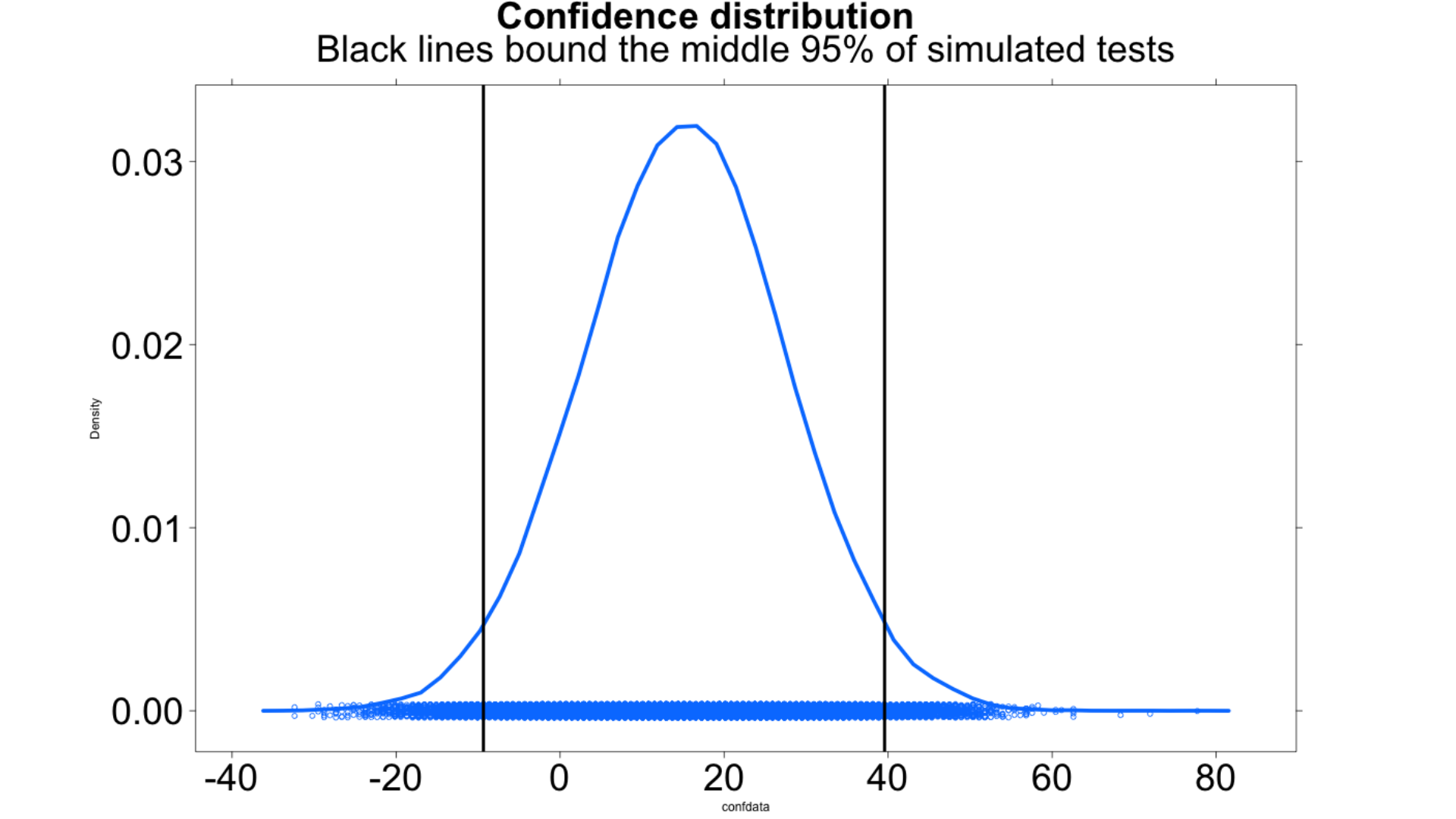
Trust p. Trust confidence intervals.

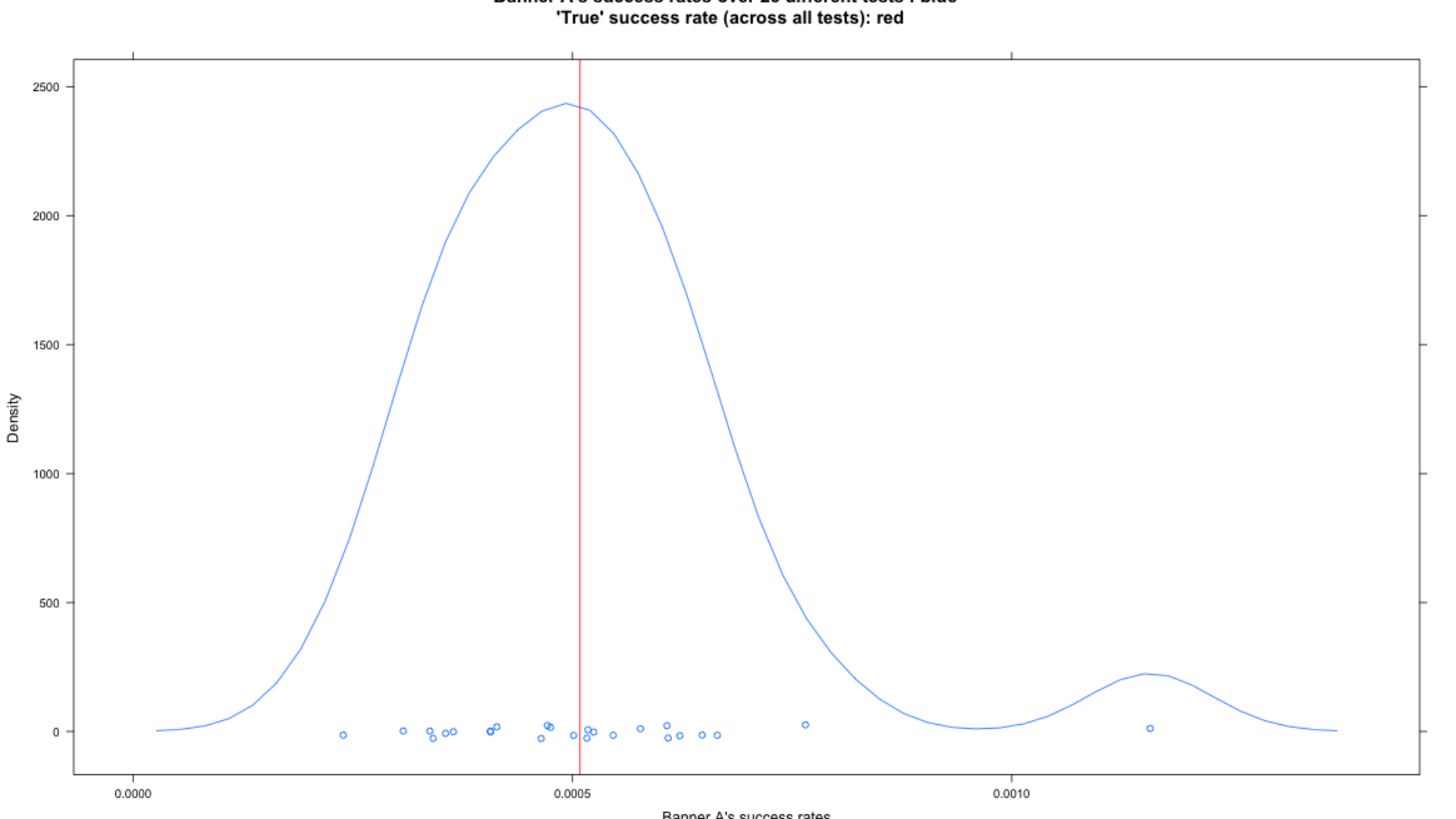
Wide confidence intervals and p values that never get to .05 are signals to move on to a new test. But don’t ignore the results just because you didn’t “get confidence.”

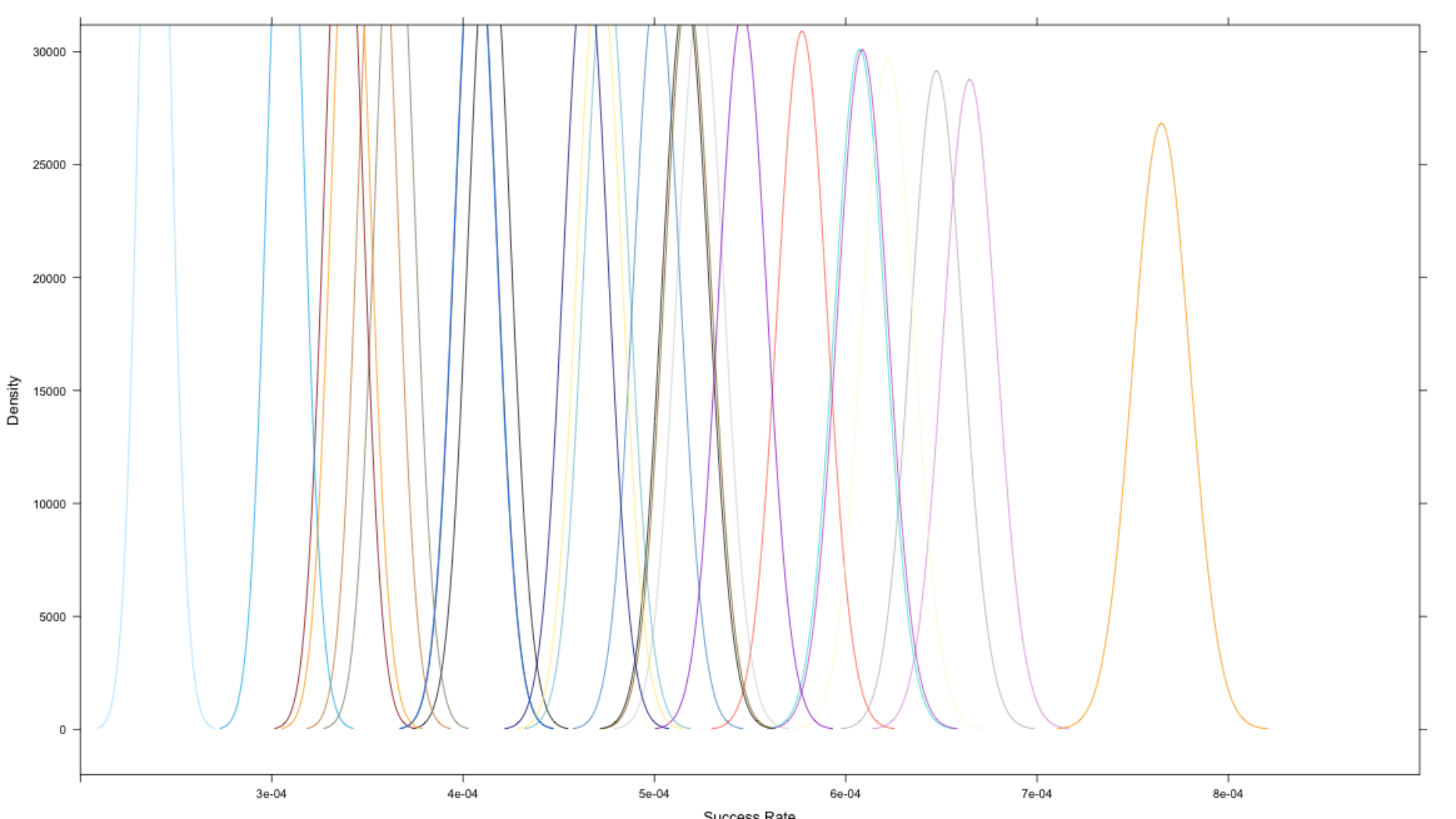
Most AB testing mistakes are caused by stupid errors in your own data or testing, not stats. Make diagnostic visualizations to spot problems in your underlying data that could be causing misleading tests.


1.0
0.5
0.0
-0.5
p-value (orangered)
30%
15%
0%
-15%
'Winning' banner's relative success
0
20000
40000
60000
80000
100000
Total Banner Impressions


0
10
20
30
improvement


Confidence distribution
Black lines bound the middle 95% of simulated tests
0.03
0.02
0.01
0.00
Density
-40
-20
0
20
40
60
80
confdata


'True' success rate (across all tests): red
2500
2000
1500
1000
500
0
Density
0.0000
0.0005
0.0010


Density
30000
25000
20000
15000
10000
5000
0
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
Success Rate

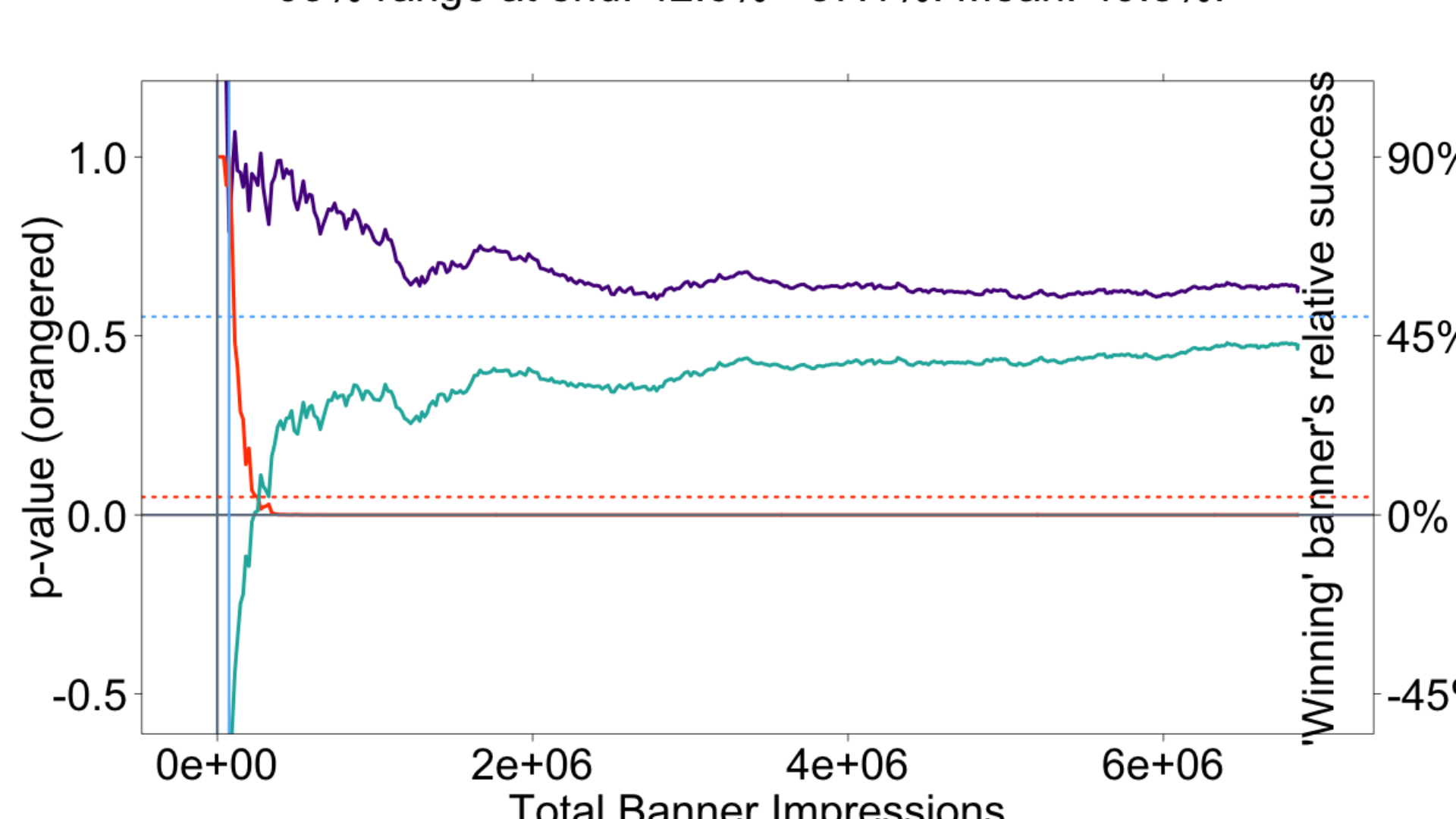
OK, everyone repeat after me...

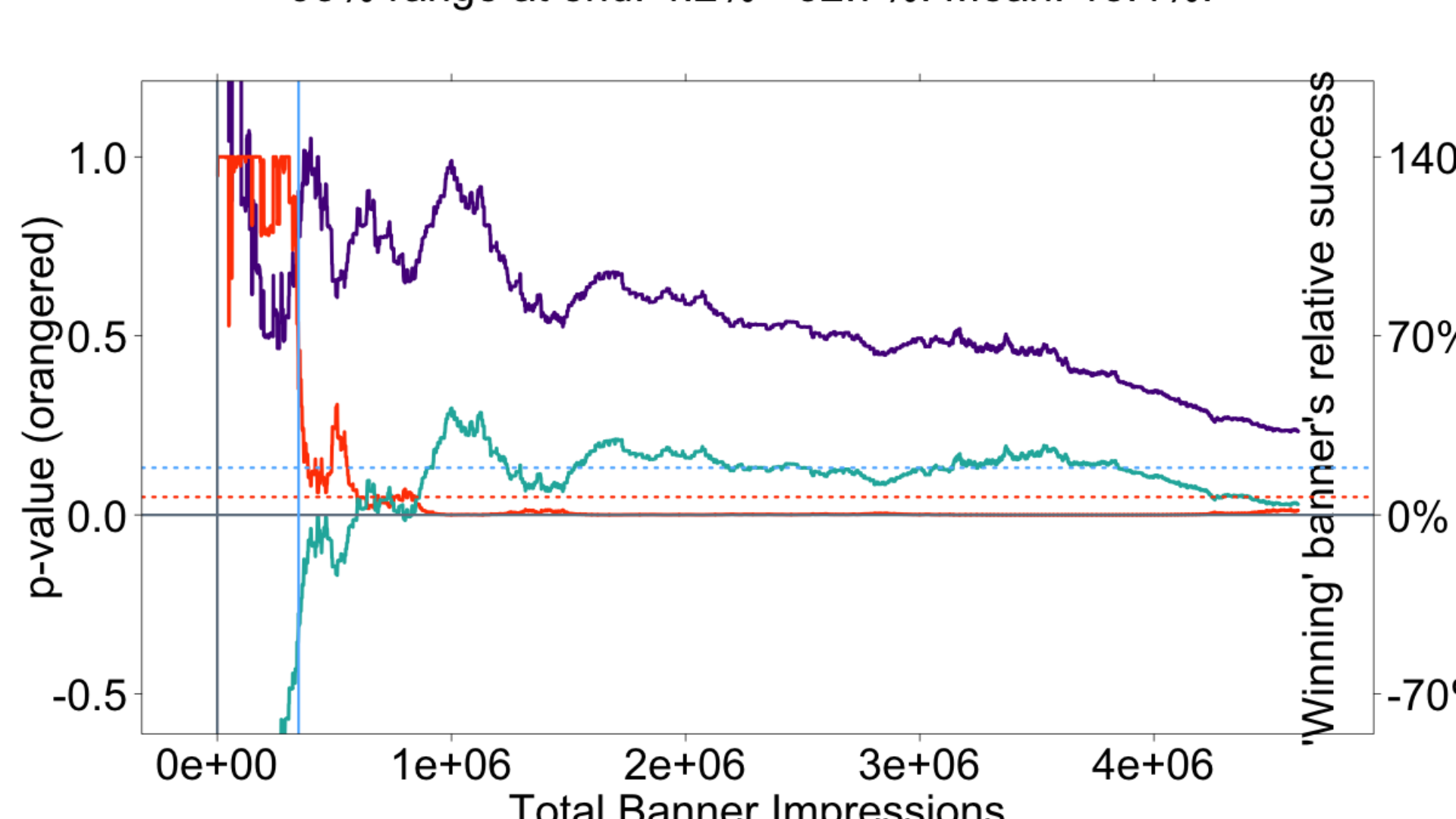
Not only is it OK to peek. You don’t even have to wait for 95% confidence!

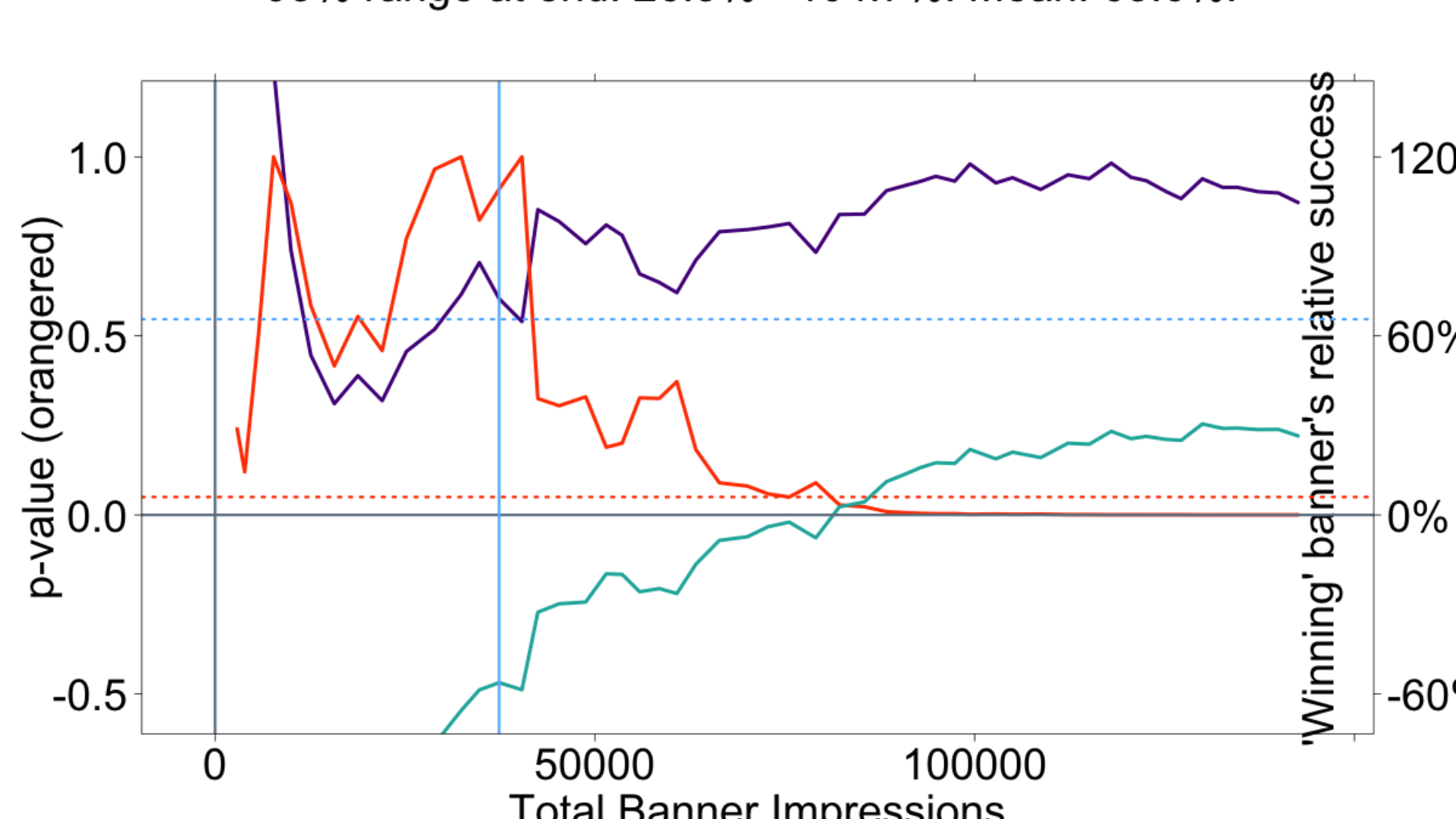
Caveat:

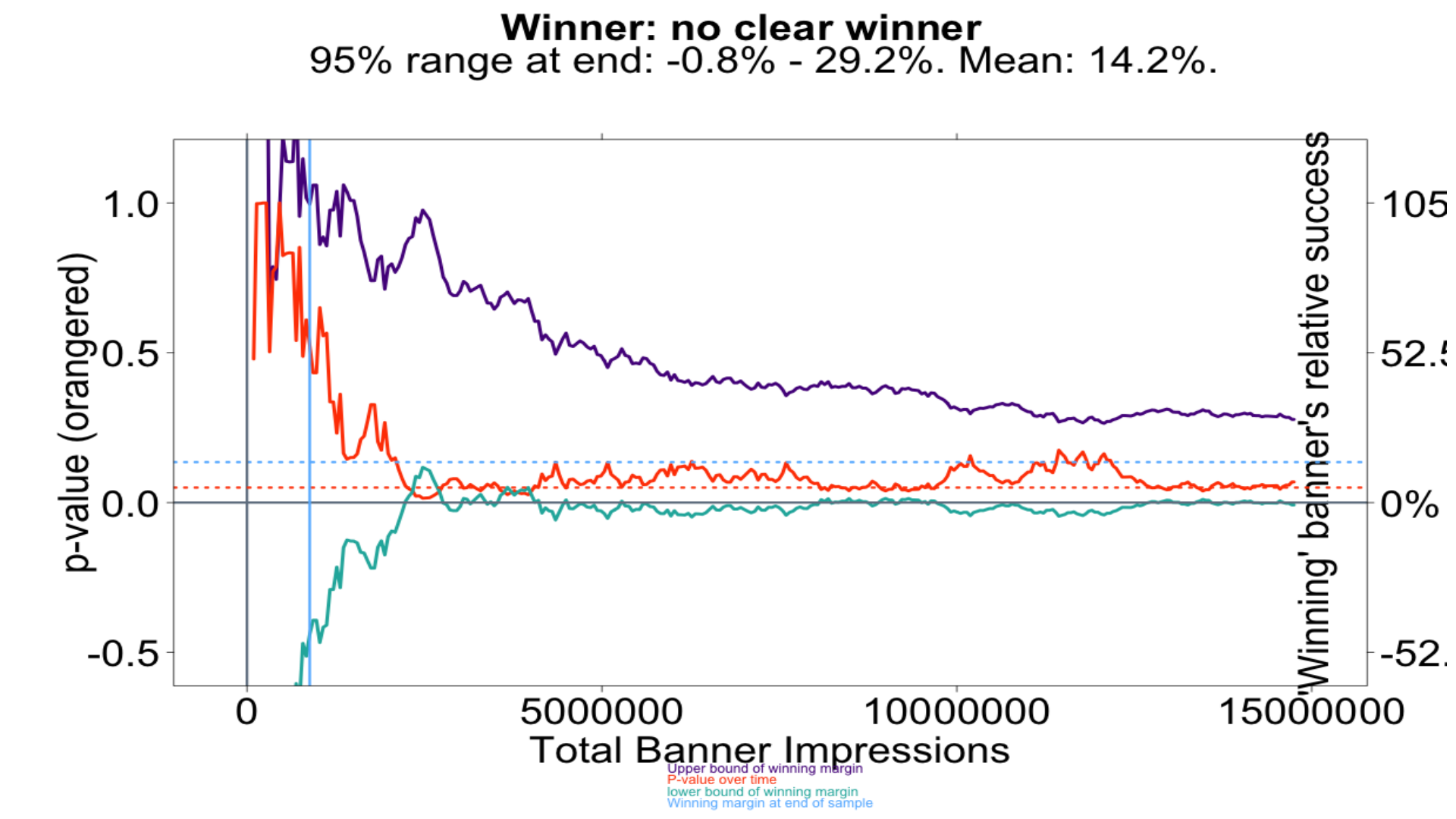
To get through the initial noise, wait until A & B have 15 successes each.
Then you can start peeking!

(There are other methods too.)


1.0
0.5
0.0
-0.5
p-value (orangered)
90%
45%
0%
-45%
'Winning' banner's relative success
0e+00
2e+06
4e+06
6e+06
Total Banner Impressions


1.0
0.5
0.0
-0.5
p-value (orangered)
140
70%
0%
-70
'Winning' banner's relative success
0e+00
1e+06
2e+06
3e+06
4e+06
Total Banner Impressions


1.0
0.5
0.0
-0.5
p-value (orangered)
120
60%
0%
-60
'Winning' banner's relative success
0
50000
100000
Total Banner Impressions


Winner: no clear winner
95% range at end: -0.8% - 29.2%. Mean: 14.2%.
p-value (orangered)
1.0
0.5
0.0
-0.5
105
52.5
0%
-52.
'Winning' banner's relative success
0
5000000
10000000
15000000
Total Banner Impressions
Upper bound of winning margin
P-value over time
lower bound of winning margin
Winning margin at end of sample

The “true” result is probably near the center of your confidence interval. Therefore, wide confidence intervals are not as useless as they might seem.

Out of 216 real Wikipedia tests we analyzed:

If we had stopped at 70% confidence (with our conservative methods of knowing when to stop):

We'd pick the winner : 93% of the time
We'd miss the winner: 5% of the time
We'd falsely find a difference: 2% of the time.
We'd pick the loser: 0% of the time.

Our total test time would be 27% of the time it'd take at 95% confidence.