We will resume at 10:07ish

Stats section outline

A/B Testing basics

Hypothesis testing: Null and alternative hypotheses

Statistical Significance and P-values

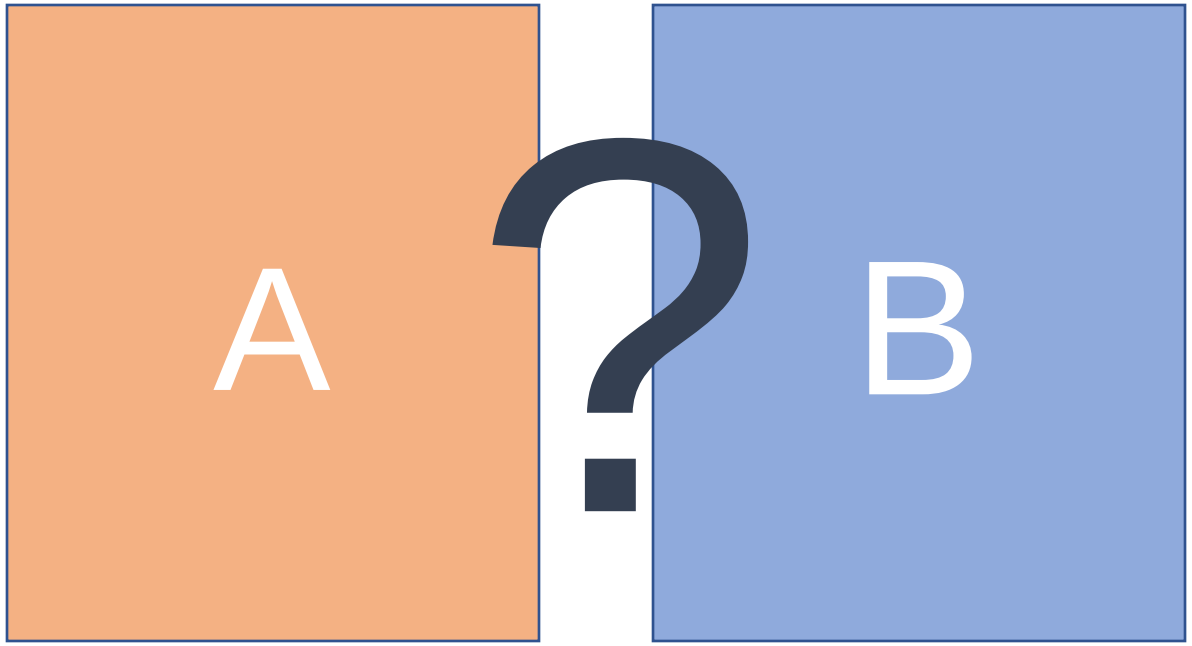
False positives and False negatives

Power

(poor) Alternatives to A/B Testing

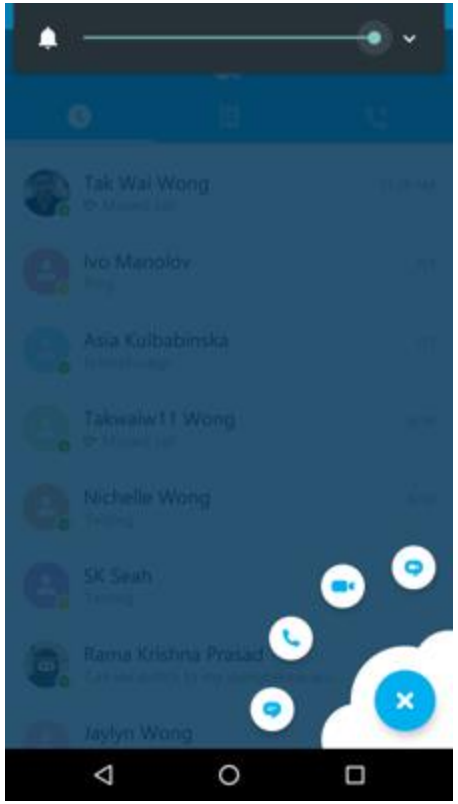
A/B Testing – Basic concepts

We want to establish which variant is better, A or B.



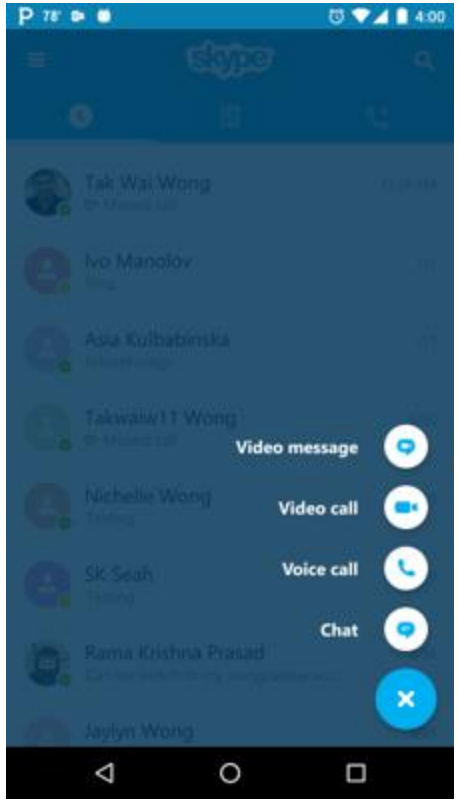
Menu on Android

A



Icons

B



Icons and names

Which variant has higher number of calls made per user?

Motivation

- You will examine scorecards for your experiments. Many of them!

	Treatment	Control	Delta	Delta %	P-Value
-- FeaturedPrimary_1 Clicks / UU	0.1389	0.1108	0.0281	+25.37%	0

- Understanding **the meaning** of the values on the scorecard will help you make more informed (and more correct) decisions.

A/B Testing – Basic concepts

How can we establish which variant is better?

Ideally, we would want to create two alternative universes.

- In the first universe, we show all people variant **A**,
- in the second universe we show all people variant **B**.
- We would then measure any **difference between the universes**.

This would ensure that any difference we observe was **caused by the treatment**, and not by extraneous effects.

A/B Testing – Basic concepts

Unfortunately, creating alternative universes is hard! So instead:

We **randomly split our users** such that some percentage experience variant **A**, and some percentage experience variant **B**.

At the end of the experiment, we measure the difference between the 2 groups.

A/B Testing – Basic concepts

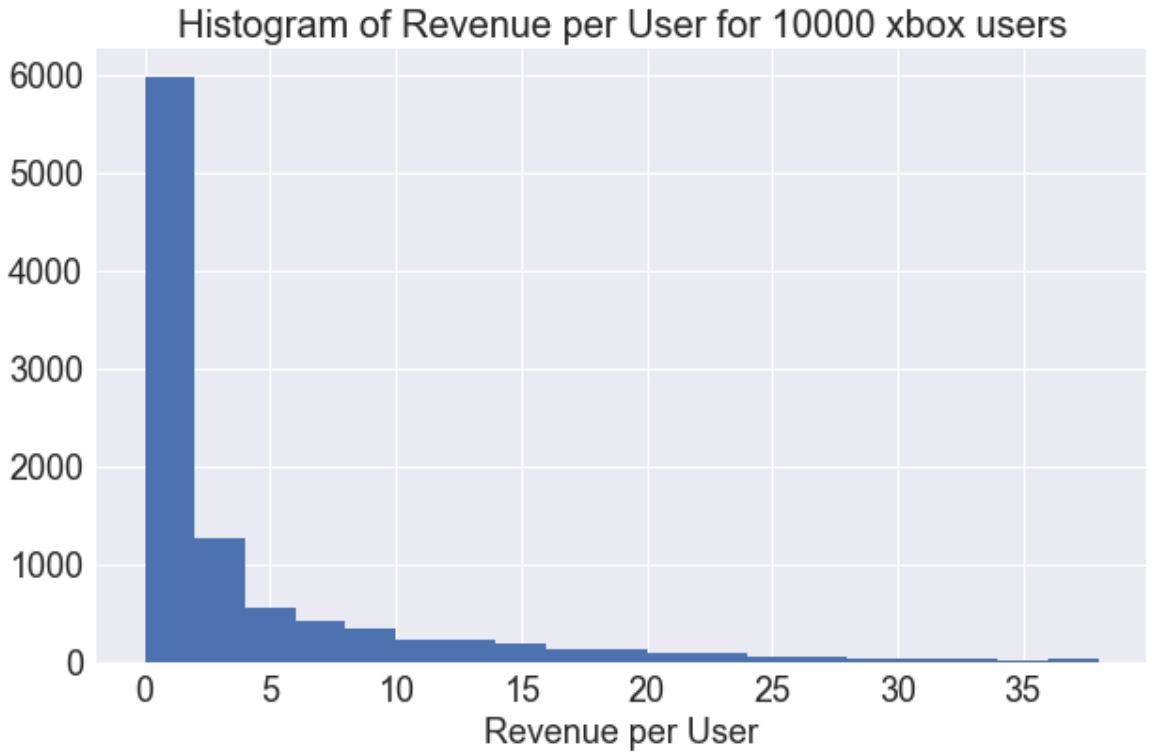
We randomly split **our original dataset** in half.

One half will experience variant “**A**” and the other half variant “**B**”.

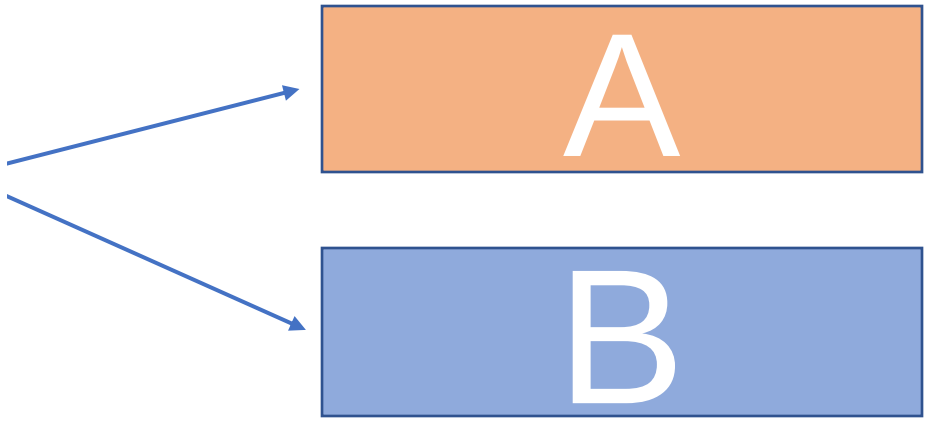
We then let the experiment run.

At the end of the experiment, we measure the difference in the means (“delta”) between treatment and control.

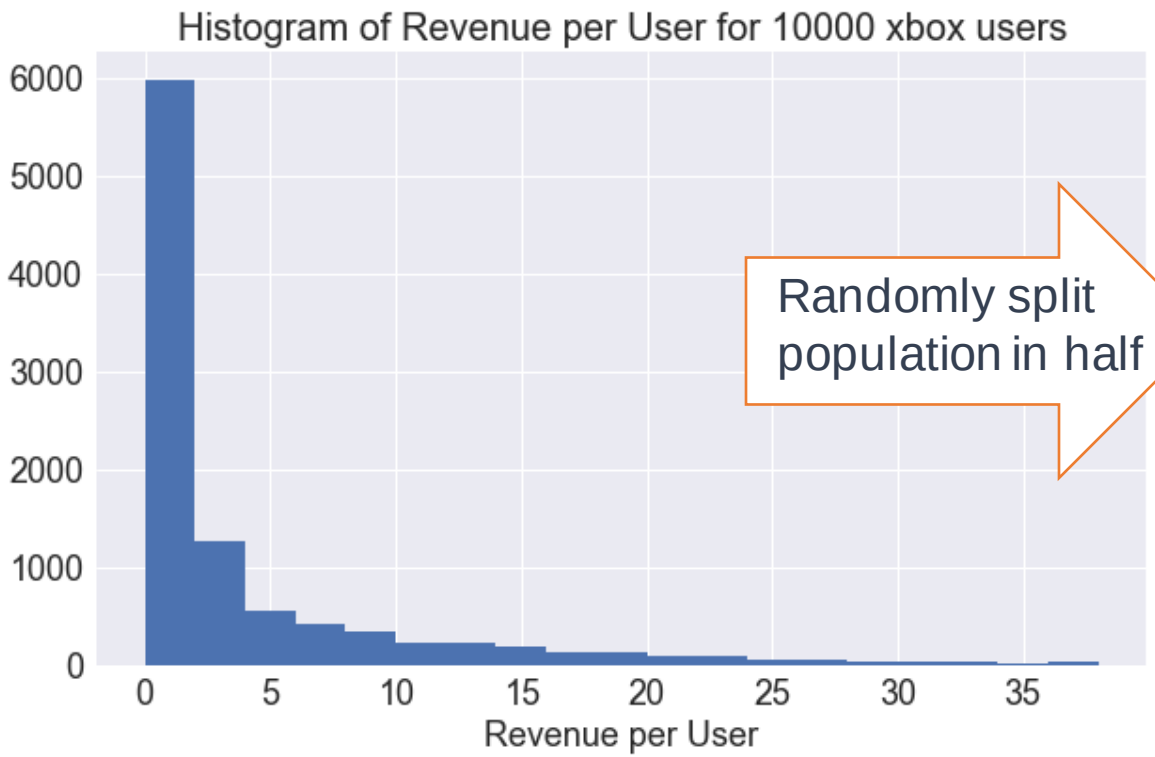
Random sampling



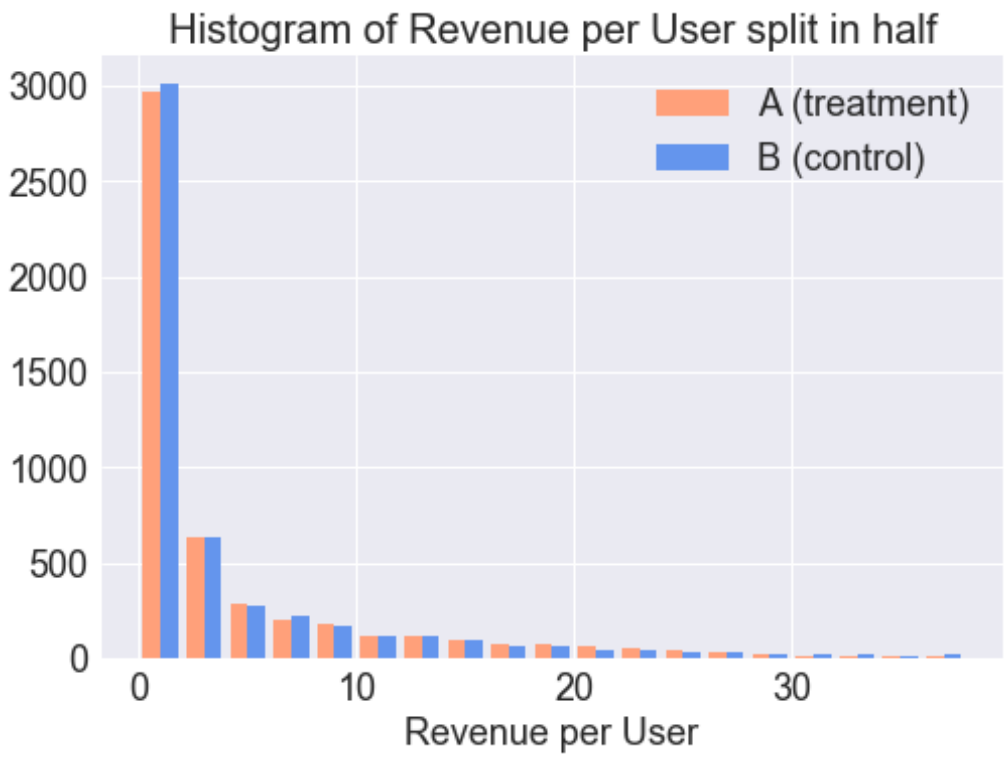
We randomly split our users into treatment and control



Random sampling



Randomly split population in half



Random sampling

In the original dataset, there were **4000 Xbox Live Gold users**. We randomly split our dataset in half.

About **how many Xbox Live Gold users** do we expect to see in each split?

Random sampling

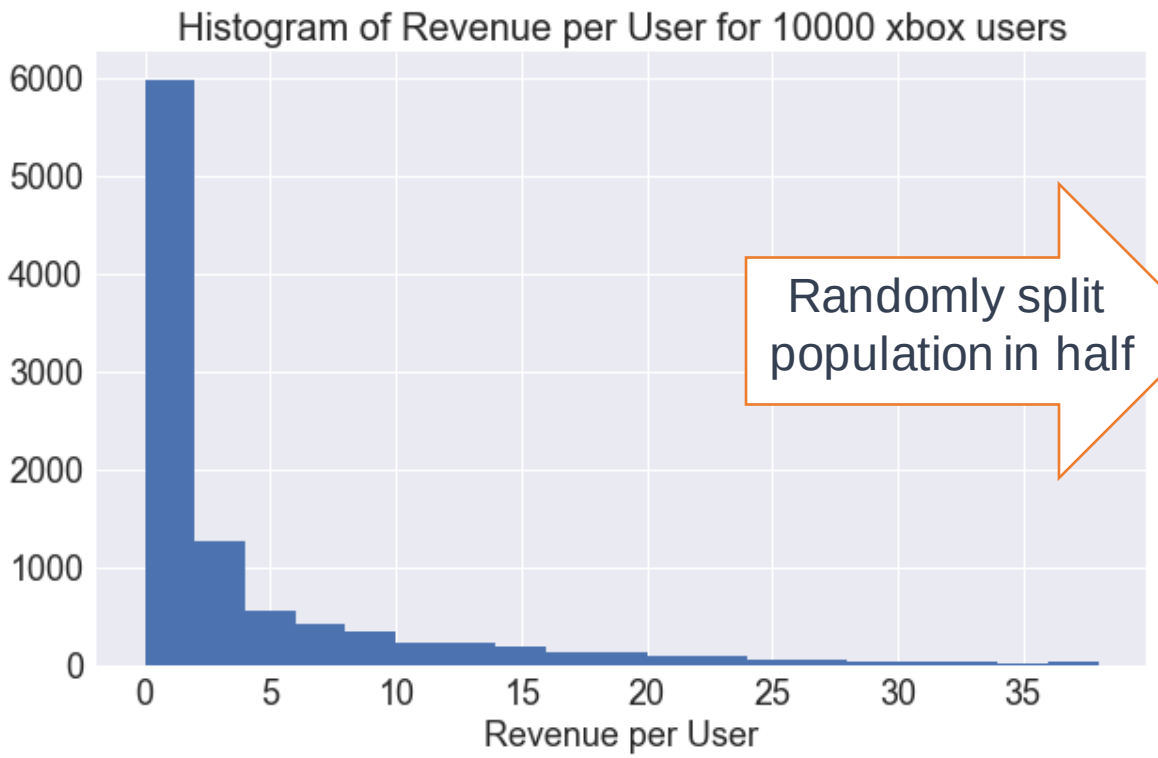
In the original dataset, there were 4000 Xbox Live Gold users. We randomly split our dataset in half.

About how many Xbox Live Gold users do we expect to see in each split?

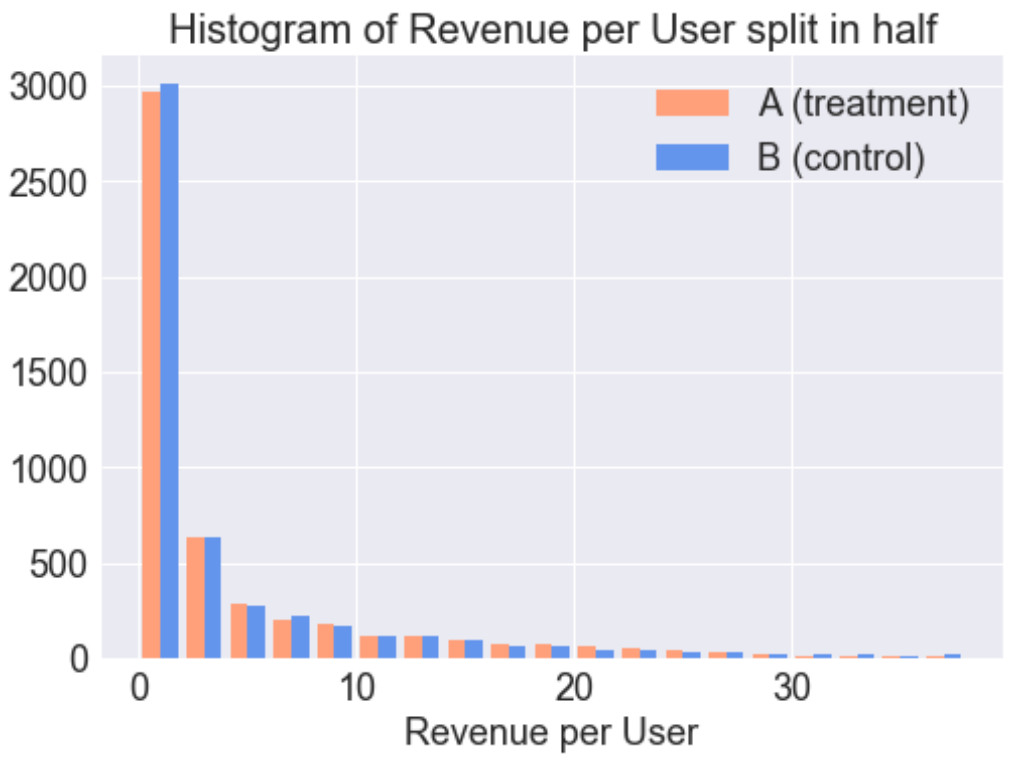
2000.

Indeed, we expect all groups, segments, markets to be approximately evenly distributed between treatment and control.

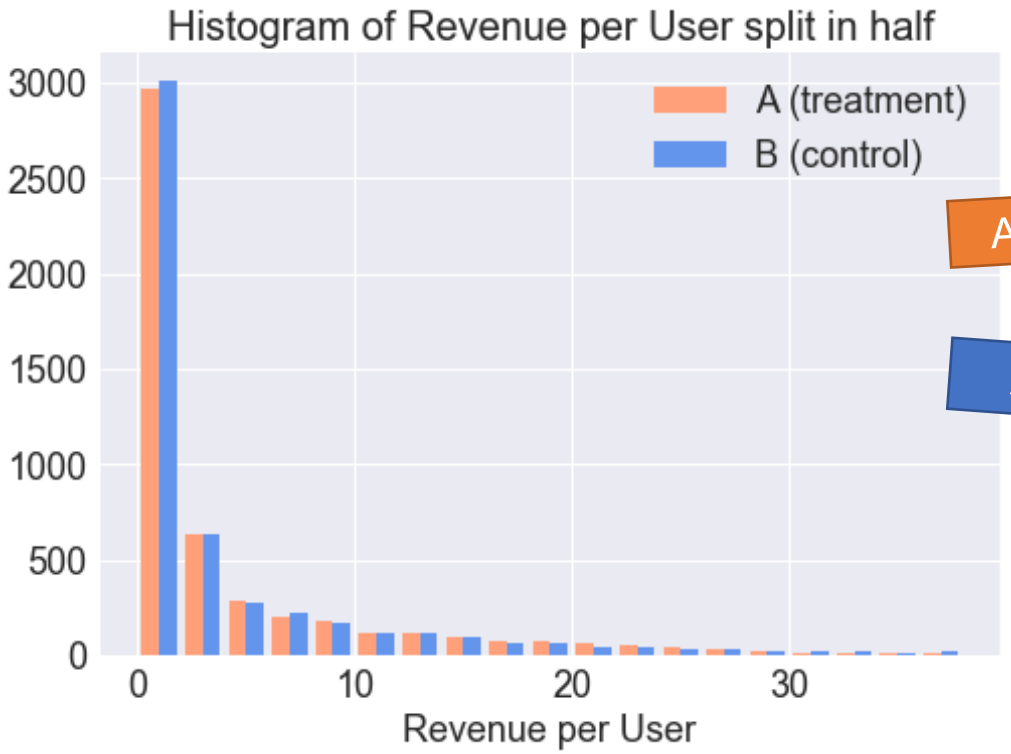
Random sampling



Randomly split population in half



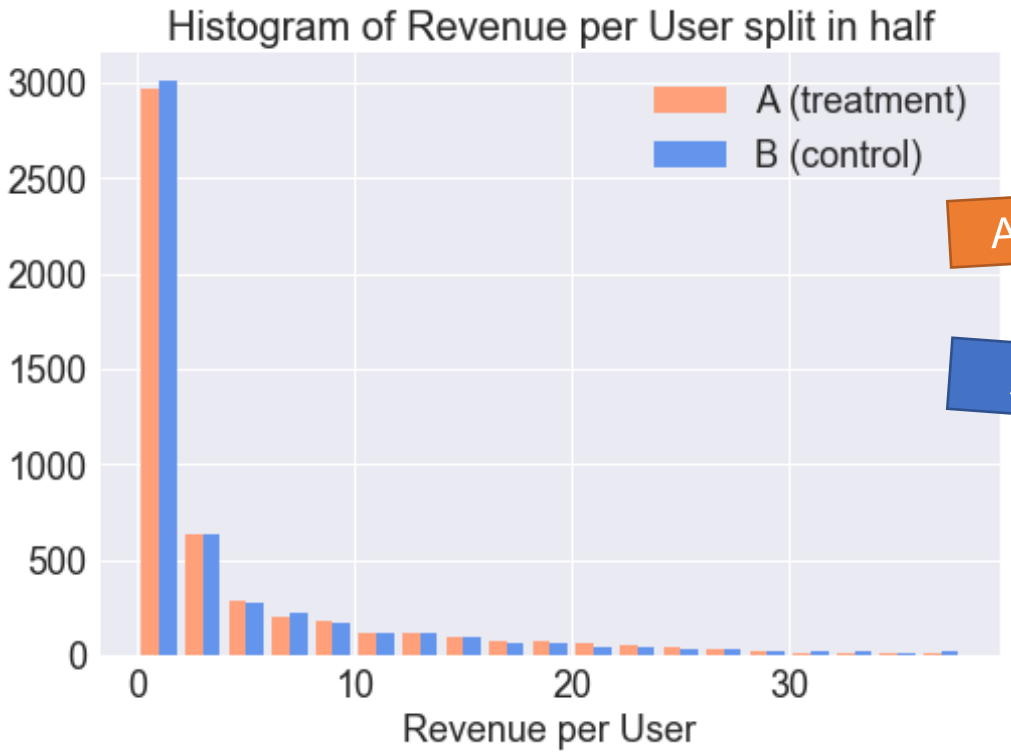
Measuring the Difference



Average(Treatment) $4.75

Average(Control) $4.43

Measuring the Difference

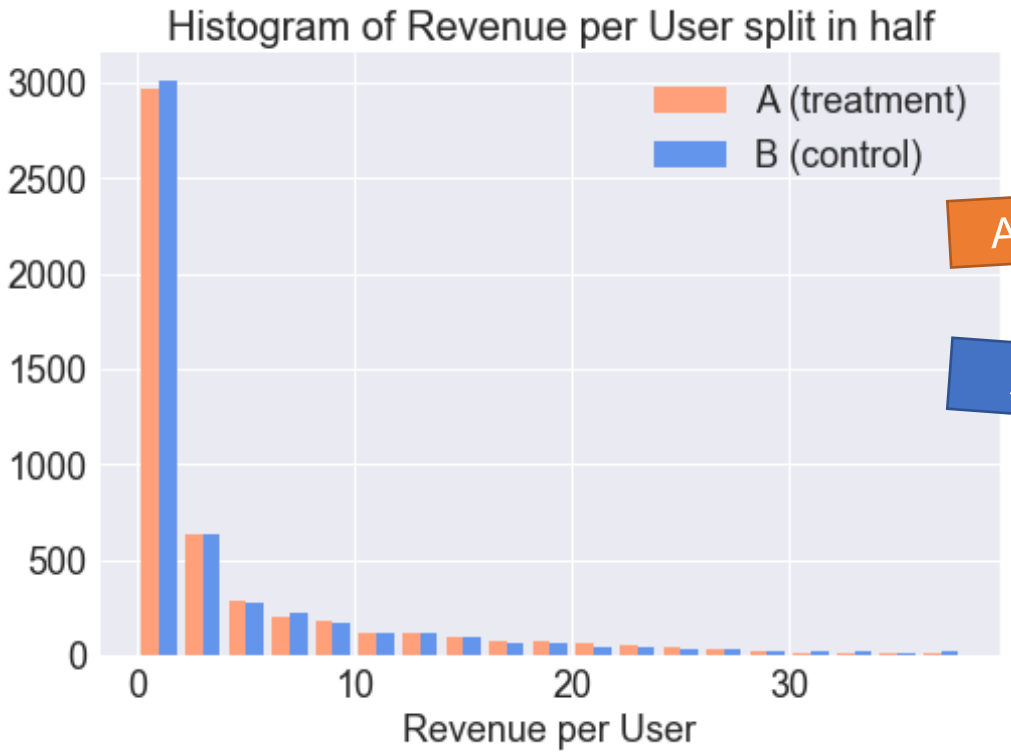


Average(Treatment) $4.75

Average(Control) -$4.43

$0.32

Measuring the Difference

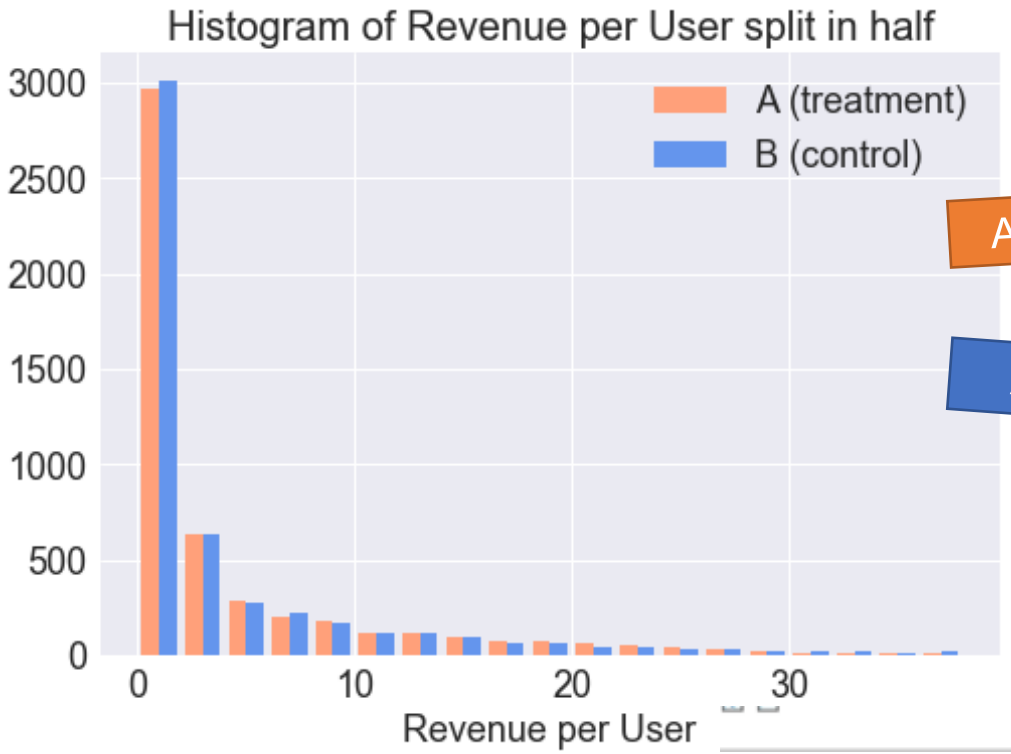


Average(Treatment) $4.75

Average(Control) -$4.43

$0.32 = “Delta” Δ

Measuring the Difference



Average(Treatment) $4.75

Average(Control) -$4.43

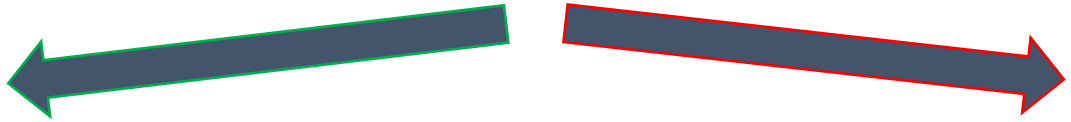
$0.32 = “Delta” Δ

	Treatment	Control	Delta	Delta %	P-Value
-- FeaturedPrimary_1 Clicks / UU	0.1389	0.1108	0.0281	+25.37%	0

Measuring the Difference

Treatment – Control = Delta

$4.75 – $4.43 = $0.32



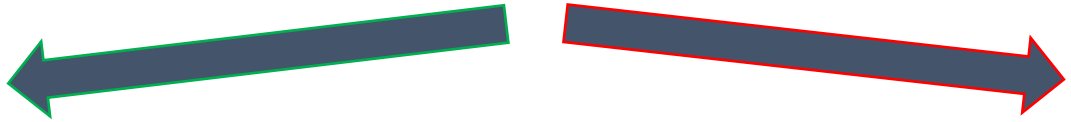
Our feature caused a difference

The difference was due to random chance

Measuring the Difference

Treatment – Control = Delta

$4.75 – $4.43 = $0.32



Our feature caused a difference

"Alternative Hypothesis"

The difference was due to random chance

"Null Hypothesis"

Which Hypothesis Do We Favor?

Our feature caused a difference

"Alternative Hypothesis"

The difference was due to random chance

"Null Hypothesis"

Which Hypothesis Do We Favor?

Our feature caused a difference

"Alternative Hypothesis"

The difference was due to random chance

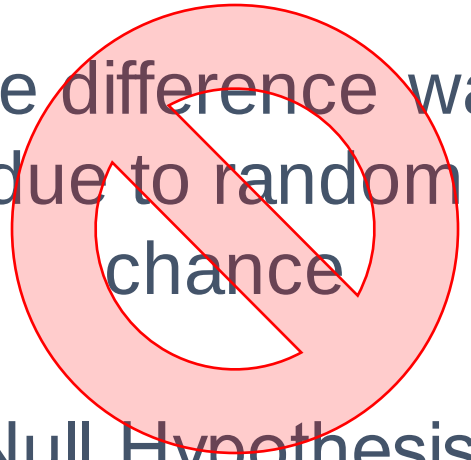
"Null Hypothesis"

These hypotheses are mutually exclusive
One of them **must** be true

Which Hypothesis Do We Favor?

Our feature caused a difference

"Alternative Hypothesis"



The difference was due to random chance

"Null Hypothesis"

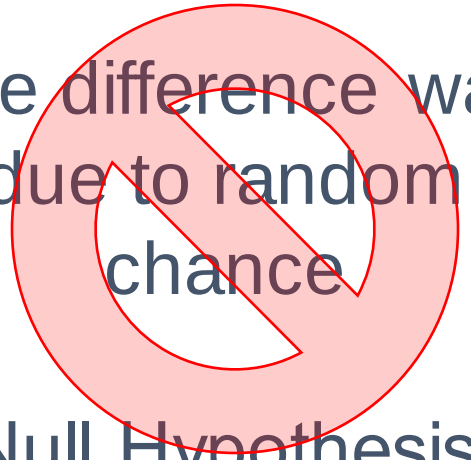
These hypotheses are mutually exclusive
One of them **must** be true

Which Hypothesis Do We Favor?

Our feature caused a difference

"Alternative Hypothesis"

The difference was due to random chance



"Null Hypothesis"

These hypotheses are mutually exclusive
One of them **must** be true

Evidence **against** the Null Hypothesis
is evidence **in favor of** the Alternative Hypothesis

Gathering Evidence Against the Null

- Assume the Null Hypothesis is true (the Treatment had no effect)

Gathering Evidence Against the Null

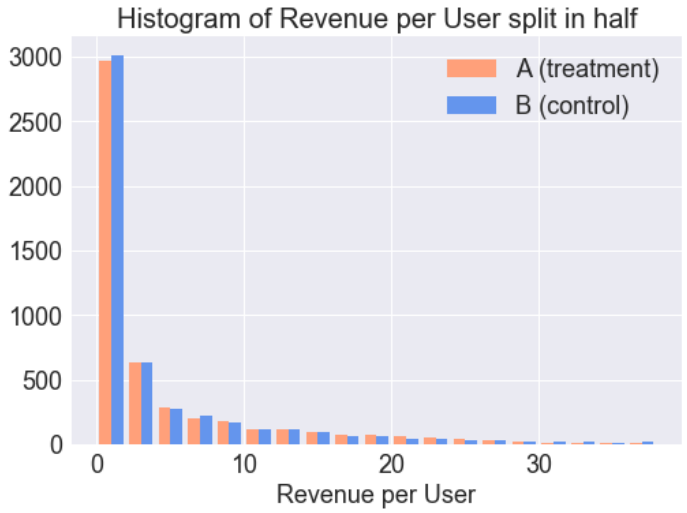
- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn’t affect them

Gathering Evidence Against the Null

- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn’t affect them
- Because this assignment didn’t matter, we can simulate the outcome of having assigned them differently

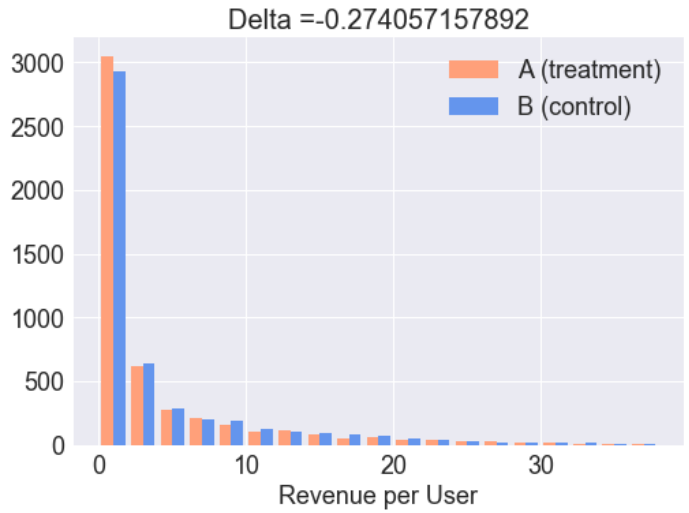
Gathering Evidence Against the Null

- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently



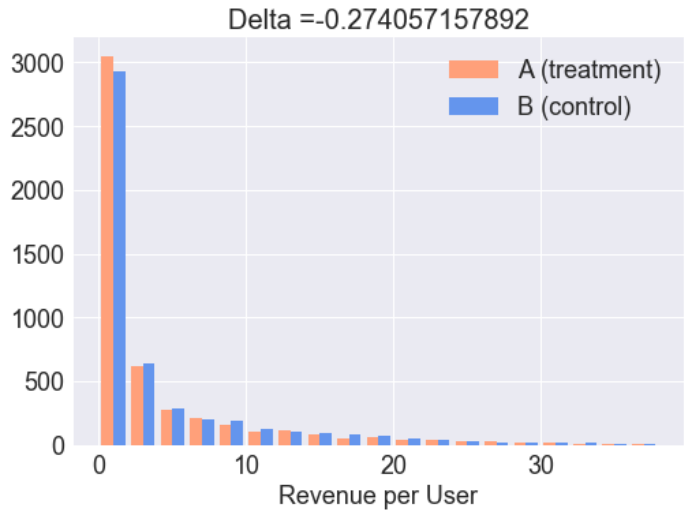
Gathering Evidence Against the Null

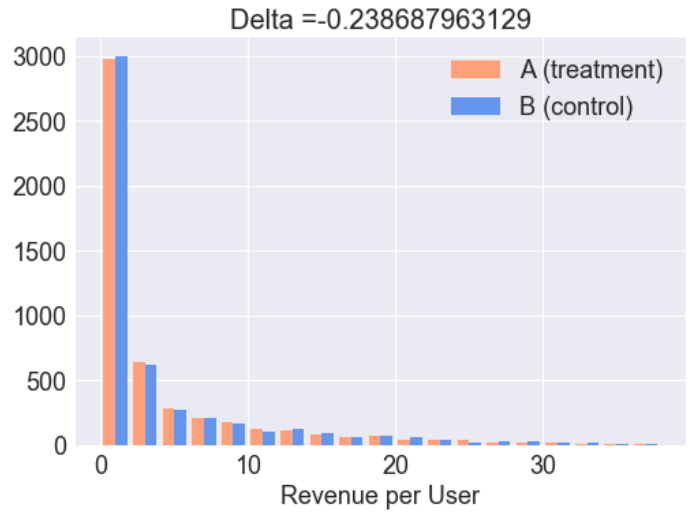
- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently



Gathering Evidence Against the Null

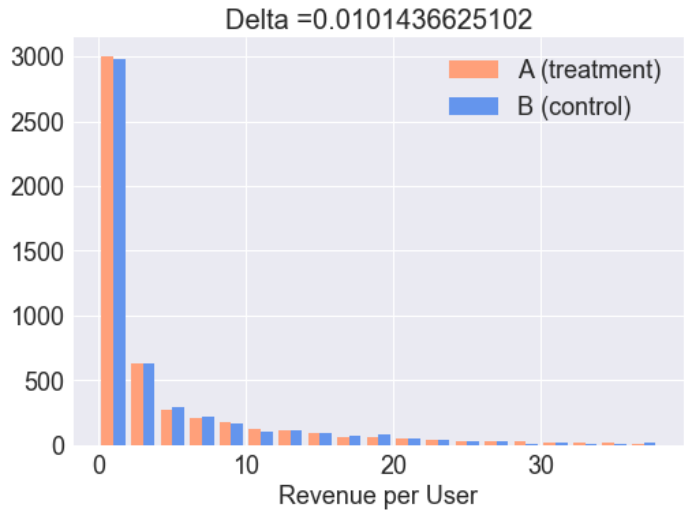
- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently

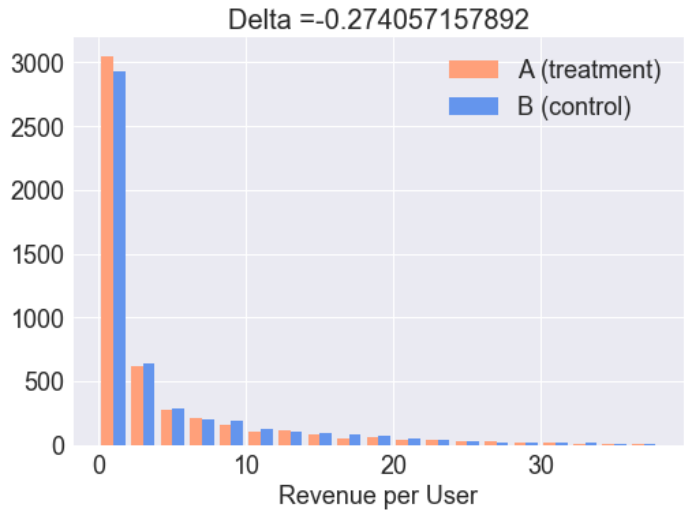


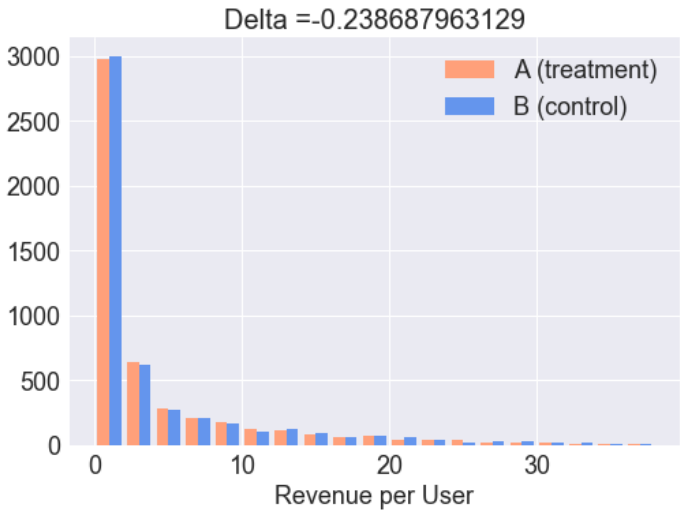


Gathering Evidence Against the Null

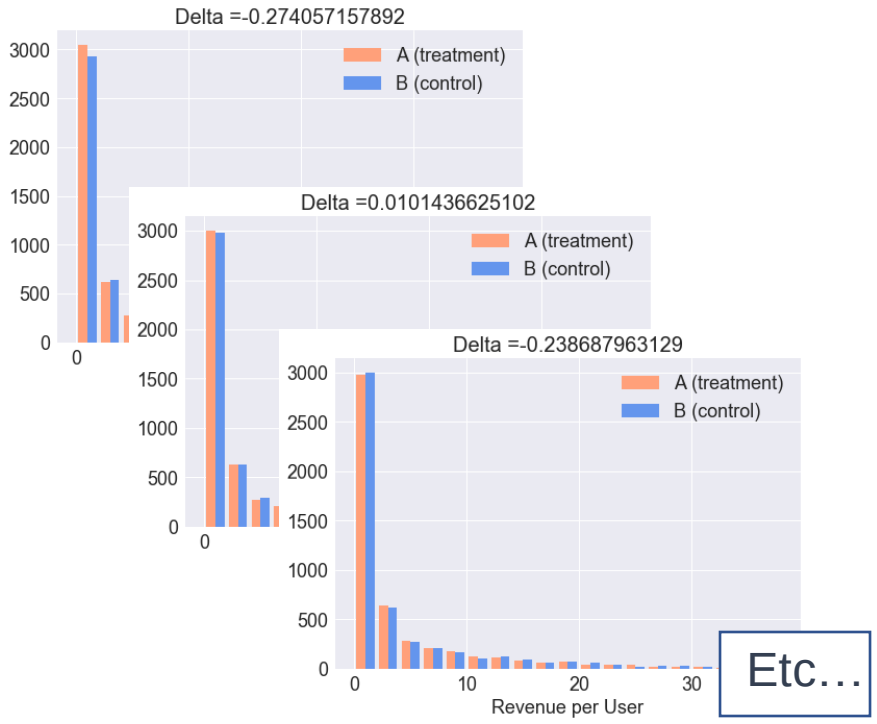
- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently



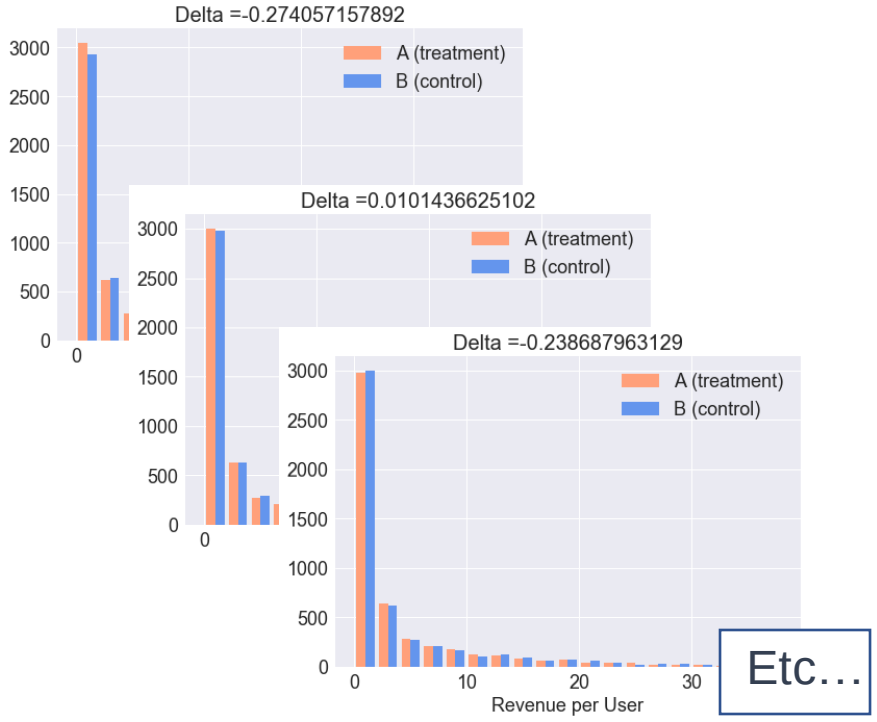




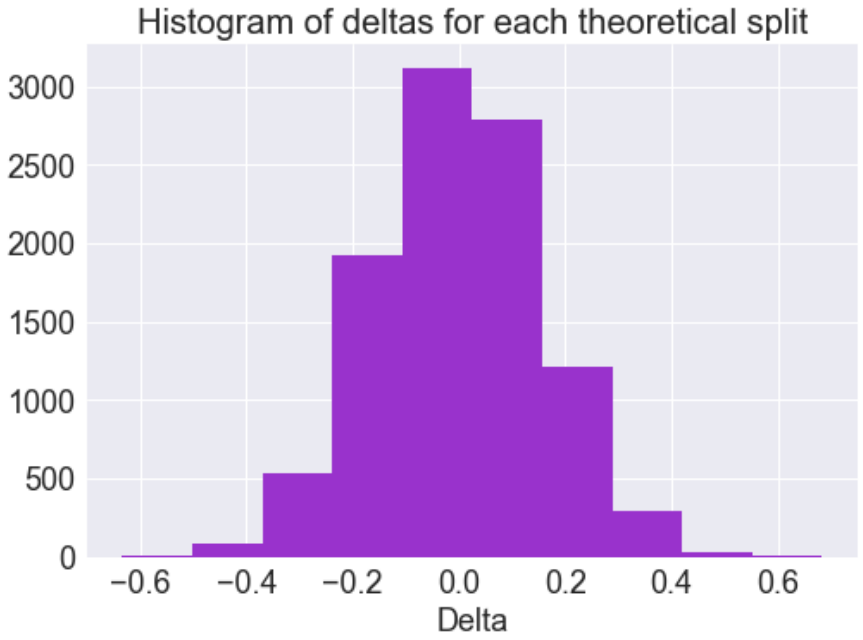
Gathering Evidence Against the Null



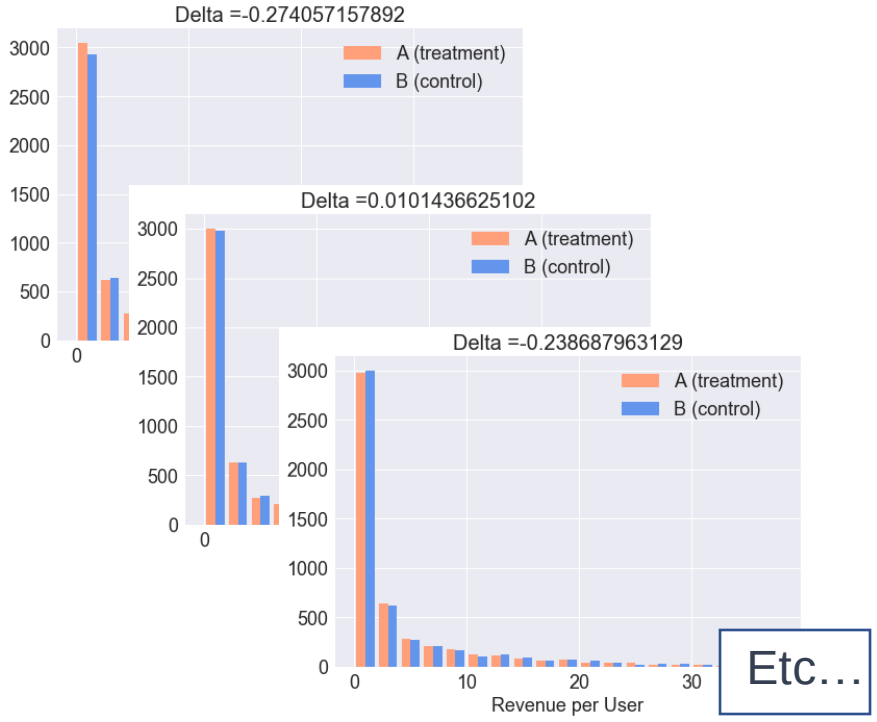
Gathering Evidence Against the Null



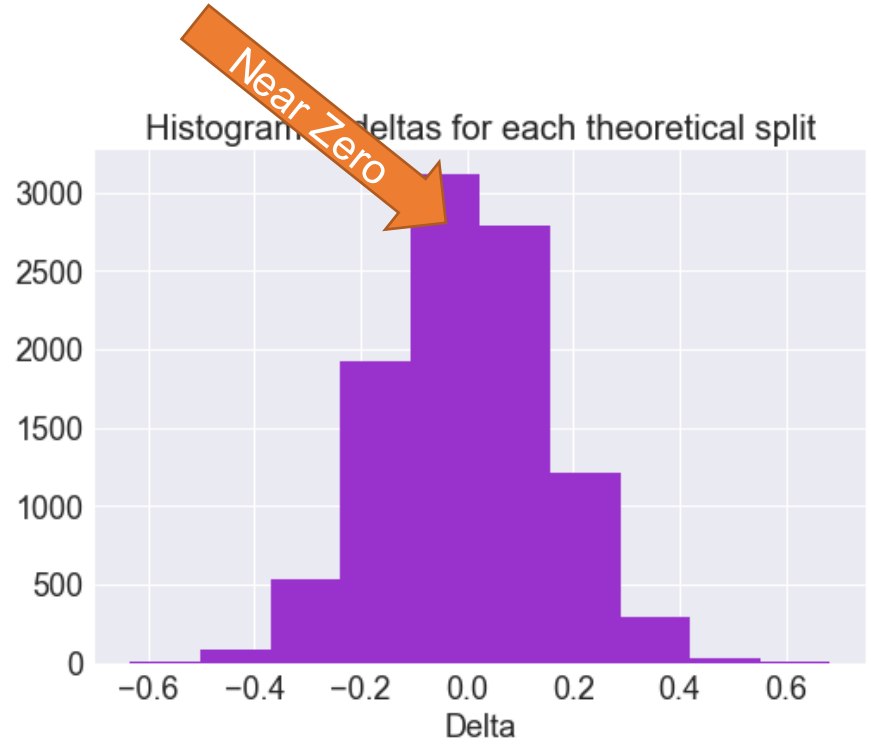
10,000x



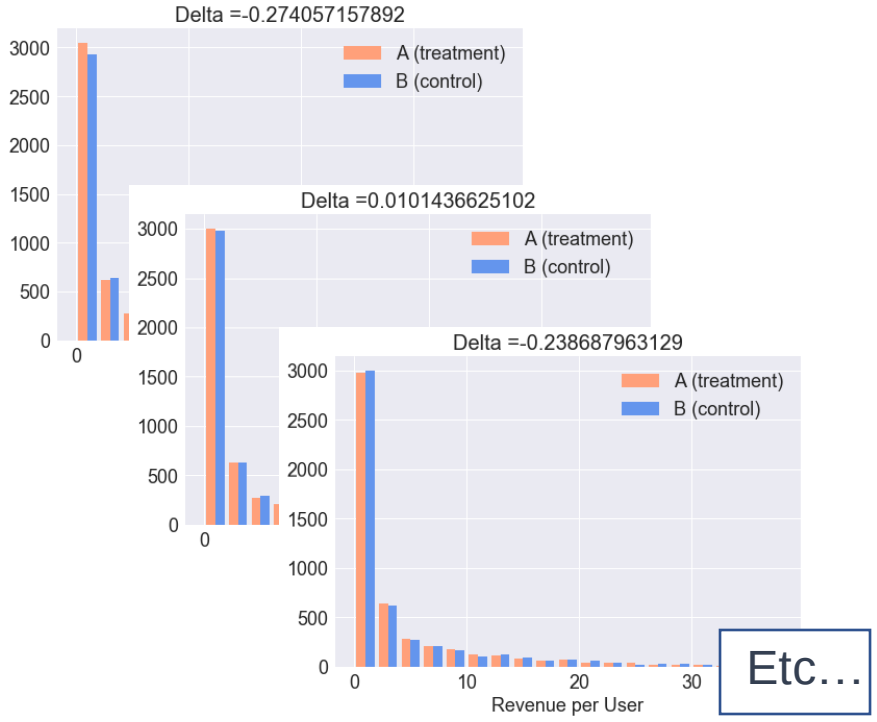
Gathering Evidence Against the Null

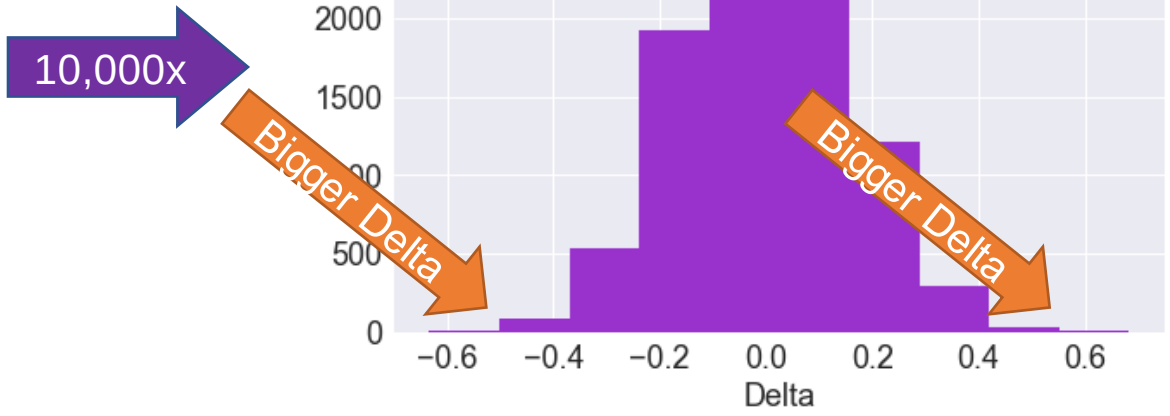


10,000x

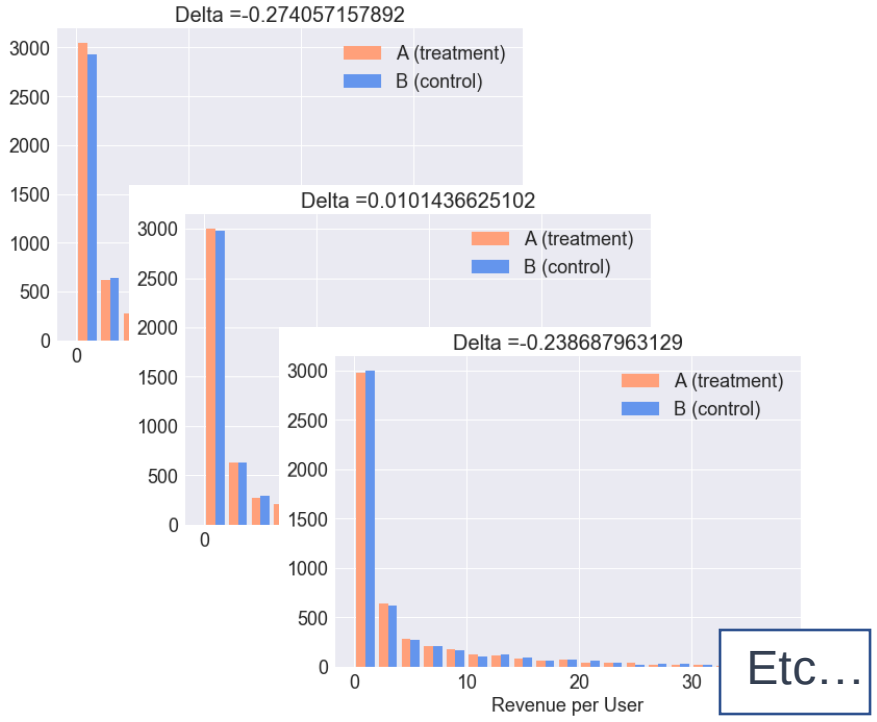


Gathering Evidence Against the Null

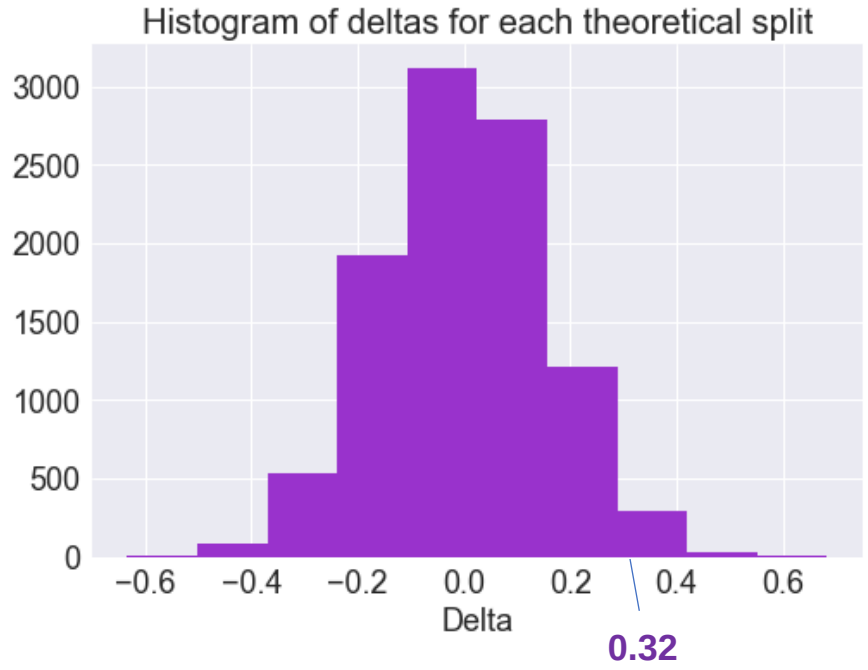




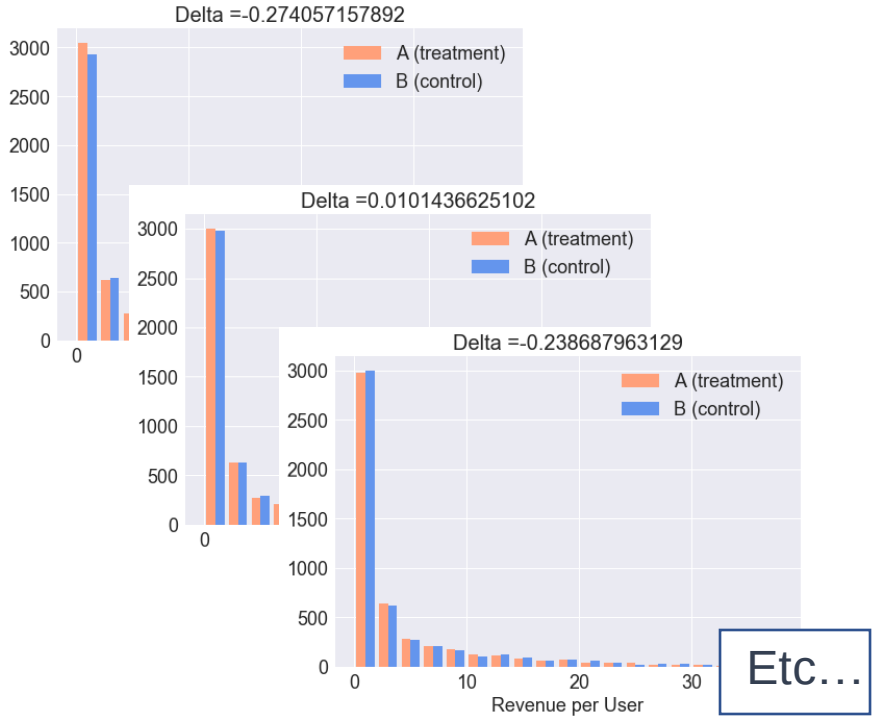
Gathering Evidence Against the Null



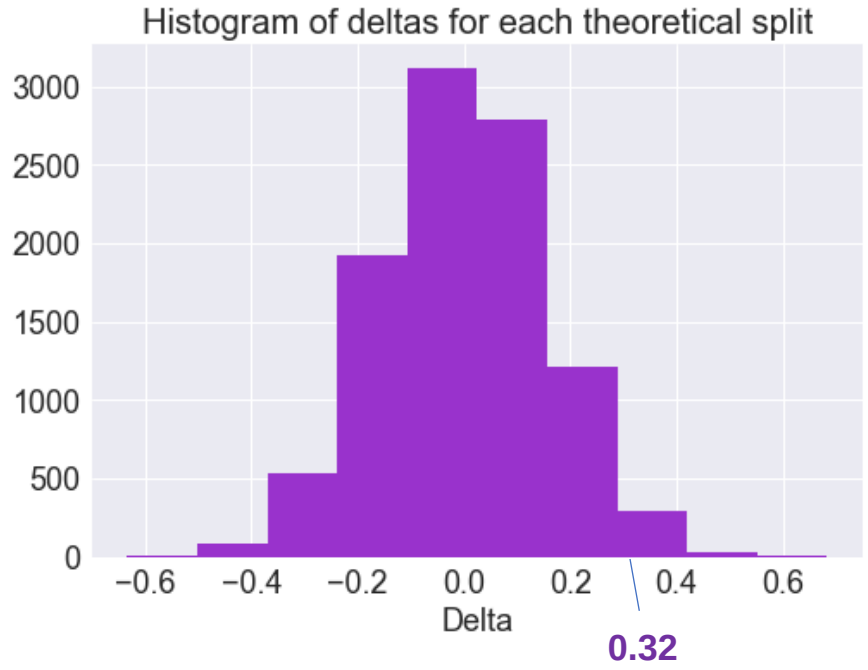
10,000x



Gathering Evidence Against the Null

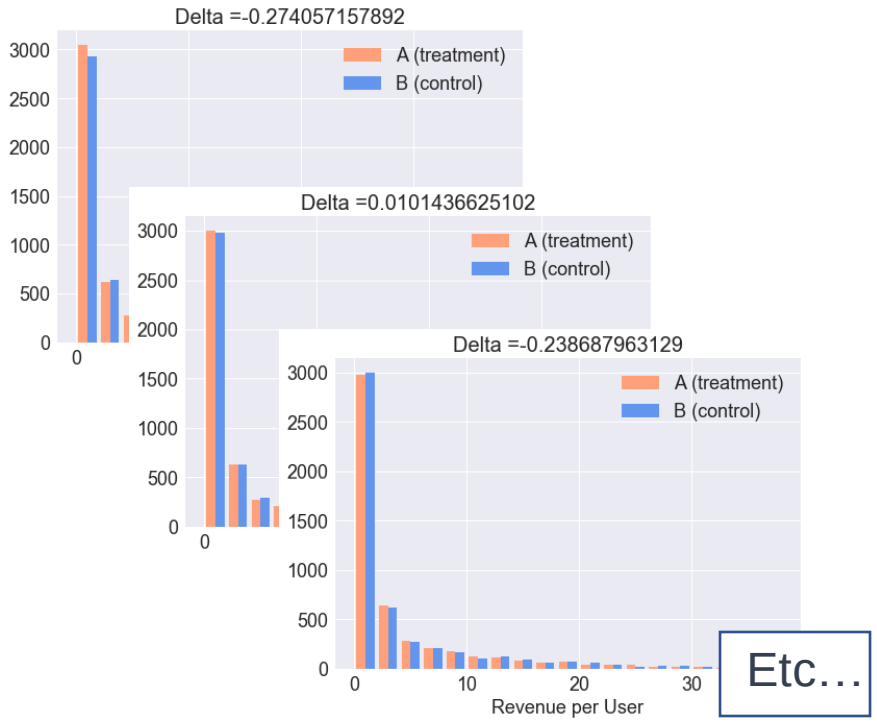


10,000x

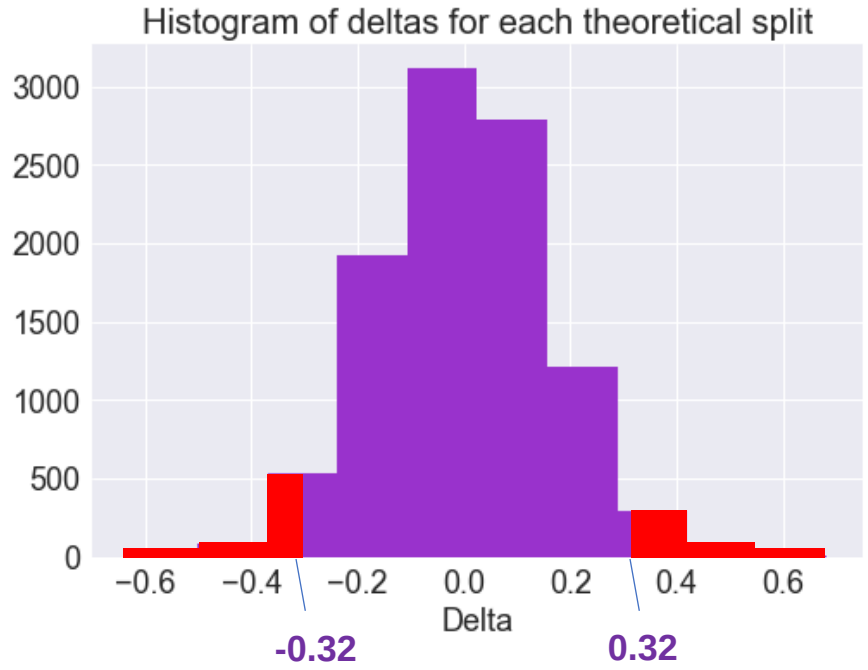


"If the null hypothesis is true, how unlikely is our delta of 0.32?"

Gathering Evidence Against the Null



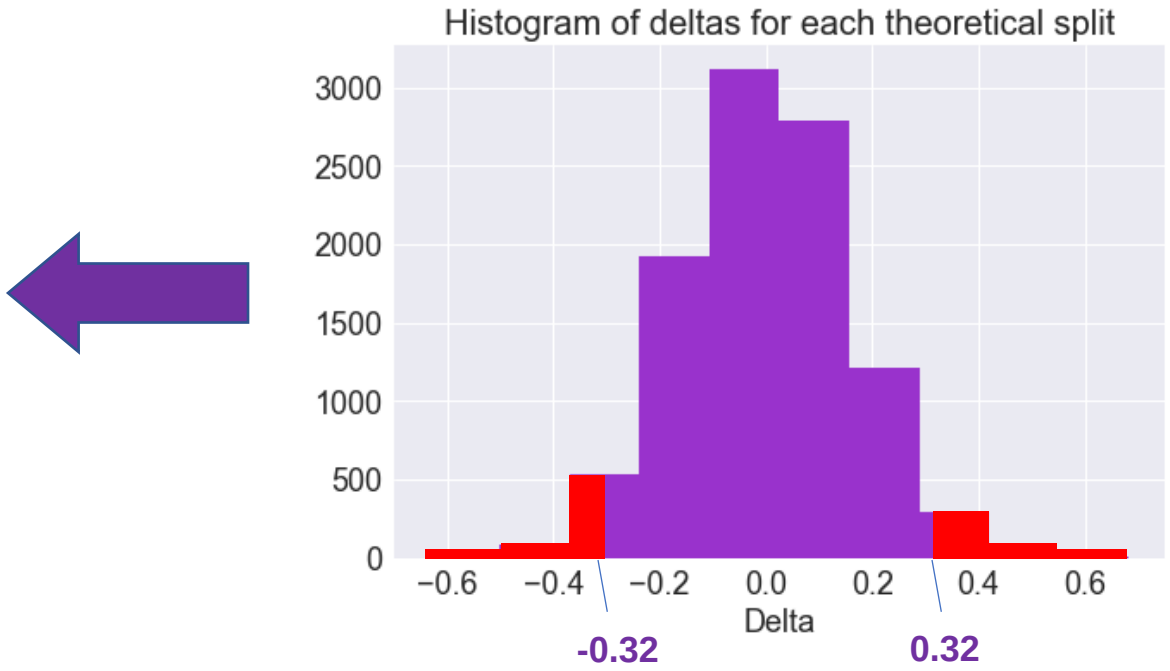
10,000x



"If the null hypothesis is true, how unlikely is our delta of 0.32?"

Gathering Evidence Against the Null

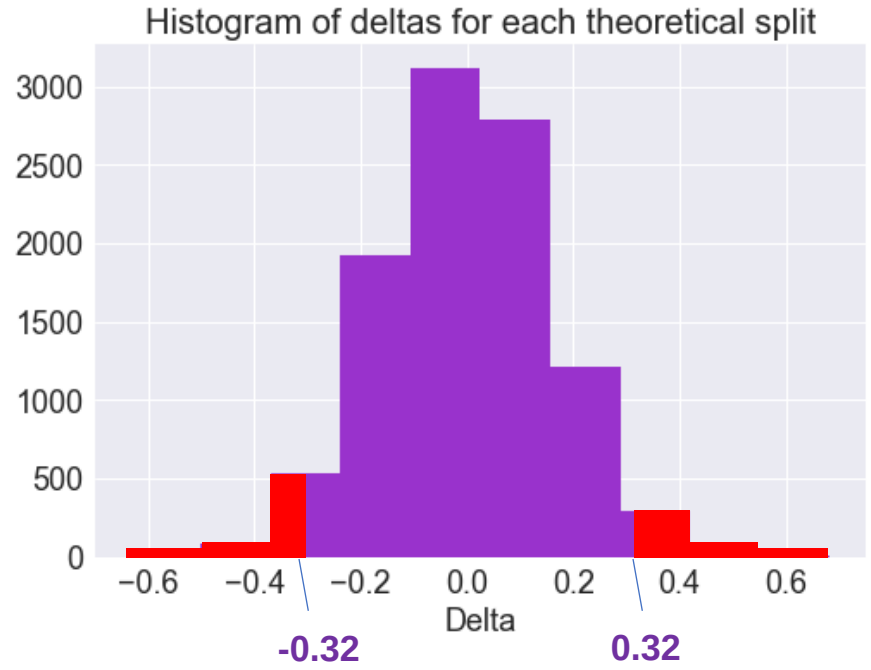
$$\frac{Red\ Area}{Total\ Area} = 0.047$$



"If the null hypothesis is true, how unlikely is our delta of 0.32?"

Gathering Evidence Against the Null

$$\frac{Red\ Area}{Total\ Area} = 0.047 = p$$



"If the null hypothesis is true, how unlikely is our delta of 0.32?"

Gathering Evidence Against the Null

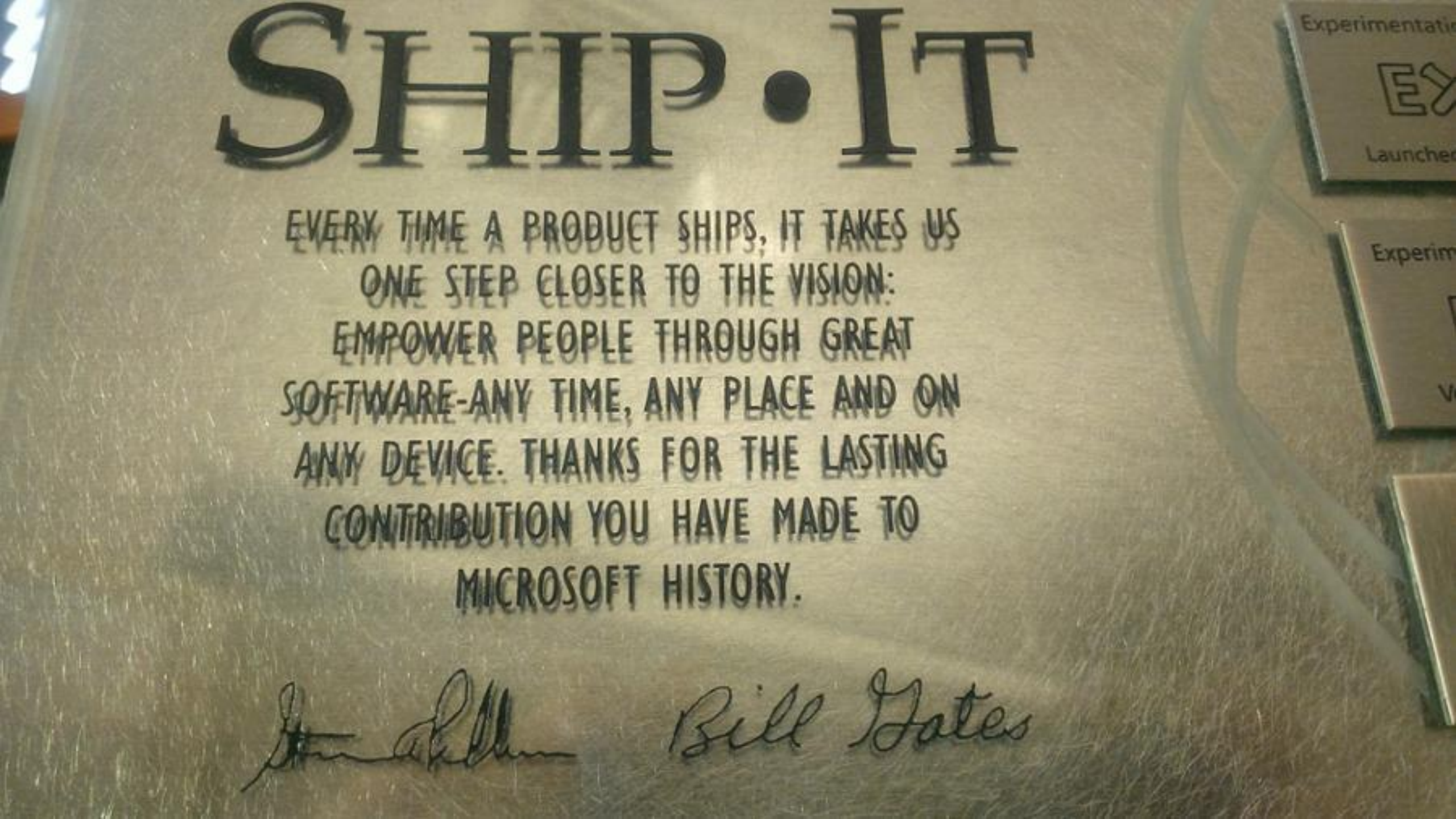
- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently
- In simulations where the Null Hypothesis is true, it's pretty rare (4.7%) to observe a delta as extreme as the one we saw

Gathering Evidence Against the Null

- **Assume** the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently
- In simulations where the Null Hypothesis is true, it's pretty rare (4.7%) to observe a delta as extreme as the one we saw
- If this *p*-value is small enough, maybe there was something wrong with our **assumption**… Maybe the Null Hypothesis isn't true!

Gathering Evidence Against the Null

- Assume the Null Hypothesis is true (the Treatment had no effect)
- If the Treatment had no effect, assigning users to the Treatment and Control groups didn't affect them
- Because this assignment didn't matter, we can simulate the outcome of having assigned them differently
- In simulations where the Null Hypothesis is true, it's pretty rare (4.7%) to observe a delta as extreme as the one we saw
- If this *p*-value is small enough, maybe there was something wrong with our assumption… Maybe the Null Hypothesis isn't true!
- **Maybe our treatment had an effect!**


SHIP • IT
EVERY TIME A PRODUCT SHIPS, IT TAKES US ONE STEP CLOSER TO THE VISION: EMPOWER PEOPLE THROUGH GREAT SOFTWARE-ANY TIME, ANY PLACE AND ON ANY DEVICE. THANKS FOR THE LASTING CONTRIBUTION YOU HAVE MADE TO MICROSOFT HISTORY.
Bill Gates

P-values in practice

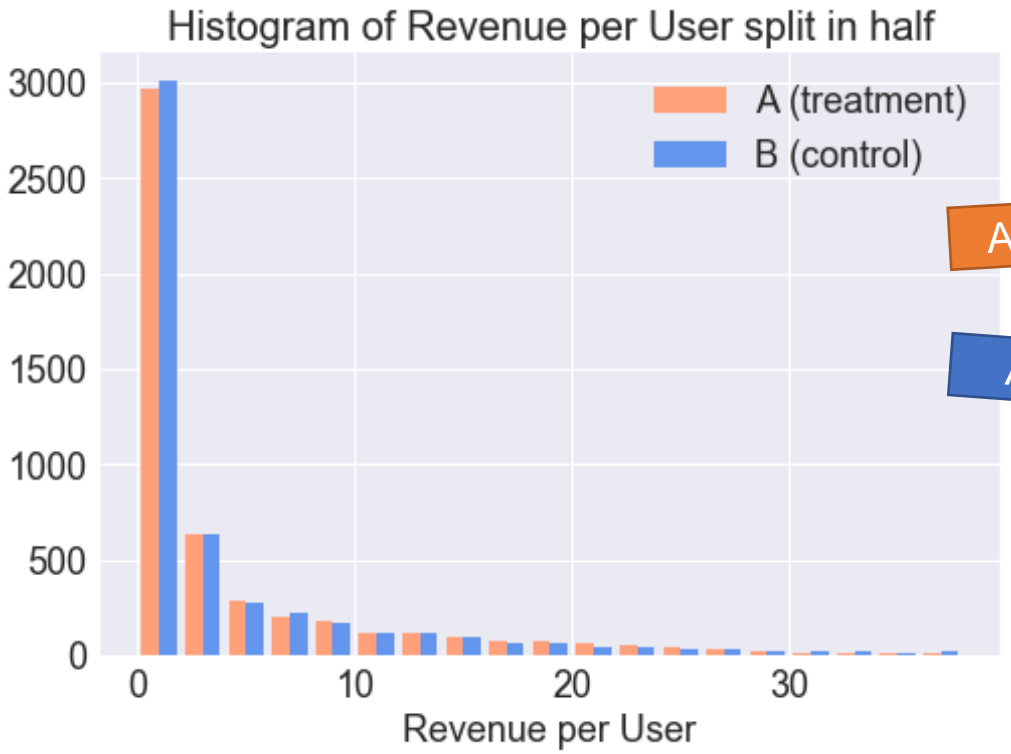
- In our score cards, we don't calculate p-values by randomly splitting our dataset many times – this introduction was meant to give you some intuition.
- In reality, we make certain assumptions about the distribution of the means of treatment and control, and this allows us to use a statistical test known as a *z*-test to obtain the p-value.
- This is much more efficient and robust than re-randomizing our dataset many times.

P-values general intuition

P-values essentially tell you **how unlucky you would have had to be to observe the values you observed, if there were no treatment effect**.

The lower the p-value, the less likely the outcome under the null, ie. the more evidence we have that the null is not true.

And now, a confession…



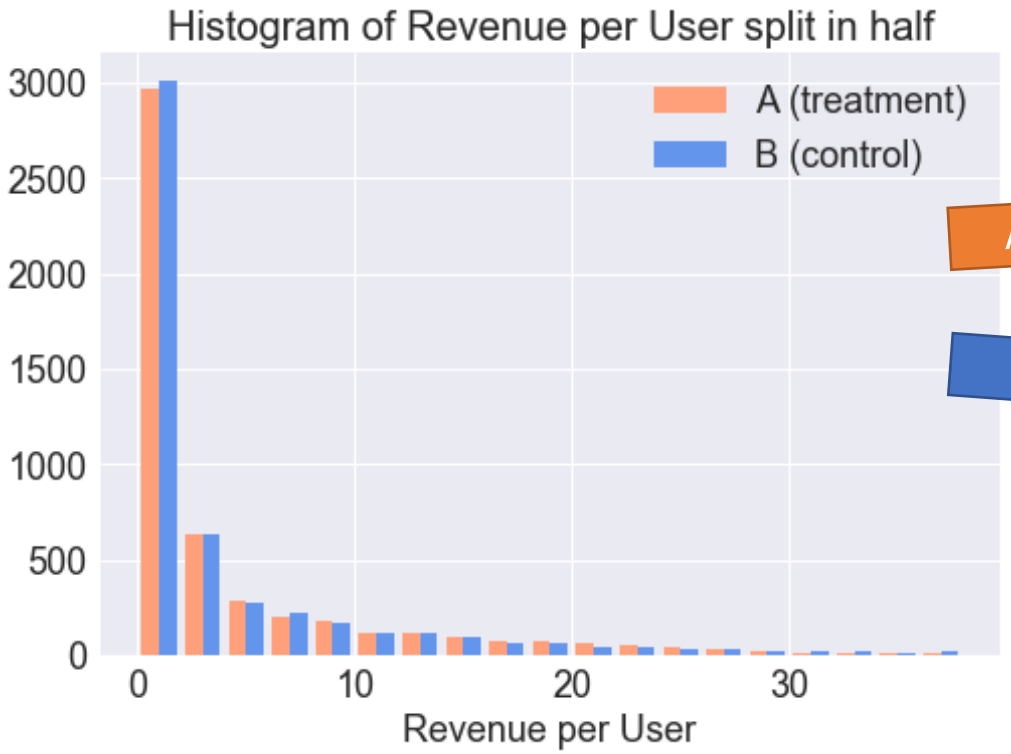
Average(Treatment) $4.75

Average(Control) -$4.43

$0.32

$p <$ 0.05!

And now, a confession… This was an A/A test!



Average(Treatment) $4.75

Average(Control) -$4.43

$0.32

$p <$ 0.05?

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”
- In this A/A test, the *p*-value was about 0.047, or about 1-in-21

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”
- In this A/A test, the *p*-value was about 0.047, or about 1-in-21
- A common threshold for claiming a *p*-value is “significant” is 0.05

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”
- In this A/A test, the *p*-value was about 0.047, or about 1-in-21
- A common threshold for claiming a *p*-value is “significant” is 0.05
- Using this threshold, false positives will occur 1-in-20 times

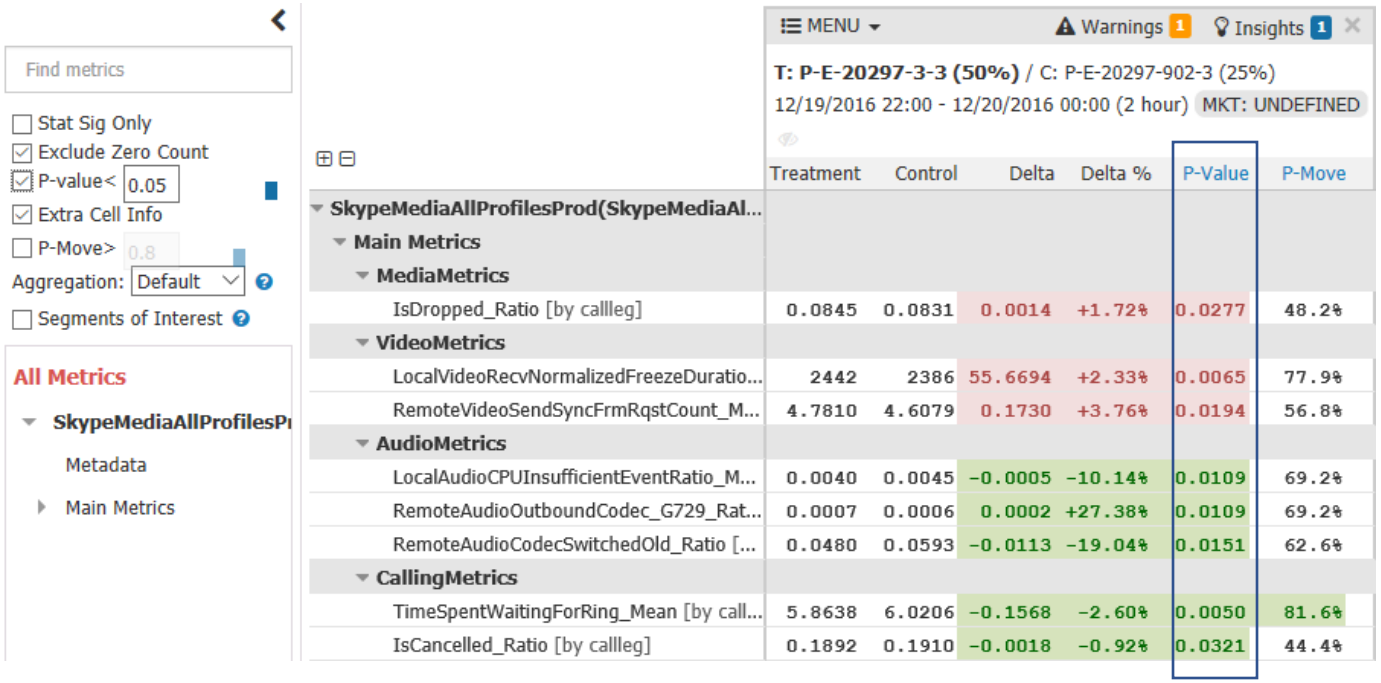
False Positives – a.k.a. “Type I Errors”

- Assume no treatment effect, in a score card with 1000 **metrics in a given experiment**, how many would you expect to show a stat sig ($p < 0.05$) movement?

False Positives – a.k.a. “Type I Errors”

- Assume no treatment effect, in a score card with 1000 **metrics in a given experiment**, how many would you expect to show a stat sig ($p < 0.05$) movement? **50!**

False Positives – a.k.a. “Type I Errors”



	Treatment	Control	Delta	Delta %	P-Value	P-Move
SkypeMediaAllProfilesProd(SkypeMediaAl...						
Main Metrics						
MediaMetrics						
IsDropped_Ratio [by callleg]	0.0845	0.0831	0.0014	+1.72%	0.0277	48.2%
VideoMetrics						
LocalVideoRecvNormalizedFreezeDuratio...	2442	2386	55.6694	+2.33%	0.0065	77.9%
RemoteVideoSendSyncFrmRqstCount_M...	4.7810	4.6079	0.1730	+3.76%	0.0194	56.8%
AudioMetrics						
LocalAudioCPUInsufficientEventRatio_M...	0.0040	0.0045	-0.0005	-10.14%	0.0109	69.2%
RemoteAudioOutboundCodec_G729_Rat...	0.0007	0.0006	0.0002	+27.38%	0.0109	69.2%
RemoteAudioCodecSwitchedOld_Ratio [...	0.0480	0.0593	-0.0113	-19.04%	0.0151	62.6%
CallingMetrics						
TimeSpentWaitingForRing_Mean [by call...	5.8638	6.0206	-0.1568	-2.60%	0.0050	81.6%
IsCancelled_Ratio [by callleg]	0.1892	0.1910	-0.0018	-0.92%	0.0321	44.4%

This scorecard has 200 metrics.

Here, for a given experiment, I filtered by metrics with a p-value < .05

Expect some number of these metrics to correspond to false positives, especially metrics with p-values that aren’t extremely small.

False Positives – a.k.a. “Type I Errors”

- Assume no treatment effect, in a score card with 1000 **metrics in a given experiment**, how many would you expect to show a stat sig ($p < 0.05$) movement? **50!**

- Again, assume no treatment effect. If we ran 100 **experiments**, how many would show a stat sig change in revenue?

False Positives – a.k.a. “Type I Errors”

- Assume no treatment effect, in a score card with 1000 **metrics in a given experiment**, how many would you expect to show a stat sig ($p < 0.05$) movement? **50!**

- Again, assume no treatment effect. If we ran 100 **experiments**, how many would show a stat sig change in revenue? **5!**

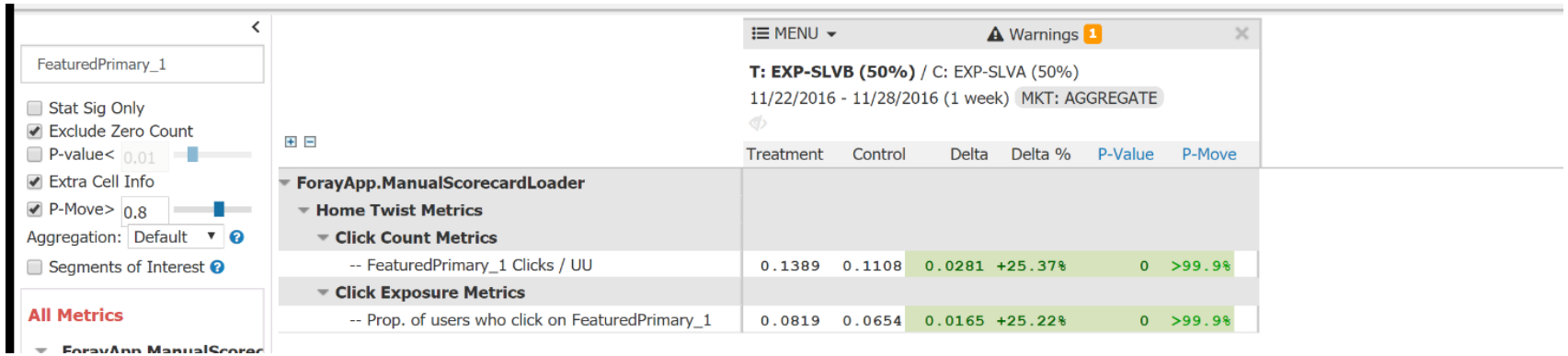
False Positives – a.k.a. “Type I Errors”

- Assume no treatment effect, in a score card with 1000 **metrics in a given experiment**, how many would you expect to show a stat sig ($p < 0.05$) movement? **50!**

- Again, assume no treatment effect. If we ran 100 **experiments**, how many would show a stat sig change in revenue? **5!**

- Remember: half of those will be positive movements – half will be negative.

Trusting a stat sig movement

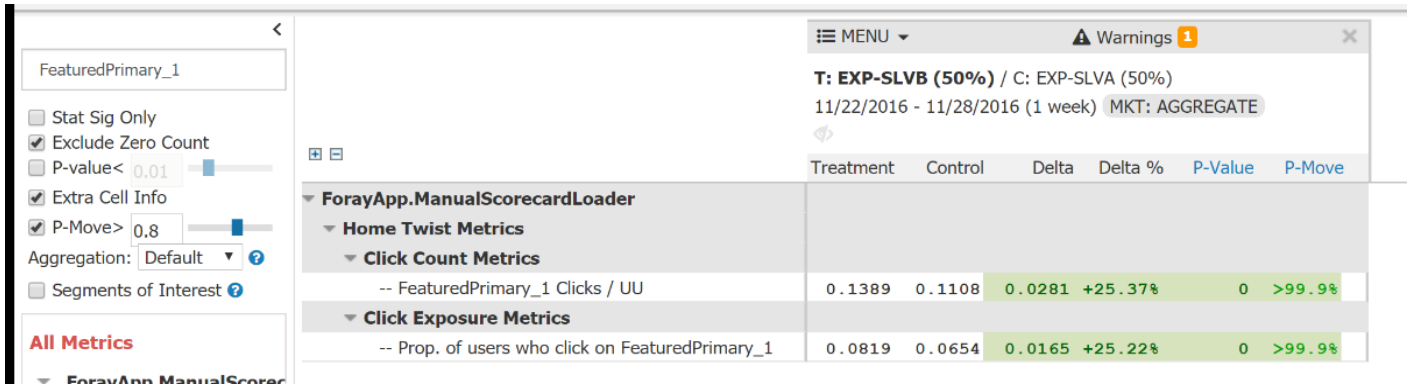


Background: This experiment involved modifying the FeaturedPrimary_1 tile in order to increase user engagement with it.

These are the results for FeaturedPrimary_1 related metrics.

Do you believe there is a difference between treatment and control? Why?

Trusting a stat sig movement



- P-values are very low
- These are metrics we expected to move!

Yes.

Trusting a stat sig movement

- If you are unsure whether or not to trust a movement (borderline p-value):
  - run a replication experiment (best way to make sure a movement is truly stat sig)!
  - or see if this replicates the findings of a previous experiment

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”
- In this A/A test, the *p*-value was about 0.047, or about 1-in-21
- A common threshold for claiming a *p*-value is “significant” is 0.05
- Using this threshold, false positives will occur 1-in-20 times
- If we use a different threshold, we can change our false positive rate
- A *p*-value threshold of 0.01 would result in a 1-in-100 Type I error rate

False Positives – a.k.a. “Type I Errors”

- *p*-value: “If the null hypothesis is true, how unlikely is our delta?”
- In this A/A test, the *p*-value was about 0.047, or about 1-in-21
- A common threshold for claiming a *p*-value is “significant” is 0.05
- Using this threshold, false positives will occur 1-in-20 times
- If we use a different threshold, we can change our false positive rate
- A *p*-value threshold of 0.01 would result in a 1-in-100 Type I error rate
- Tradeoff: Loss of “Power” – we will miss more real treatment effects

Power and False Negatives (Type II Errors)

- *p*-value: “If the null hypothesis is true, how unlikely is our data?”
- Power: “If the null hypothesis is false, how likely are we to detect it?”

Power and False Negatives (Type II Errors)

- *p*-value: “If the null hypothesis is true, how unlikely is our data?”
- Power: “If the null hypothesis is false, how likely are we to detect it?”
- By lowering our *p*-value threshold, we decreased our power

Power and False Negatives (Type II Errors)

- *p*-value: “If the null hypothesis is true, how unlikely is our data?”
- Power: “If the null hypothesis is false, how likely are we to detect it?”
- By lowering our *p*-value threshold, we decreased our power
- False Negatives – There was a real treatment effect we failed to find

Power and False Negatives (Type II Errors)

- *p*-value: “If the null hypothesis is true, how unlikely is our data?”
- Power: “If the null hypothesis is false, how likely are we to detect it?”
- By lowering our *p*-value threshold, we decreased our power
- False Negatives – There was a real treatment effect we failed to find
- If we can increase our **Power**, we can decrease our False Negative rate

Factors Which Affect Power

Power is the probability of rejecting the null hypothesis, given that it is false (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. When the treatment effect is large
2. When the treatment effect is small

Factors Which Affect Power

Power is the probability of rejecting the null hypothesis, given that it is false. (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. **When the treatment effect is large**
2. ~~When the treatment effect is small~~

If there is a bigger difference, I am more likely to detect it!

Factors Which Affect Power

- The larger the treatment effect, the greater your power

Factors Which Affect Power

Power is the probability of rejecting the null hypothesis, given that it is false (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. When the sample size is large
2. When the sample size is small

Factors Which Affect Power

Power is the probability of rejecting the null hypothesis, given that it is false (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. **When the sample size is large**
2. ~~When the sample size is small~~

With a bigger sample size, we can more accurately estimate the means of the populations.

Factors Which Affect Power

- The larger the treatment effect, the greater your power
- The more users in your experiment, the greater your power

Factors Which Affect Power

- The larger the treatment effect, the greater your power
- The more users in your experiment, the greater your power
- If you want to detect a large effect, you won’t need many users
- If you want to detect a small effect, you need a lot of users

Factors Which Affect Power

- The larger the treatment effect, the greater your power
- The more users in your experiment, the greater your power
- If you want to detect a large effect, you won’t need many users
- If you want to detect a small effect, you need a lot of users
- This means some effects will always be too small to measure

Factors Which Affect Power

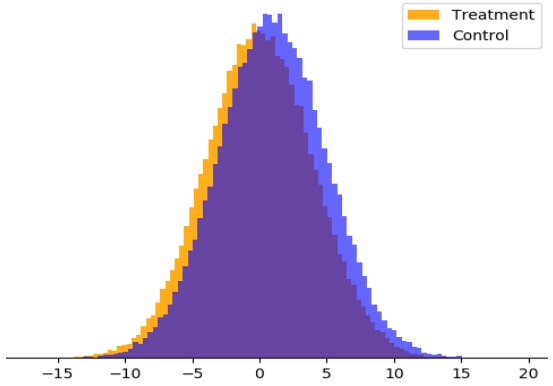
- The **larger the treatment effect**, the greater your power
- The **more users in your experiment**, the greater your power
- The **lower your metric's variance**, the greater your power

Factors Which Affect Power

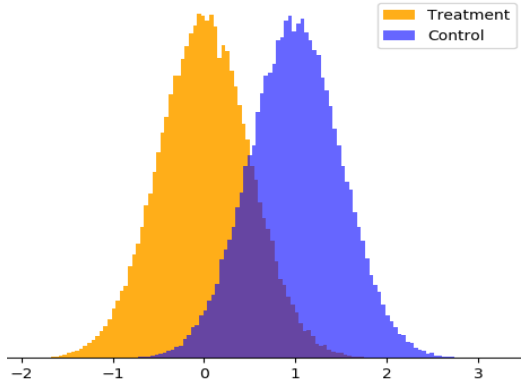
Power is the probability of rejecting the null hypothesis, given that it is false (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. When variances (spread) of the underlying populations are small
2. When the variances (spread) of the underlying populations are large



High Variance



Low Variance

Factors Which Affect Power

Power is the probability of rejecting the null hypothesis, given that it is false (i.e. our ability to detect a non-zero treatment effect)

In which scenario do I have the most power?

1. **When variances (spread) of the underlying populations are small**
2. ~~When the variances (spread) of the underlying populations are large~~

As with a bigger sample size, when the variances of the underlying populations are small, we can more accurately estimate the means.

False Positives and False Negatives Summary

- **False Positives a.k.a. Type I Errors**
  - There was no real treatment effect but we think we detected one
  - Limit False Positive rate by decreasing significance threshold for *p*-values
  - Tradeoff: loss of power

False Positives and False Negatives Summary

- **False Positives a.k.a. Type I Errors**
  - There was no real treatment effect but we think we detected one
  - Limit False Positive rate by decreasing significance threshold for *p*-values
  - Tradeoff: loss of power
- **False Negatives a.k.a. Type II Errors**
  - There was a real treatment effect but we failed to detect it
  - Limit False Negative rate by increasing power
    - Larger treatment effects
    - Larger user counts
    - Smaller metric variances

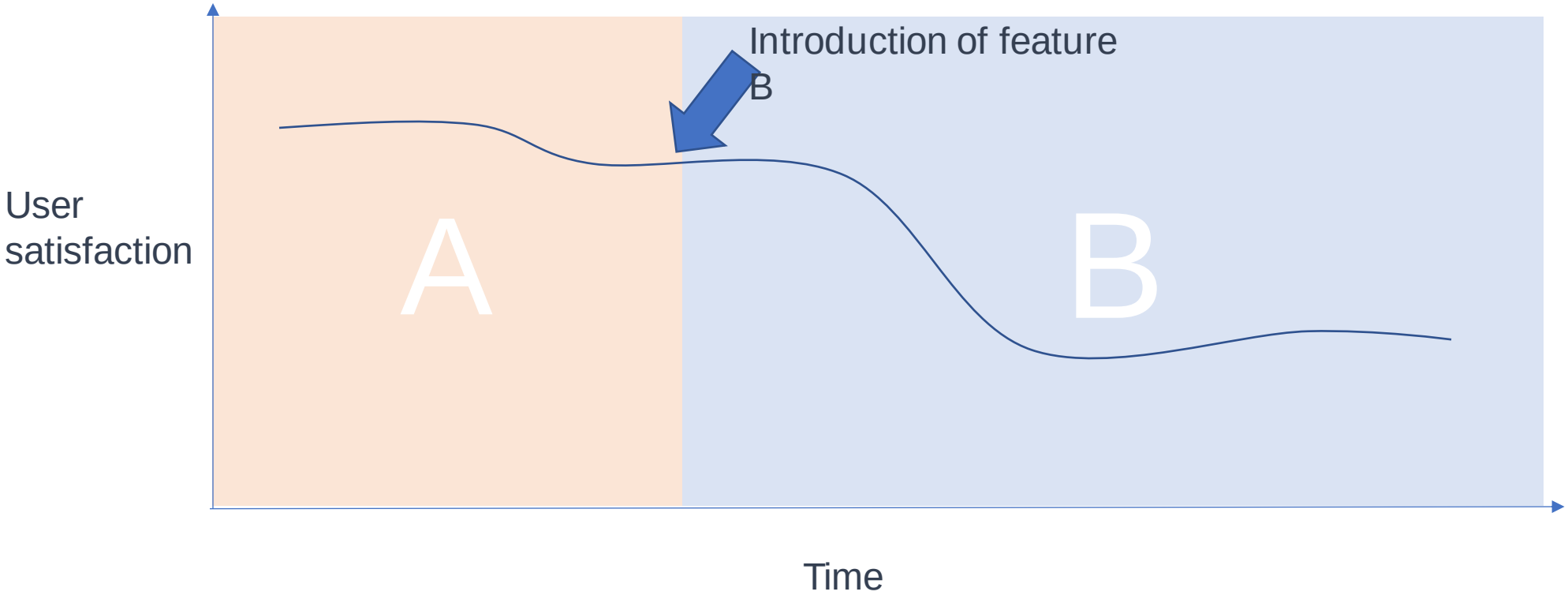
Recap

- We randomly split our population into treatment and control.
- We run the experiment.
- We calculate the p-value of the observed difference between treatment and control.
- If the observed p-value is very low, it means that the outcome we observed would have been highly unlikely under the null, and therefore we claim that we have enough evidence to reject the null.
- If the observed difference is not statistically significant, that means that the observed outcome is fairly likely under the null, so we don't have enough evidence to reject the null.

A/B testing alternatives that are not as good

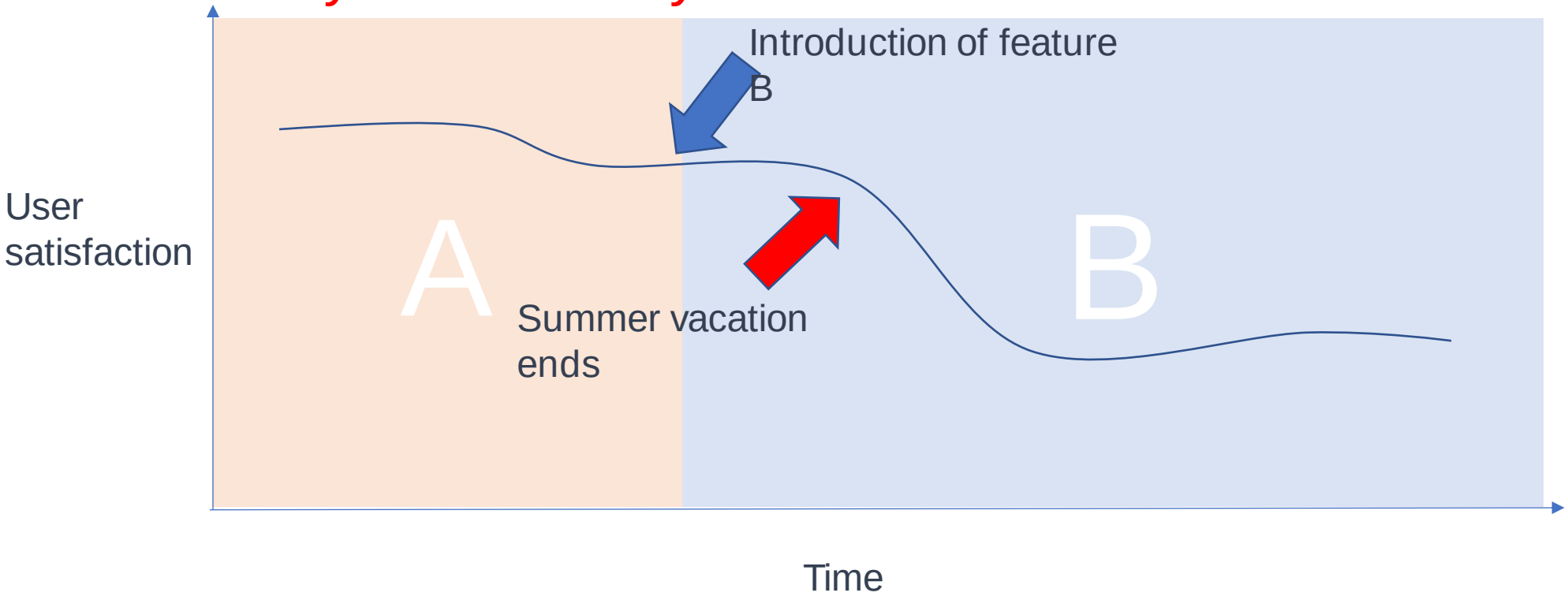
A/B testing alternatives that are not as good

Why not sequentially test each variant?



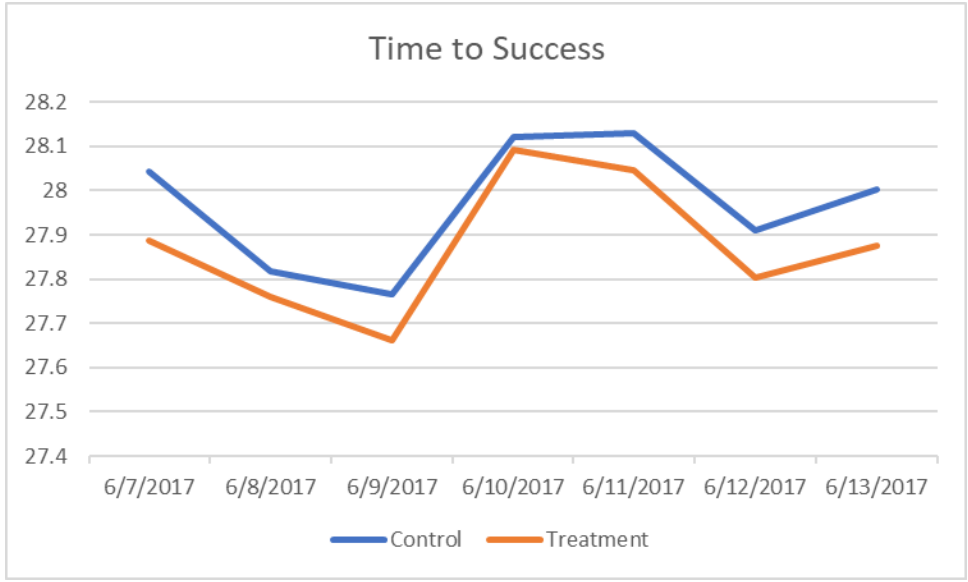
A/B testing alternatives that are not as good

Why not sequentially test each variant? This will not allow us to definitively attribute any difference to the treatment.



Advantage: Sensitivity!

- If you ran version A, then launched a change B on 6/11/2017, could you say if it was good/bad?
- If it were a controlled experiment, you could!



A/B testing alternatives that are not as good

Why are we **randomly** assigning individuals to be in each variant, A or B? Why not for example assign everyone in Paris to variant A, and everyone in London to variant B?

A/B testing alternatives that are not as good

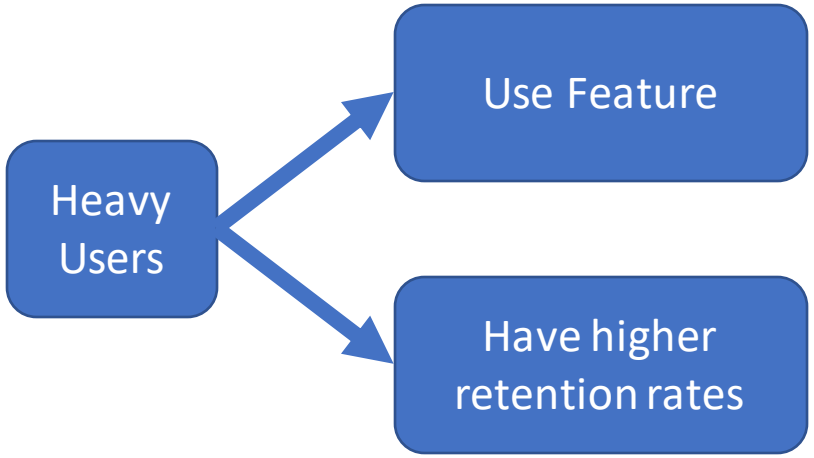
Why are we **randomly** assigning individuals to be in each variant, A or B? Why not for example assign everyone in Paris to variant A, and everyone in London to variant B?

Again, this does not allow us to definitively attribute any difference to the treatment.

A/B testing alternatives that are not as good

- Release you feature to everyone and observe differences between users using/not-using your feature.
- Let's say you observe:
  - 25% of new users that do NOT use your feature churn (stop using product 30 days later)
  - 10% of new users that use your feature churn
- Does your feature reduces churn?

Not necessarily – maybe people who use your feature are heavier users in the first place.



Generalizing experiment results

It is difficult to generalize experiments run in a specific market (e.g. en-US) to other markets (e.g. FR).

It is also tricky to generalize experiments run in a specific timeframe (e.g. Christmas) to other time frames.

Proceed with caution!

When in doubt, rerun your experiment in the new market / new timeframe

Stats section outline

A/B Testing basics

Hypothesis testing: Null and alternative hypotheses

Statistical Significance and P-values

False positives and False negatives

Power

(poor) Alternatives to A/B Testing

QUESTIONS?