A/B TESTING INCORRECTLY

Games Analytics Workshop August 2nd 2021

Overview

- Introduction
  - *Bio*
  - *Motivation & Goals*
- What's an AB Test
- Failing to define a user: Identification
- Failing to get the right result: Peeking
- Failing to lift off: Selection Bias
- Failing to Design: Interference
- Conclusion

About me (Professionally)

- Director of Data Science at The Meta (Kovaak)
- Assistant Professor of Data Science at USF
  - *2014-2020*
- Director of Analytics at Sega
  - *2014-2015*
- Director of Analytics and User Acquisition at TinyCo
  - *2011-2014*
- Senior Consultant Bates White
  - *2002-2006*

Some Games I've worked on

About me Academically

- PhD at UCLA in Management (2012)
- Masters in Economics at UC Davis (2007)
- BA in Applied Math/Statistics at UC Berkeley (2002)

Motivation

- Currently building an AB-testing system
- I think we (as academics) tend to “gloss over” a lot of the implementation details which can have an outsized effect on actual performance
- So I’m going to cover a few, relatively random, things

Goals of the talk

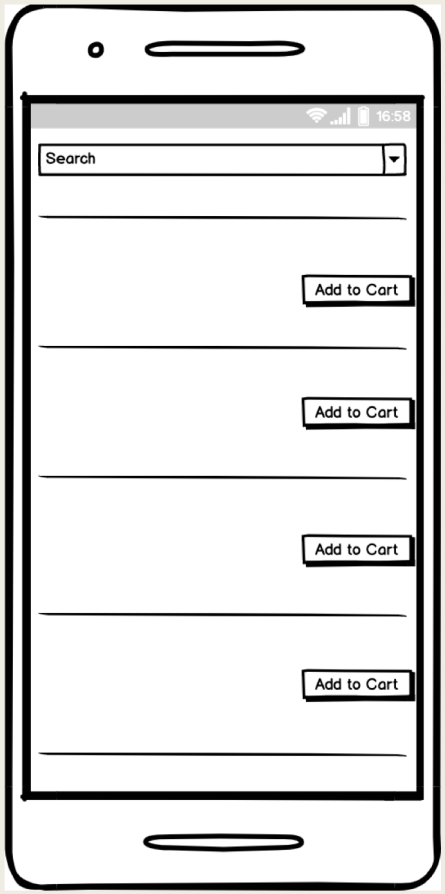
- Critical Success Factors of the talk:
  1. *Learn something new*
  2. *Increase Interest in Experimentation*
  3. *Present Solutions*
     - Tend to be non-technical
       1. *Organizational momentum*
       2. *Presentation*

WHAT'S AN A/B TEST

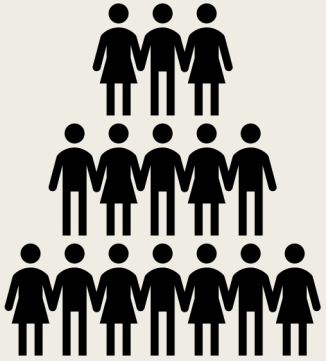
A/B Testing: What?

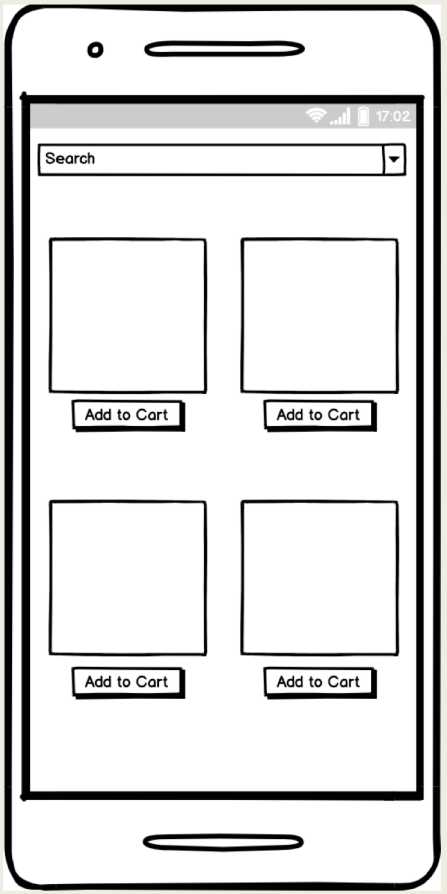
- An A/B test is a **controlled experiment**
- Interest lies in determining how some metric of interest is **causally** related to one or more factors
- Different levels of these factors define two or more **experimental conditions**
- Experimental **units** are **randomly** assigned to these conditions
- **Randomization principle:**
  - Random assignment ensures that users in different conditions will be homogenous and the only collective difference among them is the fact that they're in different conditions.
  - So any difference observed among the conditions should be due only that which the experimenter is controlling

A/B Testing: What?



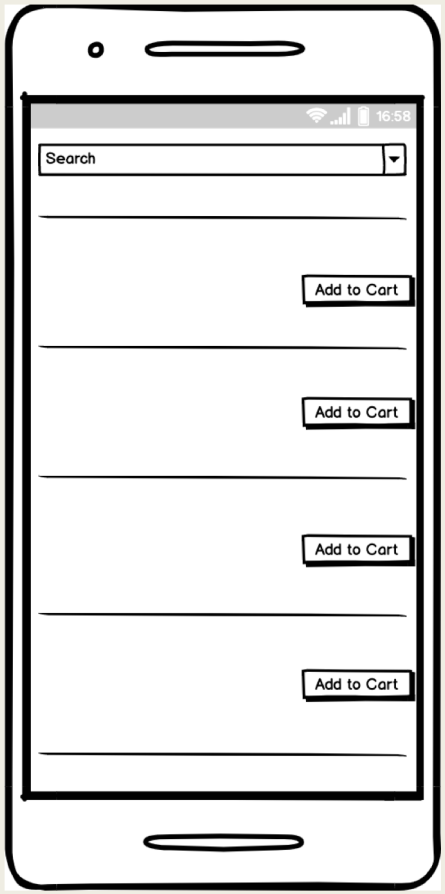
A





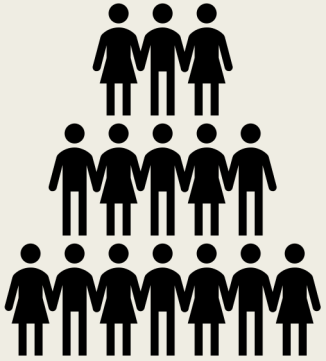
B

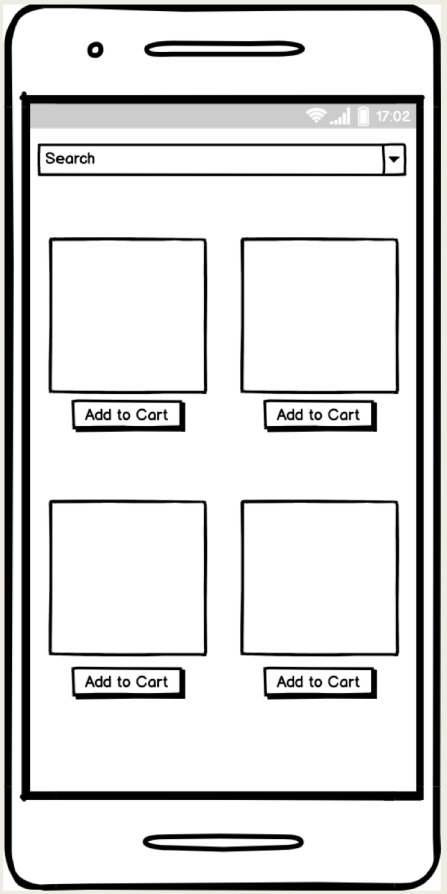
A/B Testing: What?



5% Pay

A





2% Pay

B

A/B Testing: Statistics

- After running the experiment we use Statistics to try to determine if our results:
  - Are the result of Randomness
  - Are the result of a more fundamental truth
- Is the difference between 5% and 2% payers “real”

FAILING TO DEFINE A USER

User identification

- What is a user?
- Identification strategy:
  - *Software based identification*
    - Web cookie, file on the hard drive
  - *Hardware based identification*
    - Serial numbers (IDFV on mobile)
  - *Required login*
    - 3rd party (Steam, Epic)
    - 1st party (roll your own)

Most common

- Some Combination:
  - *No login required until a threshold achieved*
  - *Login “optional” but gives additional features*
  - *Different accounts that may/may not be linkable*
    - Cross-play
    - Steam vs. non-Steam login

Who cares?

- A “User” may experience multiple experimental treatments:
  - *Nick has no login to the game and is assigned to treatment group A*
  - *Nick creates an account and is assigned to group treatment B*
  - *Nick plays via Steam (logged in via Steam) and on Mobile (logged in via FB). Accounts not linked, one in group A and one in group B.*
  - *Nick sign up to the news letter with two different email addresses and look for the best deals in any A/B situation*

Result

- For knife-edge conclusions, a small percentage of users being misidentified can swing the results
- Systematic misidentification can skew a test completely
  - *3% of users cross-play, but 100% of cross-play users experience treatment B and they all do X*

The Meta

- Two “user” identification systems (Steam and “The Meta” login)
- This is a many-to-many match
- Defining tests on this is “difficult” without considering a number of edge cases

Solution

1. Well defined tests:
   - *Avoid obvious identification issues*
   - *Focus on subsets of users (post account creation, users who do not cross play, etc.)*
2. Rely on organizational momentum:
   - *Define a framework for testing which avoids these issues*
   - *Organizations tend to have momentum. Once it's done once, rely on "This is how we do it."*

FAILING TO GET THE RIGHT RESULTS

What is peeking?

- **Peeking** is the phenomenon whereby you regularly check the results of the experiment before it finishes
- Peeking can be a good thing!
  - *Make sure the experiment is not negatively impacting other important metrics*
  - *Verify experiment is running correctly*
- The problem arises when, as a result of peeking, you decide to end the experiment early

Example

- I set up my experiment:
  - *I need 1,000 users in both the Treatment and Control group*
- On the first day, I look at my data:

	Observations Collected	Conversion Rate
Treatment	150	10%
Control	150	5%

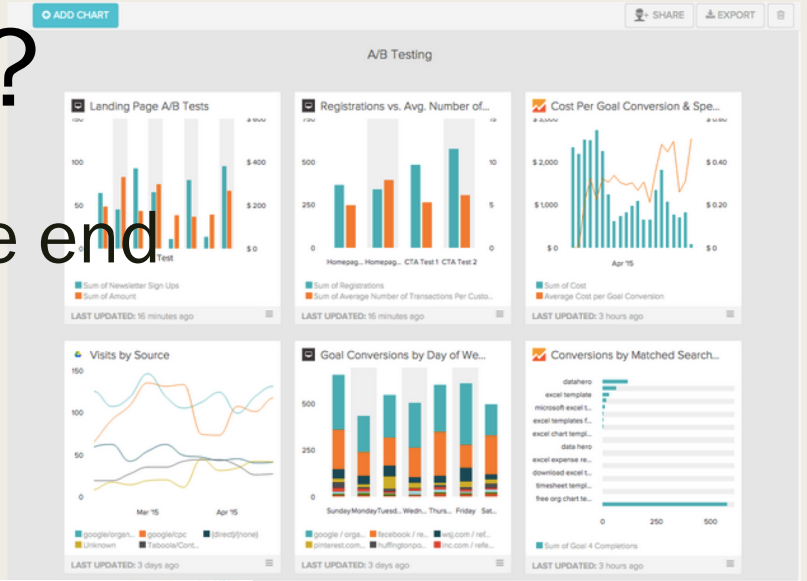
  - *The conversion rate in the Treatment group is twice that of the Control. Should we stop the test?*

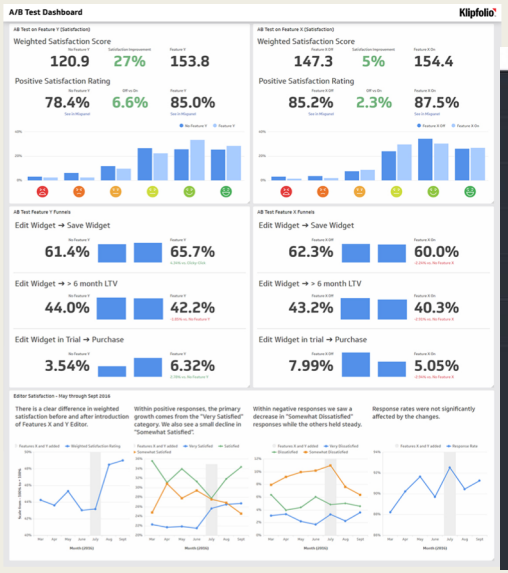
Why is this a problem?

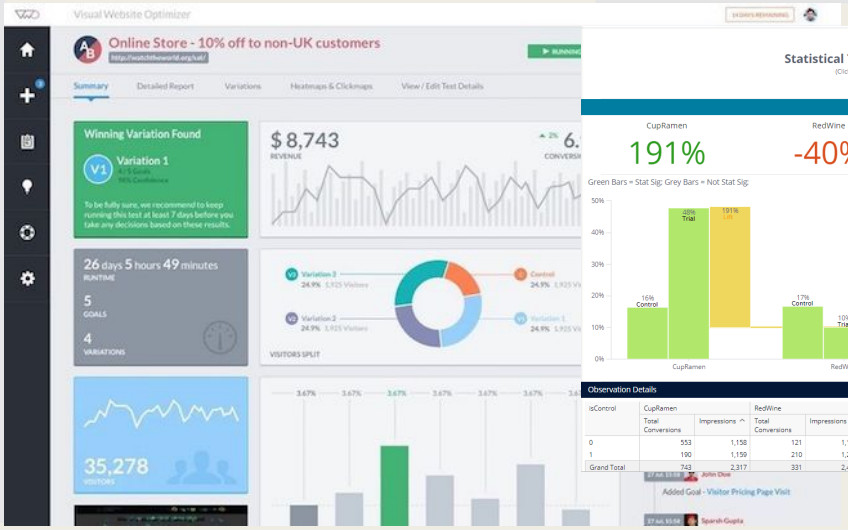
- By stopping the experiment early you have not observed enough data to be confident in your conclusion
  - Just because the results suggest a winner or a significant difference at one point in time does not mean that the results won't change as more data is collected

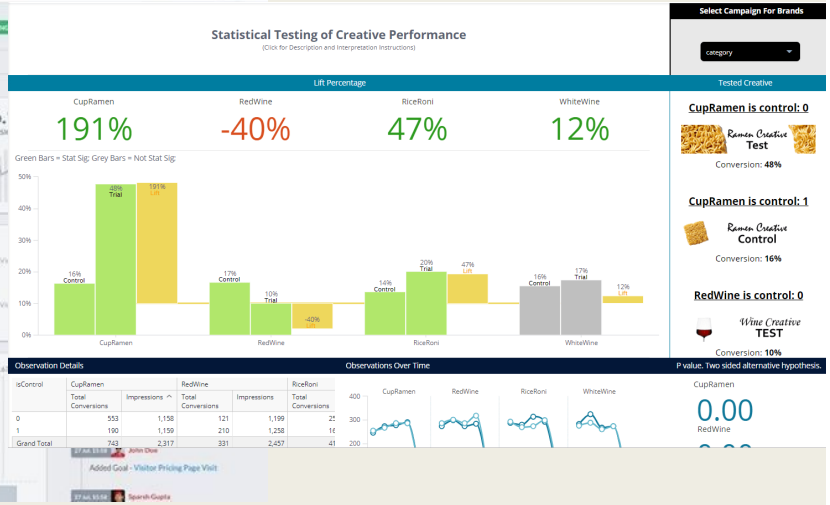
Why is this a problem?

- Okay, so just run the test to the end
- But it's hard
- So many dashboards









Why is this a problem?

- When you stop the experiment you are rejecting the null hypothesis
- Which means you might be making a Type I error
- And by stopping the experiment early the chances you make a Type I error are **much higher** than the prespecified statistical significance (α)

Why is this a problem?

Illustrative Simulation

- $n_C = n_T = 1{,}000$ data points are drawn independently from the $\mathrm{N}(0,1)$ distribution
- The observations are used to perform a Z-test of

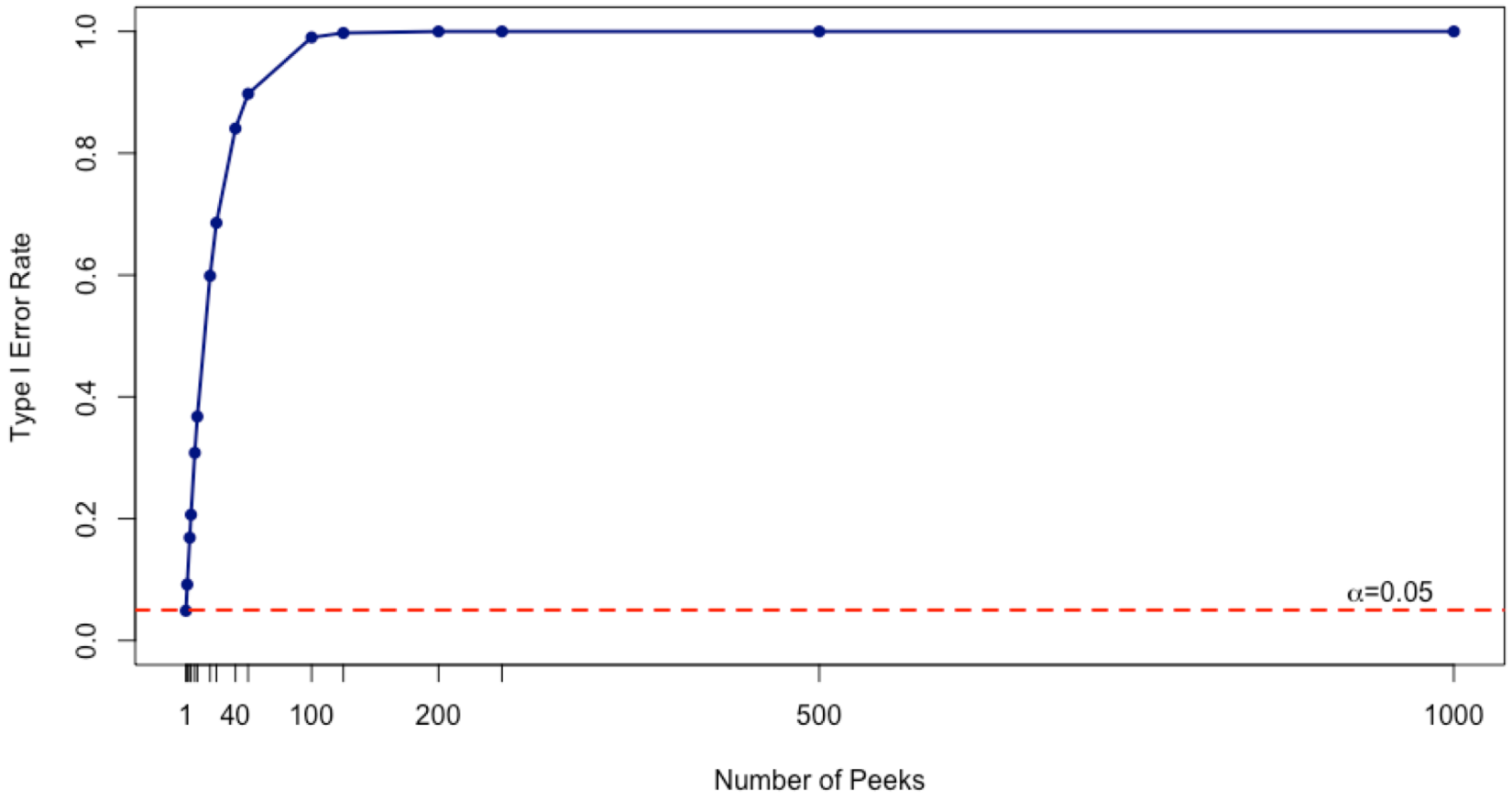
$$\mathrm{H}_0: \theta_C \leq \theta_T \text{ vs. } \mathrm{H}_\mathrm{A}: \theta_C > \theta_T$$

- Because $\theta_C = \theta_T = 0$ we should not reject H_0 very often (no more than $\alpha \times 100\%$ of the time)

Why is this a problem?

- To study the consequences of peeking, we peek – **and end the experiment if a significant result is indicated** – at regular intervals
- Repeat this 10,000 times
- The Type I Error rate is the fraction of the 10,000 simulations that an experiment is ended prematurely

Why is this a problem?



What is the solution?

- **Sequential Testing**
  - An analysis method where the sample size is not fixed a priori
  - Data are accumulated and analyzed sequentially until a stopping rule is met
  - Stopping rule is based on α and β-spending functions
  - Resulting lift estimates need to be bias-corrected
  - More complex to implement

What is the solution?

- Avoid having non-sophisticated users end tests early
  - *Presentation layer:*
    - Modify presentation with explicit warnings
    - Hide results
- Require test to have a minimum number of units (as part of the the design)

	Observations Collected	PERCENT COMPLETE	Conversion Rate
Treatment	150	30%	10%
Control	150	30%	5%

FAILING TO LIFT OFF

Hmmmm… ?

- You run a test
- Treatment effect has 5% higher revenue than control
- So you make the change, but revenue only increases by 2%
- This happens on **every** test.

	Control	Treatment	Estimated Diff	Actual Diff
Test #1	17%	12%	5%	2%
Test #2	5%	2%	3%	1.8%
Test #3	7%	3%	4%	3.2%
Test #4	9%	4.5%	4.5%	4%
Test #5	8%	6%	2%	1.4%

So what is this bias?

- Let $\delta = \theta_C - \theta_T$ be the true unknown **treatment effect** (aka: **lift**)
- This is estimated by:

$$\hat{\delta} = \hat{\theta}_C - \hat{\theta}_T = \bar{X} - \bar{Y}$$

- It is well known that this is an **unbiased estimate**:

$$\mathrm{E}[\bar{X} - \bar{Y}] = \theta_C - \theta_T = \delta$$

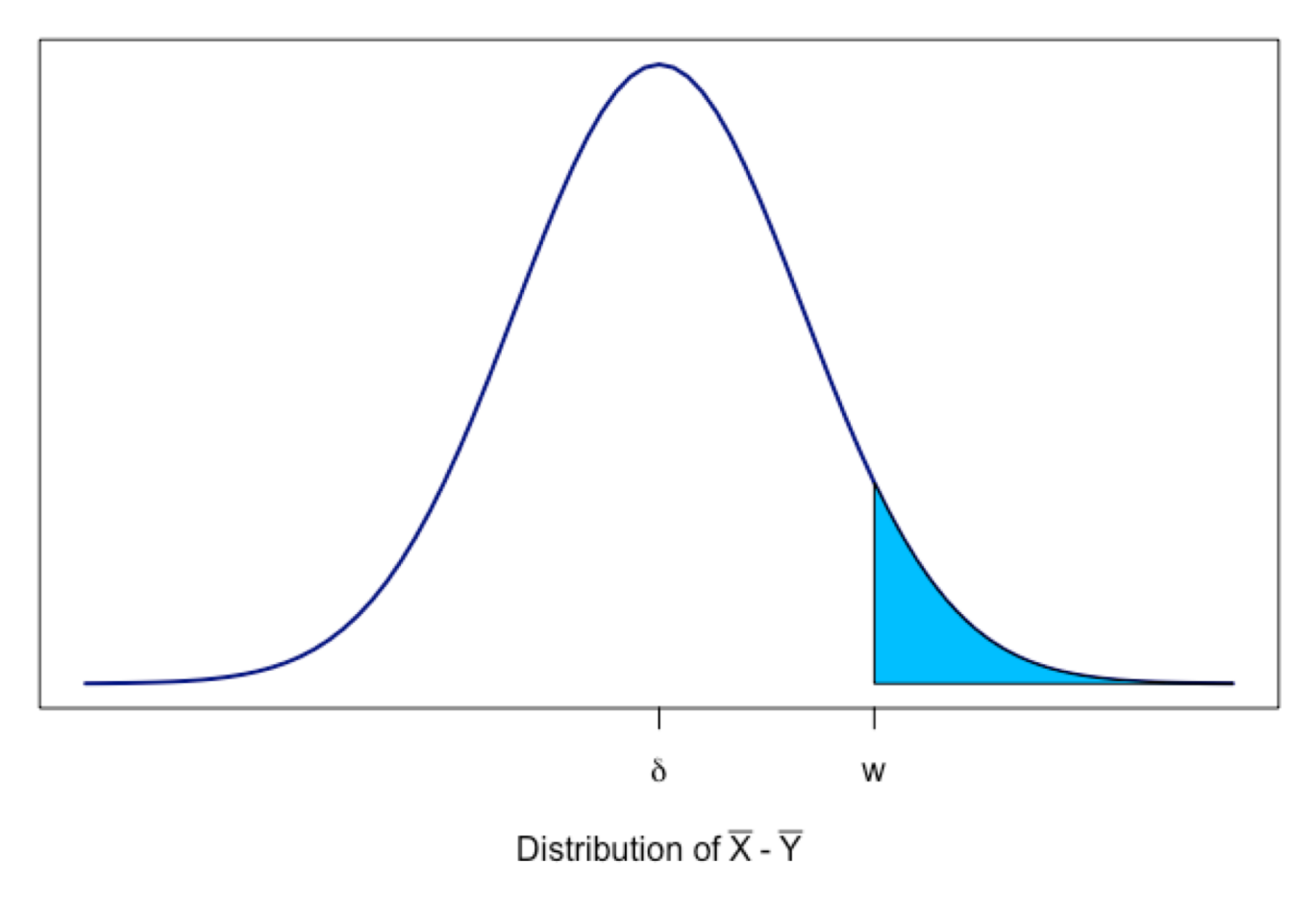
So where's the problem?

Problem:

- This isn't how we estimate lift in practice
- In practice lift is only ever estimated **if the null hypothesis is rejected**
- For illustration assume we're testing the hypothesis

$$\mathrm{H}_0: \theta_C \leq \theta_T \text{ vs. } \mathrm{H}_\mathrm{A}: \theta_C > \theta_T$$

So where's the problem?



So where's the problem?

Problem:

- So what we're actually estimating in practice is

$$\mathrm{E}[\bar{X} - \bar{Y} | \bar{X} - \bar{Y} \geq w]$$

not

$$\mathrm{E}[\bar{X} - \bar{Y}]$$

Note: $w = \sigma \times z^*$ where $\sigma = \mathrm{SD}[\bar{X} - \bar{Y}]$ and z^* is the appropriate critical value of $\mathrm{N}(0,1)$ determined by α

So where's the problem?

Problem:

When

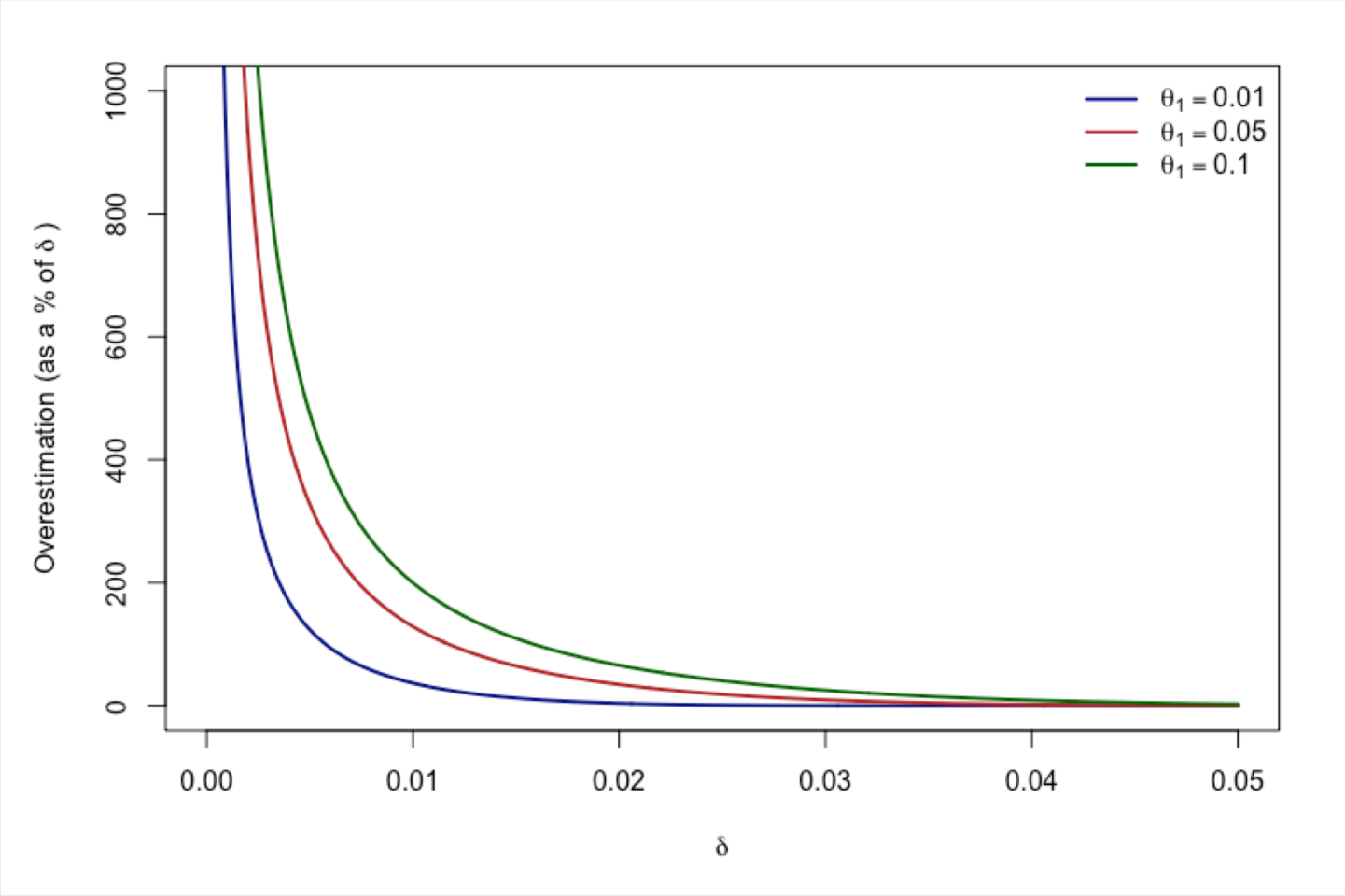
$$\mathrm{H}_0: \theta_C \leq \theta_T \text{ vs. } \mathrm{H}_\mathrm{A}: \theta_C > \theta_T$$

then

$$\mathrm{E}[\bar{X} - \bar{Y} | \bar{X} - \bar{Y} \geq w] = \delta + \sigma \frac{\phi\left(\frac{w-\delta}{\sigma}\right)}{1 - \Phi\left(\frac{w-\delta}{\sigma}\right)}$$

which is strictly greater than δ

How big a problem is this?



So what can we do?

- Accept that the lift estimated from your experiment is an overestimate
- Sadly, the statistics behind estimating this are difficult so can't just "undo" it
- **Presentation** layer:
  - *Add "Max Difference" or add an "Estimated" lift to the presentation.*

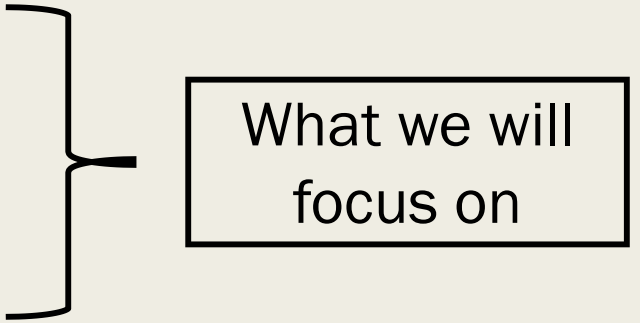
FAILING TO DESIGN

What is Interference?

- Problems of interference occur when your experimental conditions become **contaminated**
- This typically means that the **Stable Unit Treatment Value Assumption (SUTVA)** has been violated
  - SUTVA: The outcome observed on one unit should be unaffected by the treatment assignment of other units
- Your experimental conditions are no longer independent

What is Interference?

- Interference/contamination can happen for a variety of reasons:
  - Unit unidentifiability (spoken about before)
  - Colliding experiments (be careful)
  - **Network interference**
    - **Intra contamination**
    - **Inter contamination**



Network Interference

- What if my experiment effects other users and, in turn, modifies their behavior?
- Facebook does an A/B test on “People you May Know”
  - Control group sees “as is”
  - Treatment sees “new flow”
- If treatment causes more friend requests, which then increase friend requests for control users, then my lift estimates will be incorrect
- What if my users directly communicate to each other about test conditions?
  - *This will change behavior (test/control group may be unhappy and do something negative)*

Network Interference

- Academically, this is solved by modeling as a network/graph problem
  - *Many assumptions*
  - *Specific knowledge / parameter estimates, etc.*
- "Real world"
  - *Tend to either **ignore** or design around (geo-fencing + light modeling)*

How bad is ignoring?

- Given that this “only happens a little” in my product, how much does it matter?
- At the Meta, we don’t expect this to be too much of an issue (outside of leaderboards our product does not have too much of a social element)
- Go over two models and see what happens
  - *Correlation between treatment groups*
  - *Correlation within treatment groups*

Between Treatment Groups

- Standard T-test of significance
- Users from one treatment group effect the outcome of the other treatment group
- Specifically:

$$COV\left(Y_{i,A}, Y_{j,B}\right) = \begin{cases} 0 \; otherwise \\ \lambda\sigma^2 \;\; if \; i = j \end{cases}$$

- If user #1 in treatment A does something => effects the outcome of user #1 in treatment B

Between Treatment Groups

- Our T Statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

- Without Interference: $\mathrm{VAR}[T] = 1$
- With Interference: $\mathrm{VAR}[T] = 1 - \lambda$

Conclusion (correlation between groups)

- Since λ can be positive or negative, so unless we know it's value it's difficult to conduct a test.
- This is a really simple, well specified case.
- One nice thing – in this case increasing our sample size will naturally help things:
  - *While it doesn't solve the interference it does spread out our estimators ($\overline{Y}_1, \overline{Y}_2$) making the interference less costly.*

What about correlation within groups?

- Once again, standard T-test of significance
- Users from one treatment group effect the outcome of the other treatment group
- Specifically:

$$COV(Y_{i,j}, Y_{k,j}) = \begin{cases} \lambda\sigma^2 \; if \; i \neq k \\ \sigma^2 \; if \; i = k \end{cases}$$

- Note that $j \in (A, B)$. We assume zero correlation between groups

Within Treatment Groups

- Our T Statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

- Without Interference: $\mathrm{VAR}[T] = 1$
- With Interference: $\mathrm{VAR}[T] = 1 + (n-1)\,\lambda$

Conclusion (correlation between groups)

- Since λ can be positive or negative, so unless we know it's value it's difficult to conduct a test.
- This is a really simple, well specified case.
- Increasing the sample size in this case makes interference **worse**.

How to handle interference

- Academically:
  - *Network modelling*
  - *Econometric (but-for analysis)*
  - *Matched-pairs experimental design (geo-fencing)*
- All of these are difficult and costly (man power)

How to handle interference

- So… I'm still not sure
- In the short term, try to avoid experimental situations that might make it worse:
  - *Social offers*
  - *Leaderboards functionality*
  - *Tie-in testing (adding social logins, rewarding for streaming, etc.)*
- In particular – probably (technology side), make these types of test costly to do

CONCLUSION

Conclusion

- While a lot of experimentation is well-known, the details of implementation are hard.
- Concepts like peeking, estimating lift and interference are “solved” academically
- Putting that solution into practice is incredibly difficult and often results in leveraging organizational/UI/UX solutions to achieve.