

# A/B Testing

Hypothesis Testing

# OUTLINE

- Hypothesis Testing
- A/B Testing
  - Use of a simple business example to set up an A/B test end to end.
  - How to design a task, choose metrics, and analyze the results

# Statistical Hypothesis Testing

Assessing evidence provided by the data in favor of or against each hypothesis about the population.

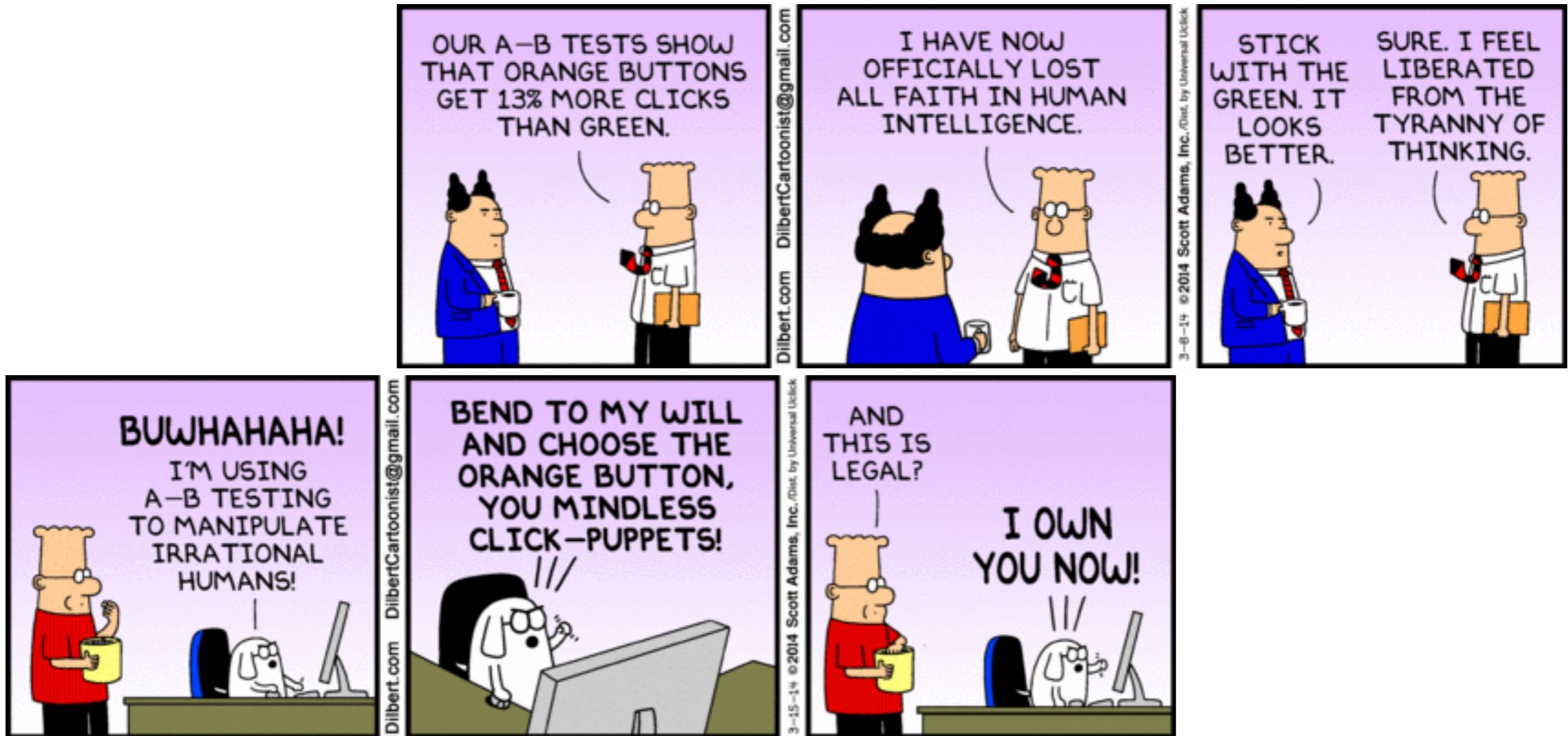
Hypothesis testing is one of the most important inferential tools when it comes to the application of statistics to real life problems.

It's used when we need to make decision concerning populations on the basis of only sample information.

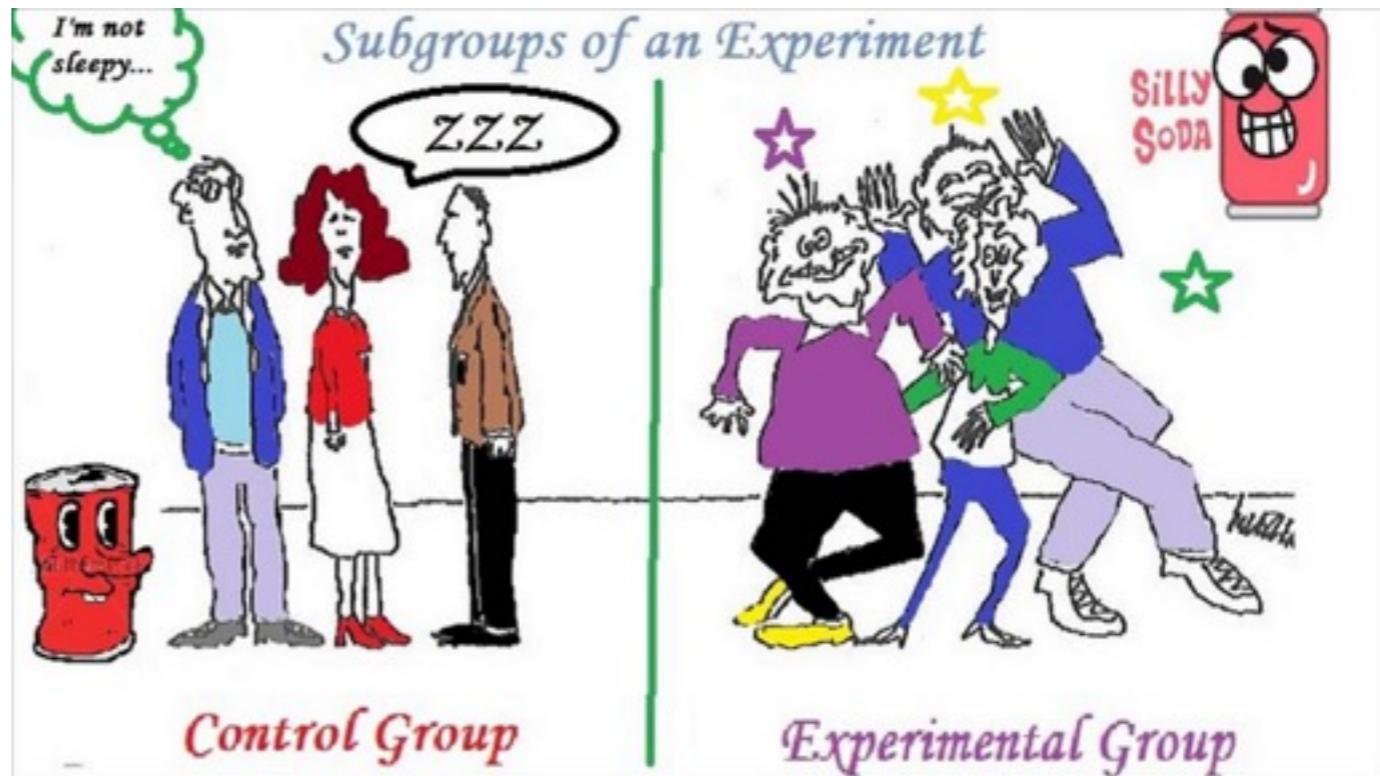
# Steps in Hypothesis Testing

- Specify the null ( $H_0$ ) and the alternative ( $H_A$ ) hypotheses
- Choose a sample
- Assess the evidence
- Draw conclusions

# A/B Testing



- **Control group**
- **Experimental group**



A/B tests allow you to determine scientifically how to optimize a website or a mobile app by trying out possible changes and seeing what performs better with your users.

# Tools you can use



mixpanel

 kissmetrics



optimizely

# Examples of A/B testing in the industry



- Amazon:
  - Personalized recommendations



- Google:
  - Tested 41 different shades of blue



- LinkedIn:
  - Tested whether to use the top slot of an user's stream for news articles or an encouragement to add more contacts



- Amazon & Google:
  - Every 100ms increase in page load time decreased revenues by 1%

# Not good for...

- Can't tell you if you are missing something...
- Testing new experiences
  - Change aversion
  - Novelty effect
- Time

# When can you use A/B Testing?

- Online shopping company:  
Is my site complete?
- Add premium service
- Movie recommendation site:  
New ranking algorithm
- Change backend-page  
Load time, results, what users see

# When can you use A/B Testing?

- Online shopping company:  
Is my site complete?
- Add premium service
- Movie recommendation site:  
New ranking algorithm
- Change backend-page  
Load time, results, etc

**Can't answer in general**

**You can't fully test:  
Could gather info**

**Great - clear control and  
clear metrics**

**Good if you have the  
computing power to run  
both simultaneously**

# When can you use A/B Testing?

- Website selling cars: will a change increase repeat customers or referrals?
- Update brand, including main logo
- Test layout of initial page

# When can you use A/B Testing?

- Website selling cars: will a change increase repeat customers or referrals?
- Update brand, including main logo
- Test layout of initial page

**Too long and don't have data**

**Very emotional...**

**Clear control and clear metrics**

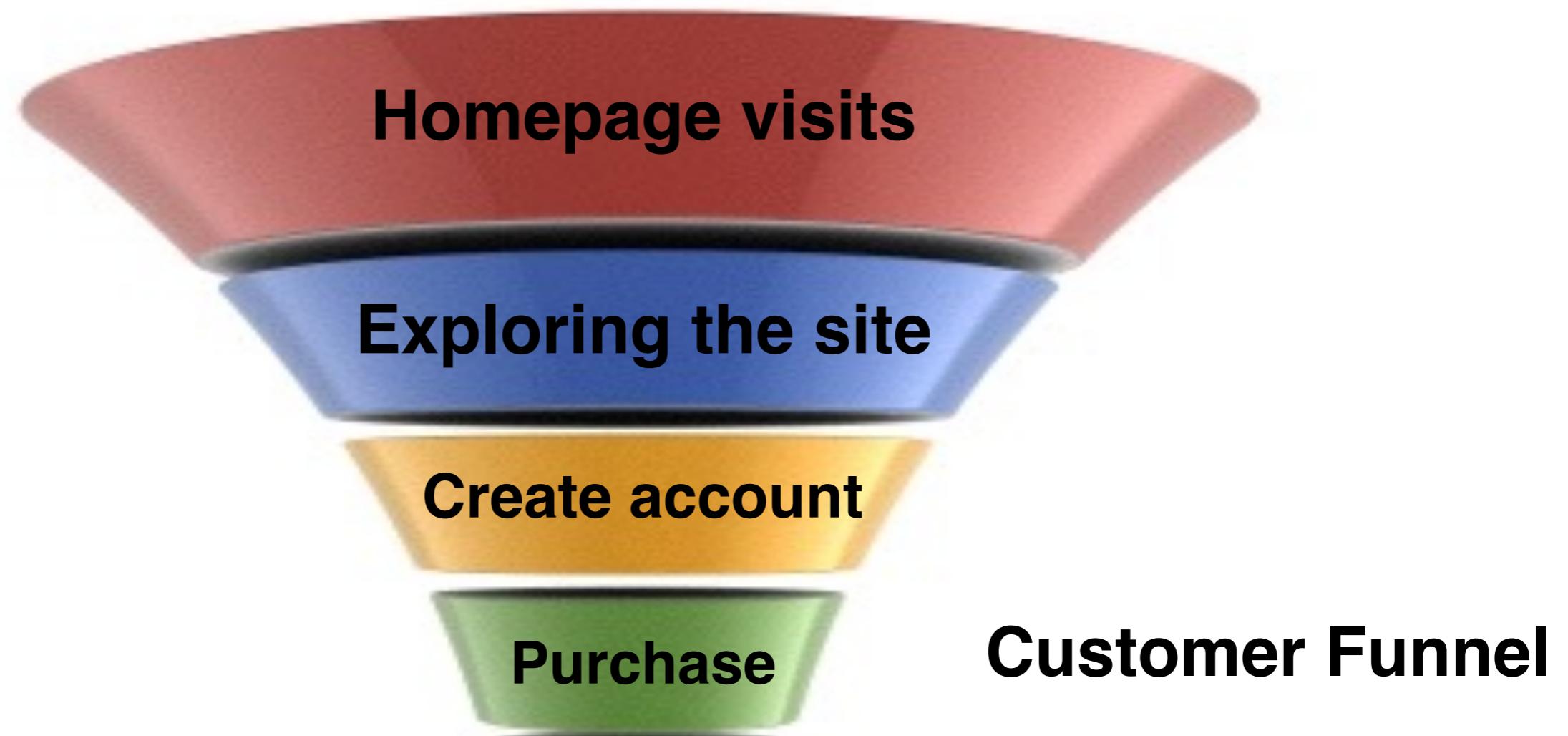
# Other techniques

- Retrospective analysis
- Prospective analysis
- User experience research
- Focus groups
- Survey
- Human evaluation

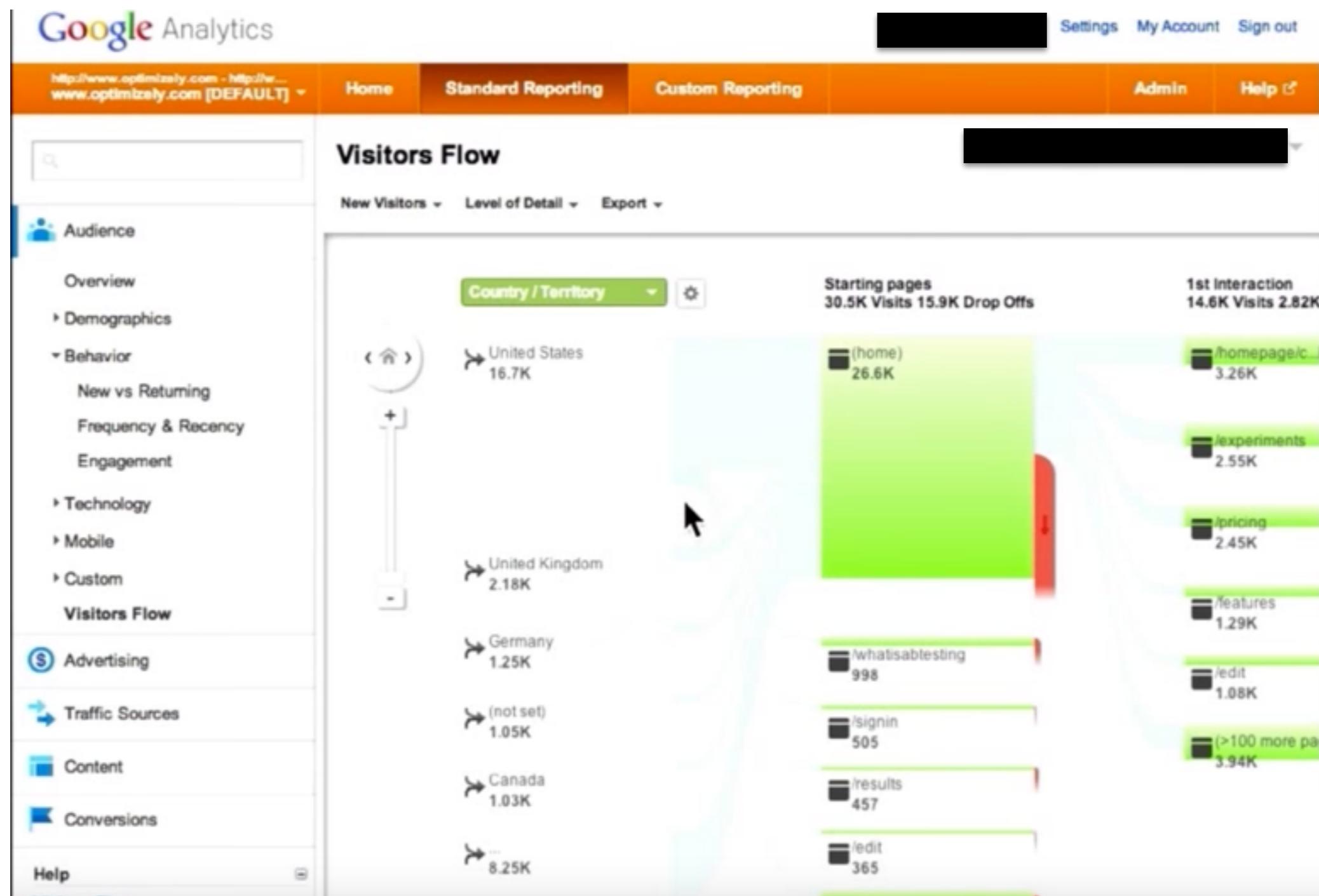
# History of A/B Testing

- Agricultural field
- Medicine

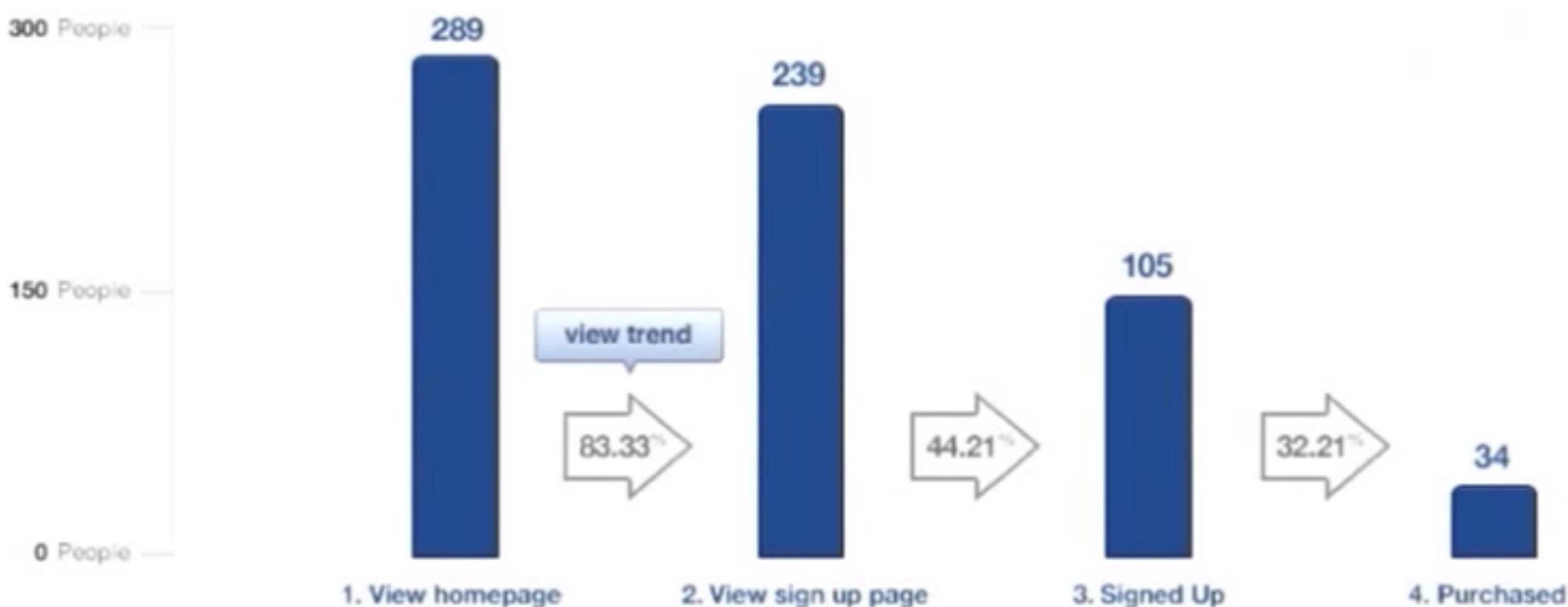
# **The Iconic:** Online company selling clothes

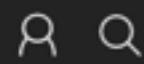


# Where to look for this?



# mixpanel





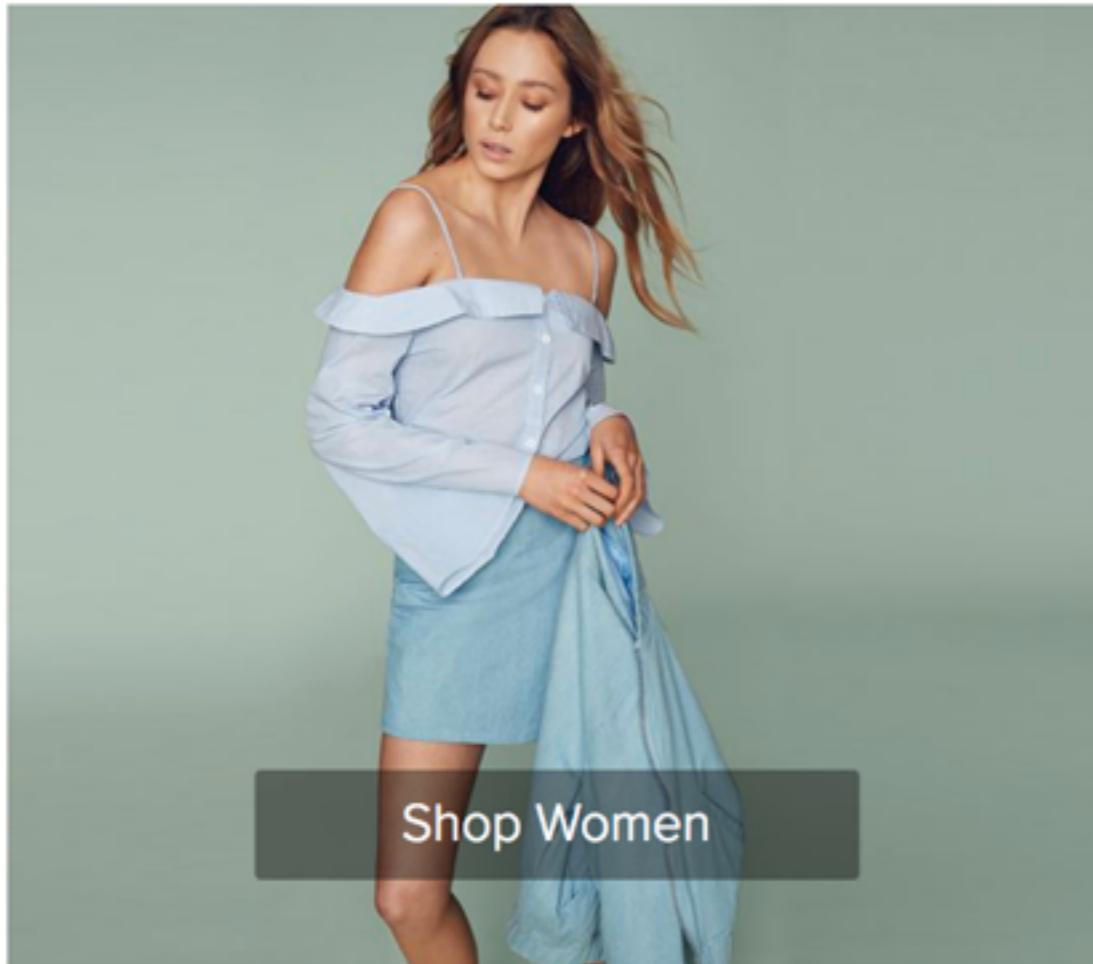
THE ICONIC

CLOTHING SHOES ACCESSORIES SPORT BRANDS DISCOVER



FREE NEXT DAY DELIVERY OVER \$70

FREE RETURNS FOR 100 DAYS



Shop Women

# Refining the Hypothesis

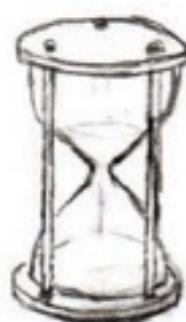
Initial Hypothesis: Changing the “Shop Women” button from **grey** to **purple** will increase how many people explore women clothing.

Which metric to use?

Define a quantifiable success metric!

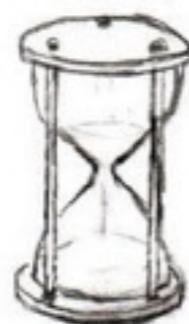
# Which metric to use

- Total number of purchases completed



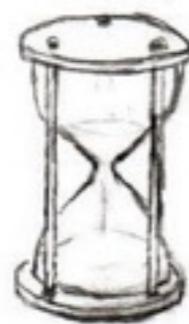
# Which metric to use

- Total number of purchases completed
- Number of clicks

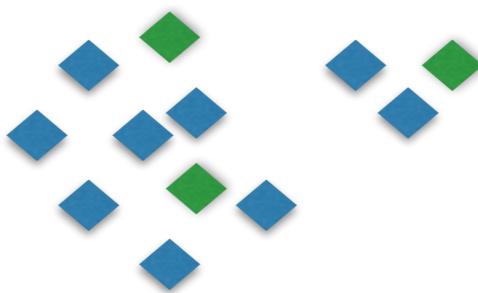


# Which metric to use

- Total number of purchases completed
- Number of clicks

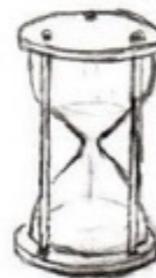


**Group 1**      **Group 2**



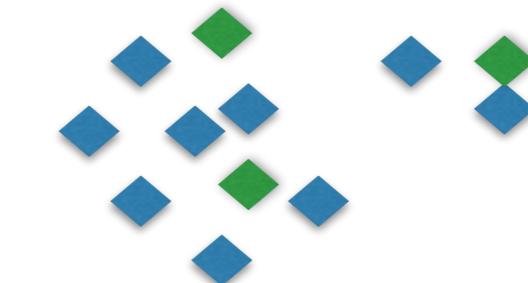
# Which metric to use

- Total number of purchases completed



**Group 1**      **Group 2**

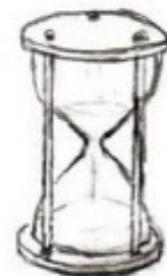
- Number of clicks
- $$\frac{\text{Number of clicks}}{\text{Number of page views}}$$



Click-Through-Rate - CTR

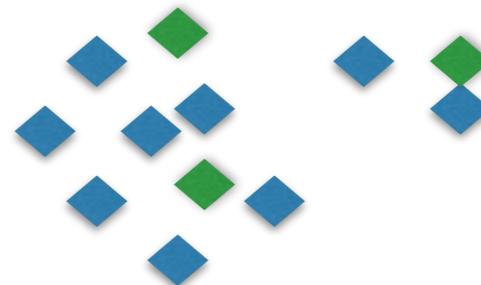
# Which metric to use

- Total number of purchases completed



- Number of clicks
- $\frac{\text{Number of clicks}}{\text{Number of page views}}$

**Group 1**      **Group 2**



Click-Through-Rate - CTR

$$\frac{\text{Unique visitors who click}}{\text{Unique visitors to page}}$$

Click-Through-Probability - CTP

# CTR vs CTP

0 clicks



5 clicks



**CTR= 2.5**

**CTP= 0.5**

# CTR vs. CTP

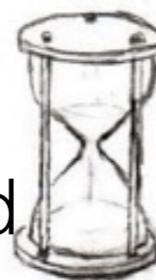
Generally speaking:

- Use “rate” when you want to measure usability of the site
- Use “probability” when you want to measure the total impact.

# Which Metric to use



- Total number of purchases completed

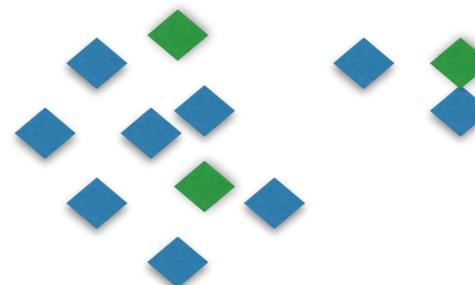


- Number of clicks



- $\frac{\text{Number of clicks}}{\text{Number of page views}}$

**Group 1**      **Group 2**



Click-Through-Rate - CTR



$\frac{\text{Unique visitors who click}}{\text{Unique visitors to page}}$

Click-Through-Probability - CTP

# Refining the Hypothesis cont'd

Initial Hypothesis: Changing the “Shop Women” button from **grey** to **purple** will increase how many people explore women clothing.

Updated Hypothesis: Changing the “Shop Women” button from **grey** to **purple** will increase the **click-through-probability** of the button.

# Measurement of CTP



**1,000 visitors**



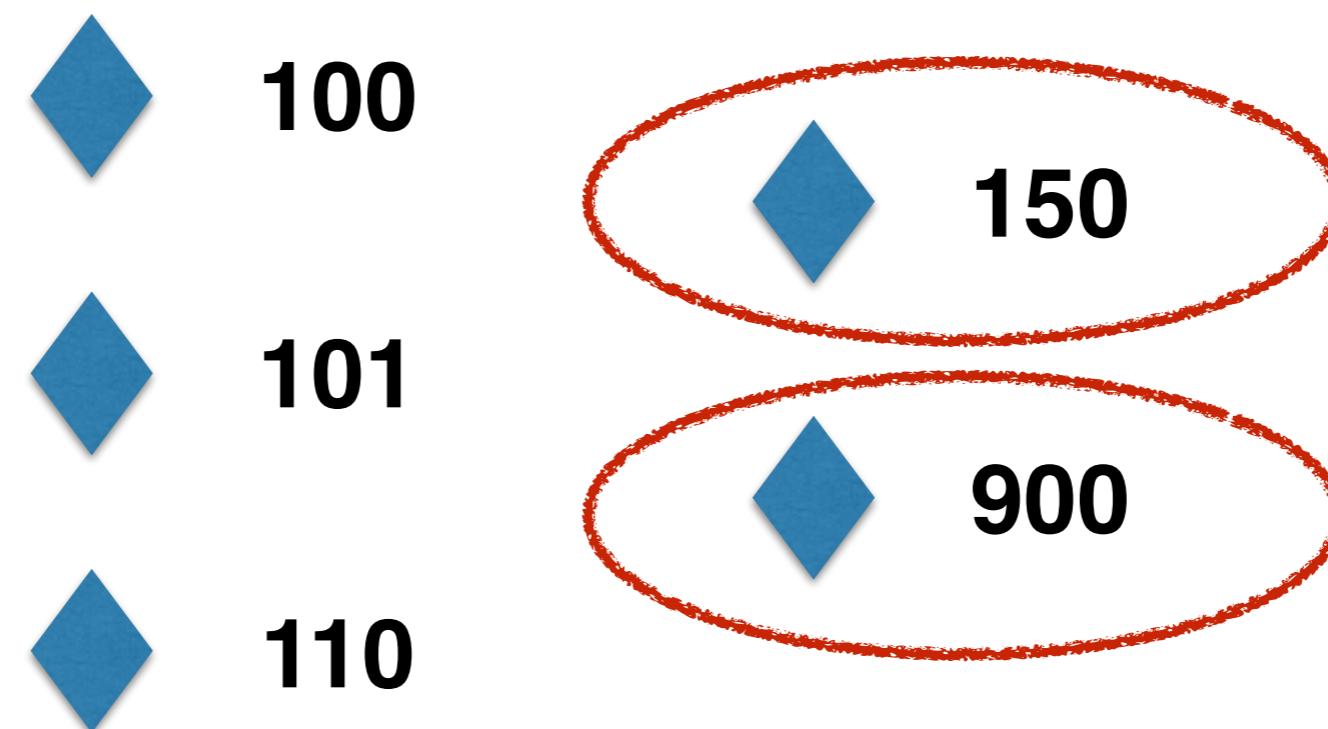
**100 unique clicks**

**Click-through-Probability = 10%**

**How sure would you be of this estimate?**

# Repeated measurement of CTP

Which results would surprise you if you repeated the measurement?



# Statistics Review

- What's the distribution?
  - Binomial Distribution



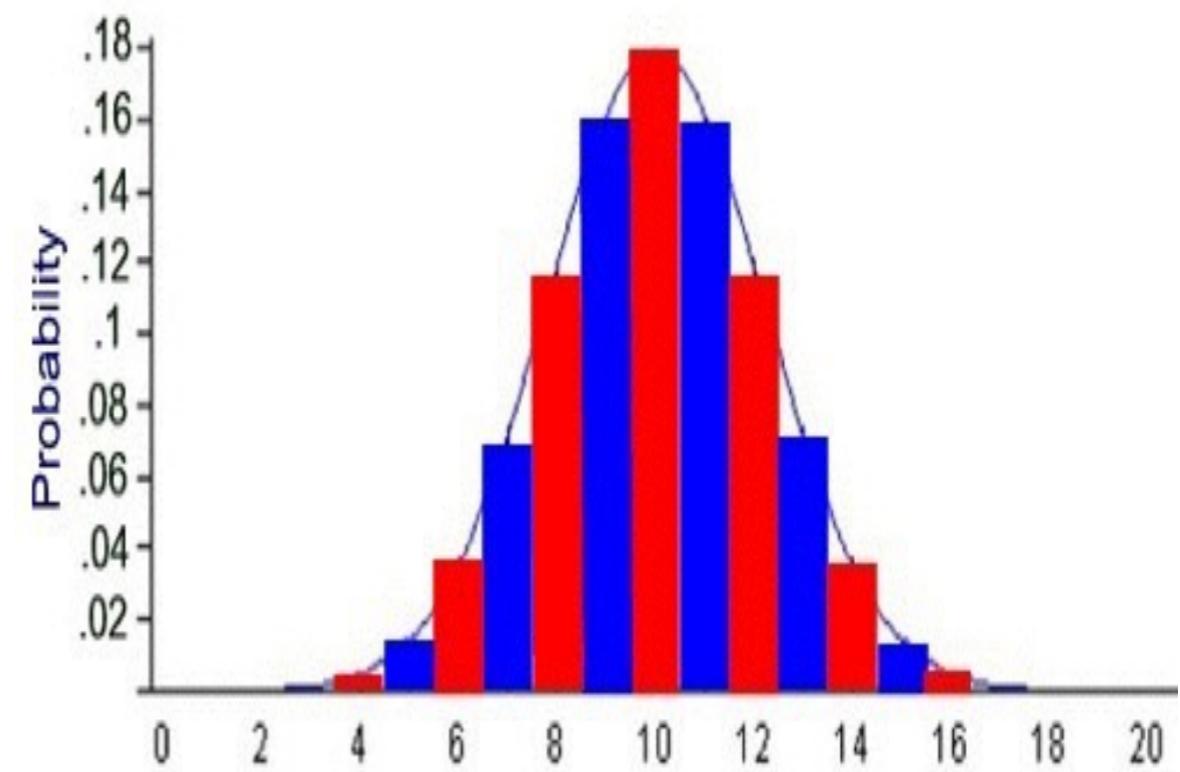
$p=0.5$   
Success = heads  
Failure= tails

$$N=20 \quad X=16$$

$$\hat{p} = \frac{16}{20} = \frac{4}{5} = 0.8$$

$$\text{mean} = p$$

$$\sigma = \sqrt{\frac{p(1-p)}{N}}$$



Flip the coin 20 times!

# Binomial Distribution

## 1. Two types of outcomes

- Success
- Failure

## 2. Independent events

## 3. Identical distribution

- $p$  must be the same for all

# Examples

- Drawing 20 cards from a shuffled deck:  
Outcomes: red and black
- Roll a die 50 times  
Outcomes: 6 or other
- Clicks on a search results page  
Outcomes: click or no click
- Purchase of items within a week  
Outcomes: Purchased or not

# Examples

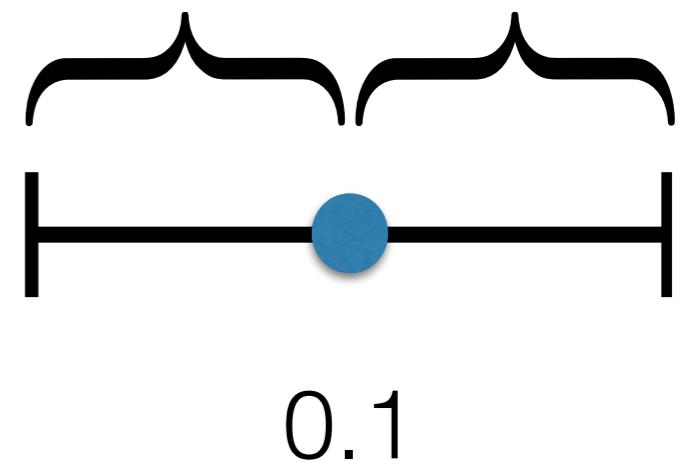
- Drawing 20 cards from a shuffled deck:  
Outcomes: red and black
  - Roll a die 50 times  
Outcomes: 6 or other
  - Clicks on a search results page  
Outcomes: click or no click
  - Purchase of items within a week  
Outcomes: Purchased or not
- Not independent**
- Independent**
- Not independent**
- Not independent**

# Confidence Interval

$$\hat{p} = \frac{X}{N} = \frac{\# \text{ users who clicked}}{\# \text{ users}}$$

$$\hat{p} = \frac{100}{1000} = 0.1$$

m= margin of error



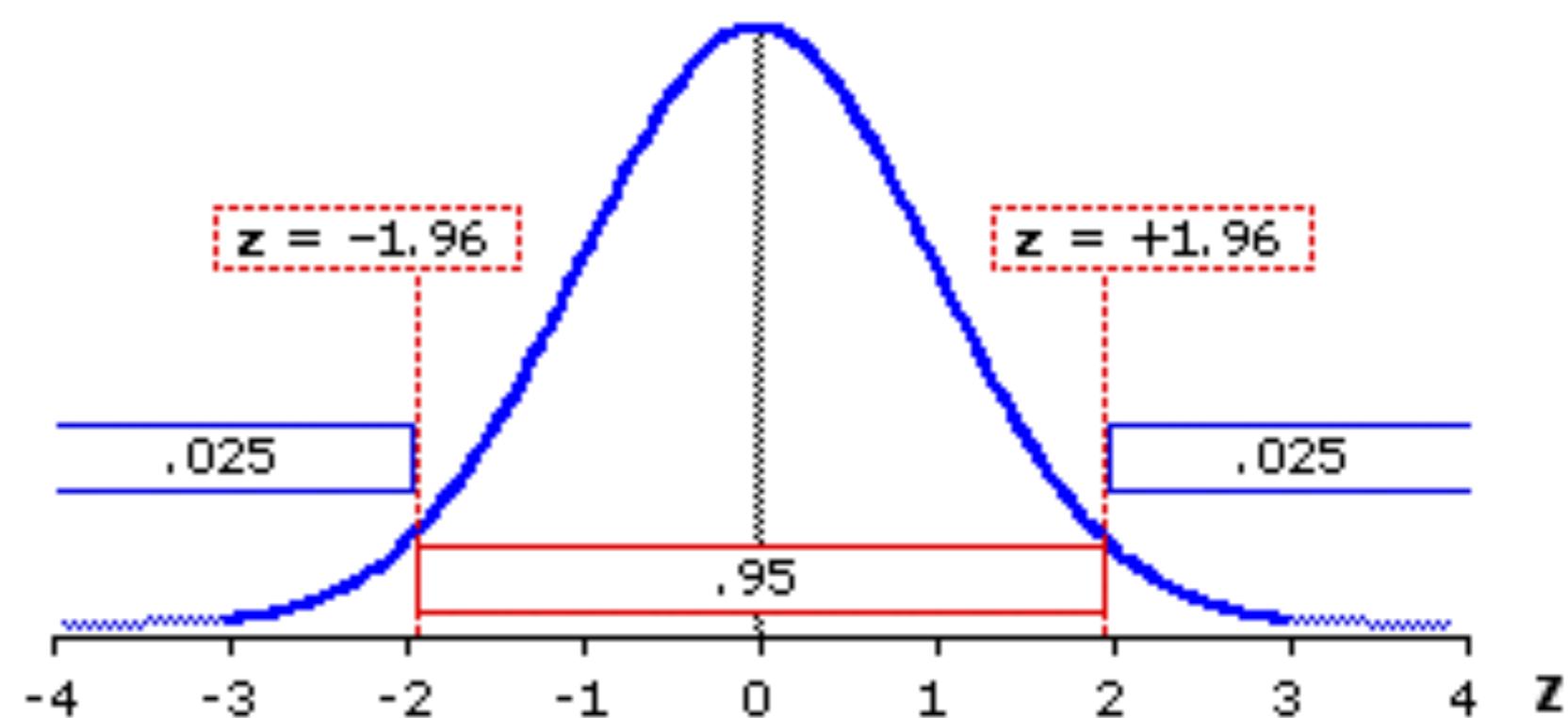
Use Normal if  $N\hat{p} > 5$

$$N(1 - \hat{p}) > 5$$

# Calculating Confidence Interval

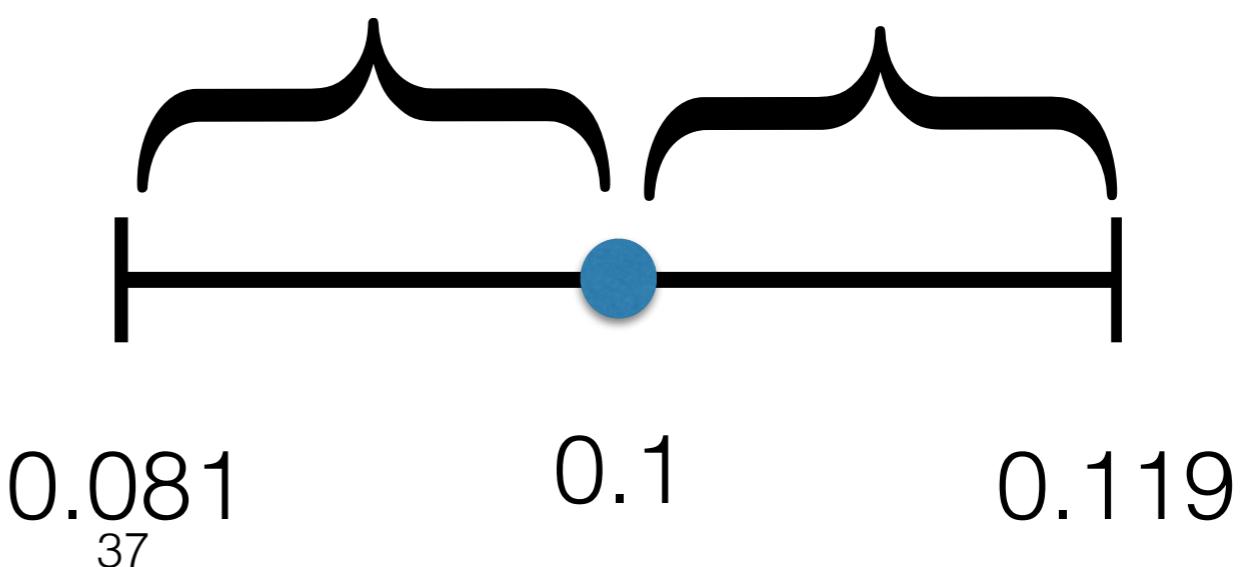
$$m = z * SE$$

$$m = z * \sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}$$



$m$  = margin of error

- $m = 0.019$



# Hypothesis Testing

$P(\text{results due to chance})$

$p_{exp}$  = (probability that someone in the experiment group clicks)

$p_{cont}$  = (probability that someone in the control group clicks)

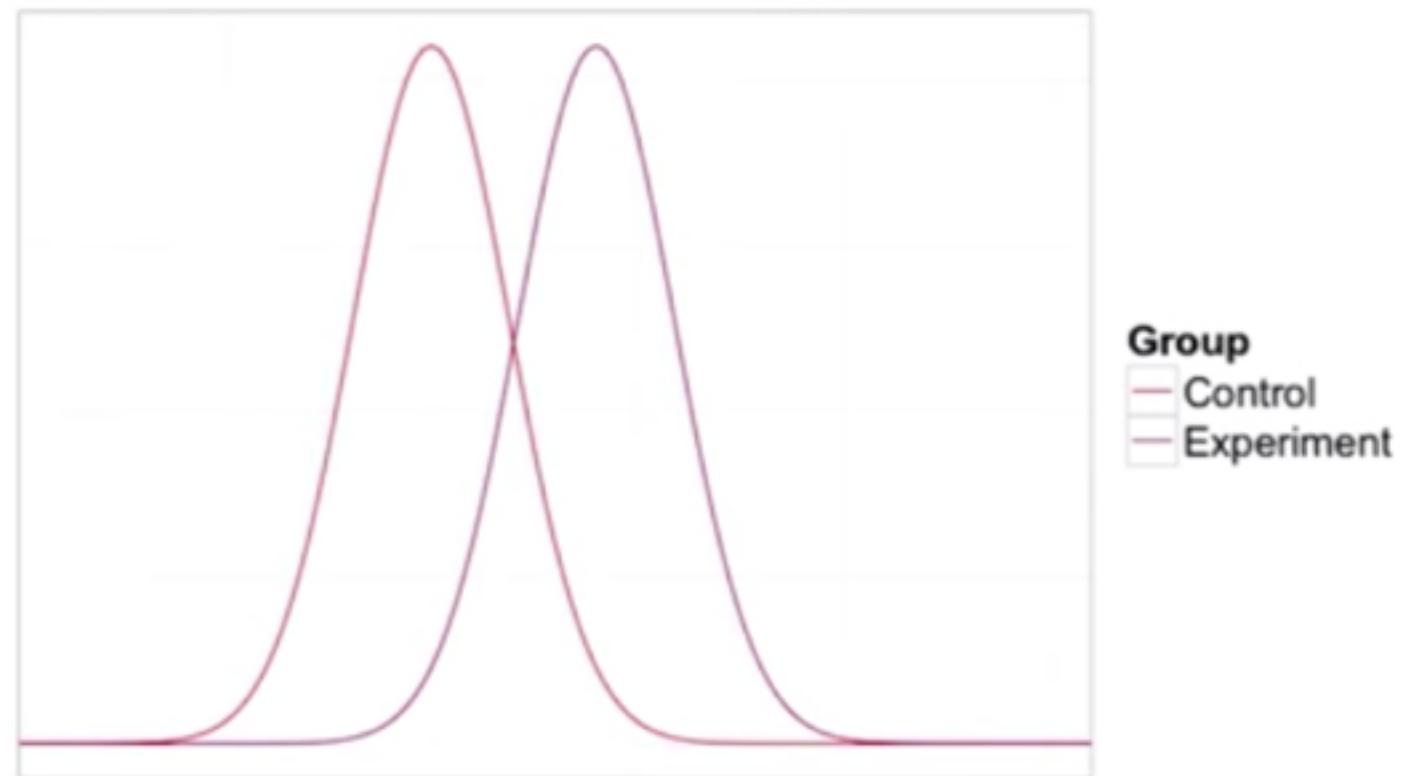
$$H_0 \left\{ \begin{array}{l} p_{cont} = p_{exp} \\ p_{exp} - p_{cont} = 0 \end{array} \right.$$

$$H_A \left\{ p_{exp} - p_{cont} \neq 0 \right.$$

Estimate  $\hat{p}_{cont}$   $\hat{p}_{exp}$

Calculate  $P(\hat{p}_{exp} - \hat{p}_{cont} | H_0)$

Reject null if **P<0.05**



# Comparing Two Samples

- Now we have set up a hypothesis test. How do we decide whether to reject the null?
  - Need to look at confidence intervals
  - Tricky! In this case we have two samples: The control side and the experiment side. These two groups may have a different number of users. We need to compare the two samples!

# Pooled Standard Error

- Measure the number of users who click in each group  $x_{cont}$   $x_{exp}$
- Measure the total number of users in each group  $N_{exp}$   $N_{cont}$

$$\hat{p}_{pool} = \frac{x_{cont} + x_{exp}}{N_{cont} + N_{exp}}$$

$$SE_{pool} = \sqrt{\hat{p}_{pool} * (1 - \hat{p}_{pool}) * \left( \frac{1}{N_{cont}} + \frac{1}{N_{exp}} \right)}$$

$$\hat{d} = \hat{p}_{exp} - \hat{p}_{cont} \quad H_0 : d = 0 \quad \hat{d} \sim \mathcal{N}(0, SE_{pool})$$

If  $\begin{cases} \hat{d} > 1.96 * SE_{pool} \\ \hat{d} < -1.96 * SE_{pool} \end{cases}$  **Reject null**

# What Significance?

- We have seen how to assess whether a difference we observe in our experiment is actually significant.

What's next?

We have to decide from a business perspective, what size change matters to us?

- Practical significance
- Substantive significance
- Iconic's practical significance is 2%

# Size vs. Power

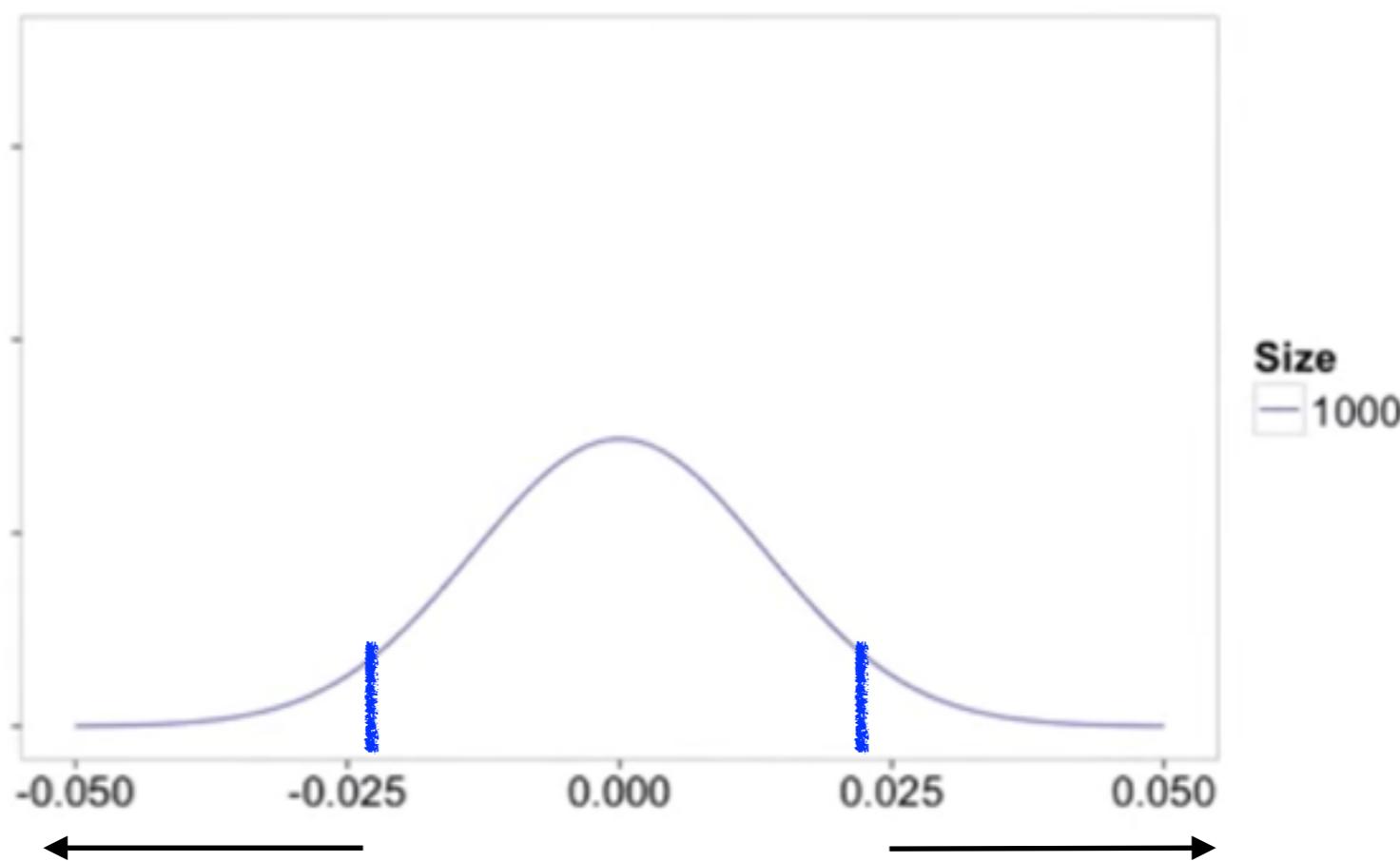
- Trade-off!
  - The smaller the change we want to detect the larger the experiment

# Hypothesis Testing Grid

		Do not reject $H_0$	Reject $H_0$
$H_0$ is true	<b>Correct Decision</b> 😊	Type 1 error $\alpha$	
$H_0$ is false	Type 2 error $\beta$	<b>Correct decision</b> $1 - \beta$	

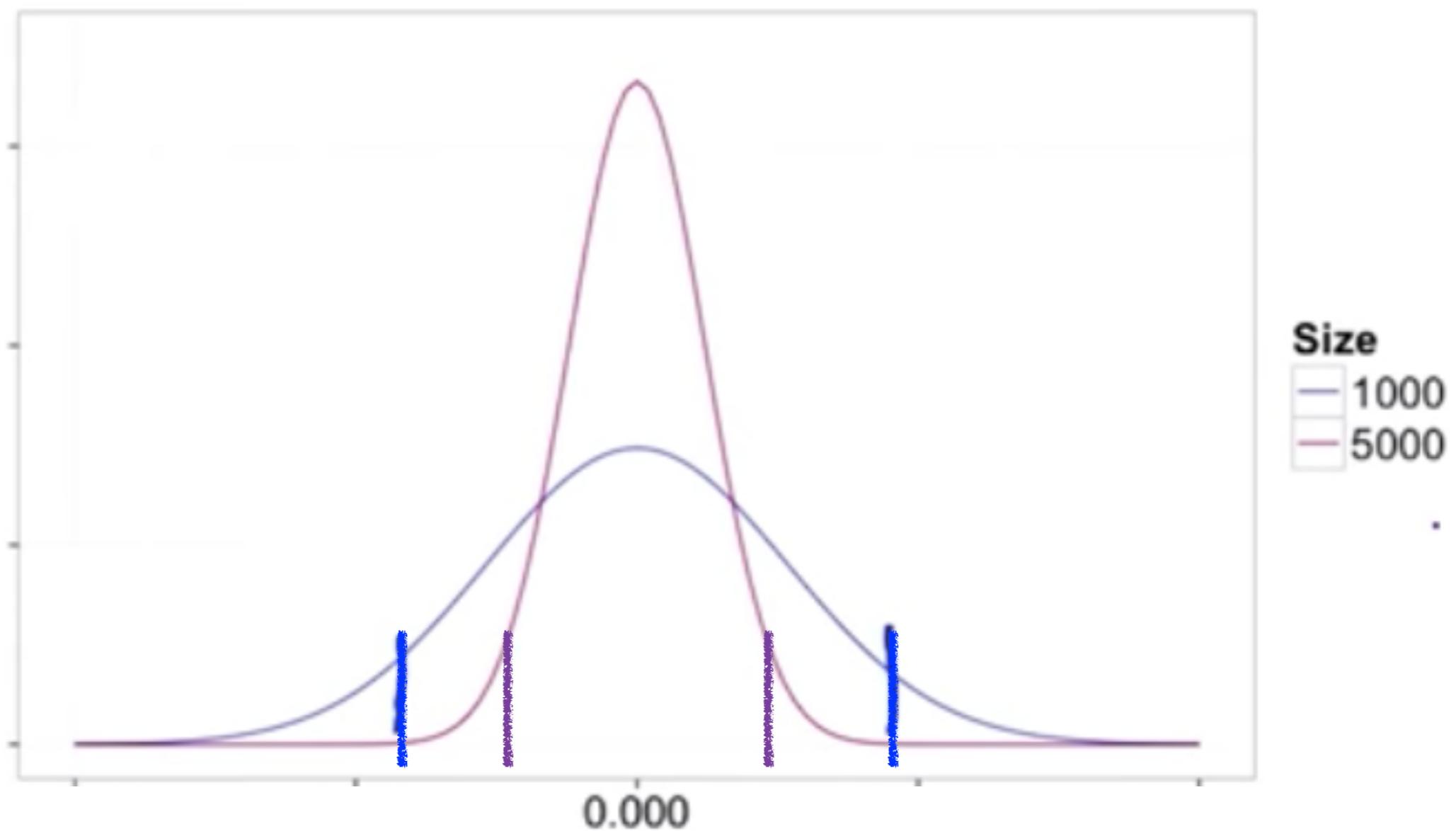
the statistical power of the test

# Sample size vs. Power



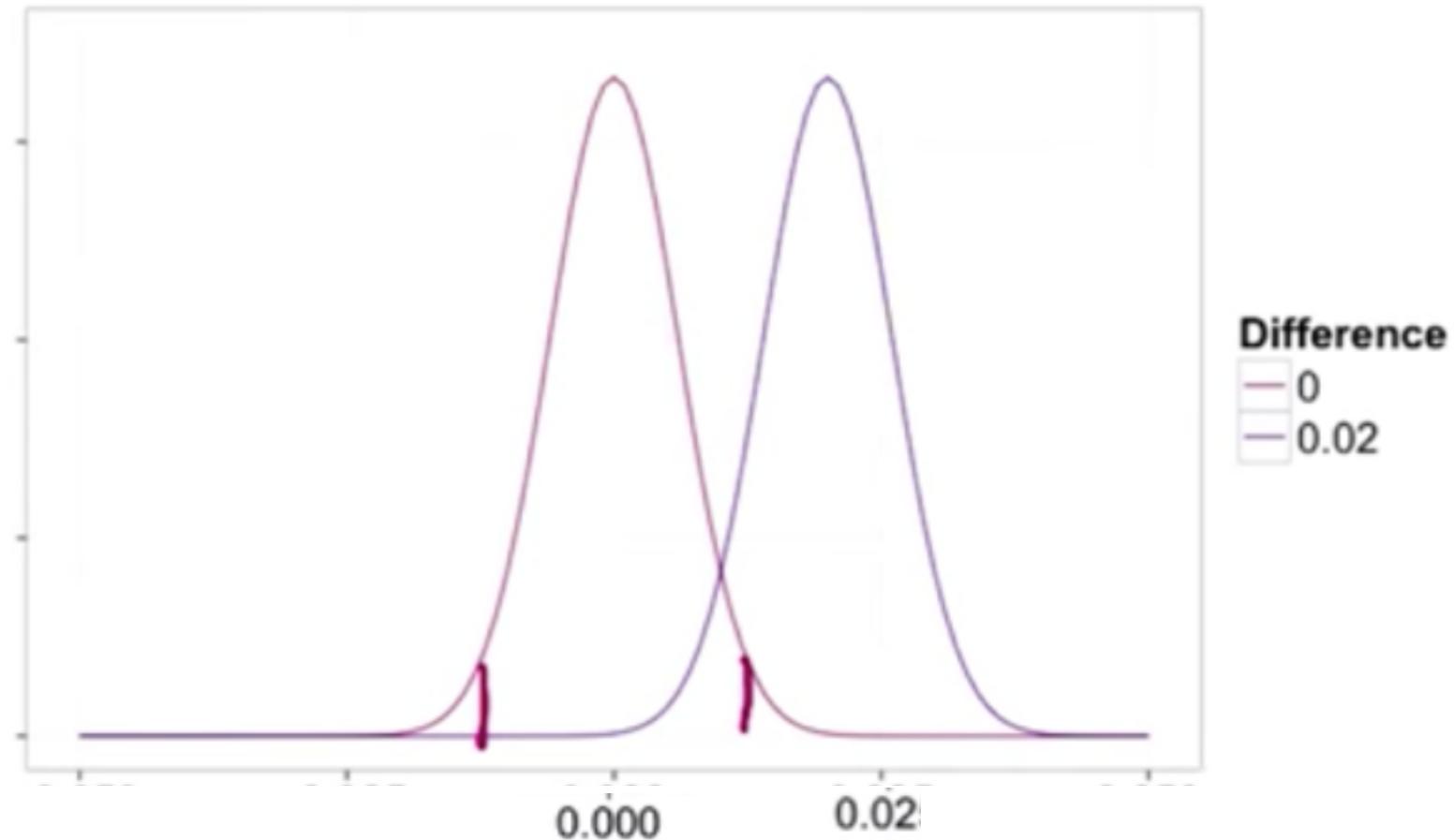
$$\alpha = P(\text{Reject null} | \text{null true})$$

Probability of falsely concluding there was a difference  
44



$$\alpha = P(\text{Reject null} | \text{null true})$$

$$\beta = P(\text{Fail reject} | \text{null false})$$



$$1 - \beta = \text{Statistical power} \sim 80\%$$

**Small Sample**  $\begin{cases} \alpha & \text{low} \\ \beta & \text{high} \end{cases}$

**Larger Sample**  $\begin{cases} \alpha & \text{same} \\ \beta & \text{lower} \end{cases}$

# Determining Sample Size

- Built-in libraries
- Look up answer in a table
- Use online calculator

# Determining Sample Size

- Built-in libraries
- Look up answer in a table
- Use online calculator

**How many page views will we need in each group?**

3,840

$$N=1000 \quad x=100 \quad d_{\min}=0.02$$

$$\alpha = 0.05 \quad \beta = 0.2$$

## Statistical Power

The statistical power of a test is defined as the probability that the test correctly rejects the null hypothesis ( $H_0$ ) when the alternative hypothesis ( $H_a$ ) is true. i.e.,

$$power = P(\text{reject } H_0 | H_a \text{ is true}) = 1 - \beta$$

**Q:** What is the power of the test if the true difference in click-through probability is 0.02? The sample size for each group is 1000, and the click-through probability for the control group is 10%.

```
In [35]: power.prop.test(n = 1000, p1 = 0.1, p2 = 0.1 + 0.02)
```

```
Two-sample comparison of proportions power calculation

n = 1000
p1 = 0.1
p2 = 0.12
sig.level = 0.05
power = 0.2977321
alternative = two.sided
```

NOTE: n is number in \*each\* group

**Example:** Now say if we know that our baseline click-through probability is 10% (the click-through probability before the new feature is introduced), and to be practically significant, we need an absolute difference of 2% in the click-through probability between the control and the experiment groups. In order to have a statistical power of 80%, what is the required sample size for each group?

```
In [36]: power.prop.test(p1 = 0.1, p2 = 0.1 + 0.02, power = 0.8)
```

```
Two-sample comparison of proportions power calculation

n = 3840.847
p1 = 0.1
p2 = 0.12
sig.level = 0.05
power = 0.8
alternative = two.sided
```

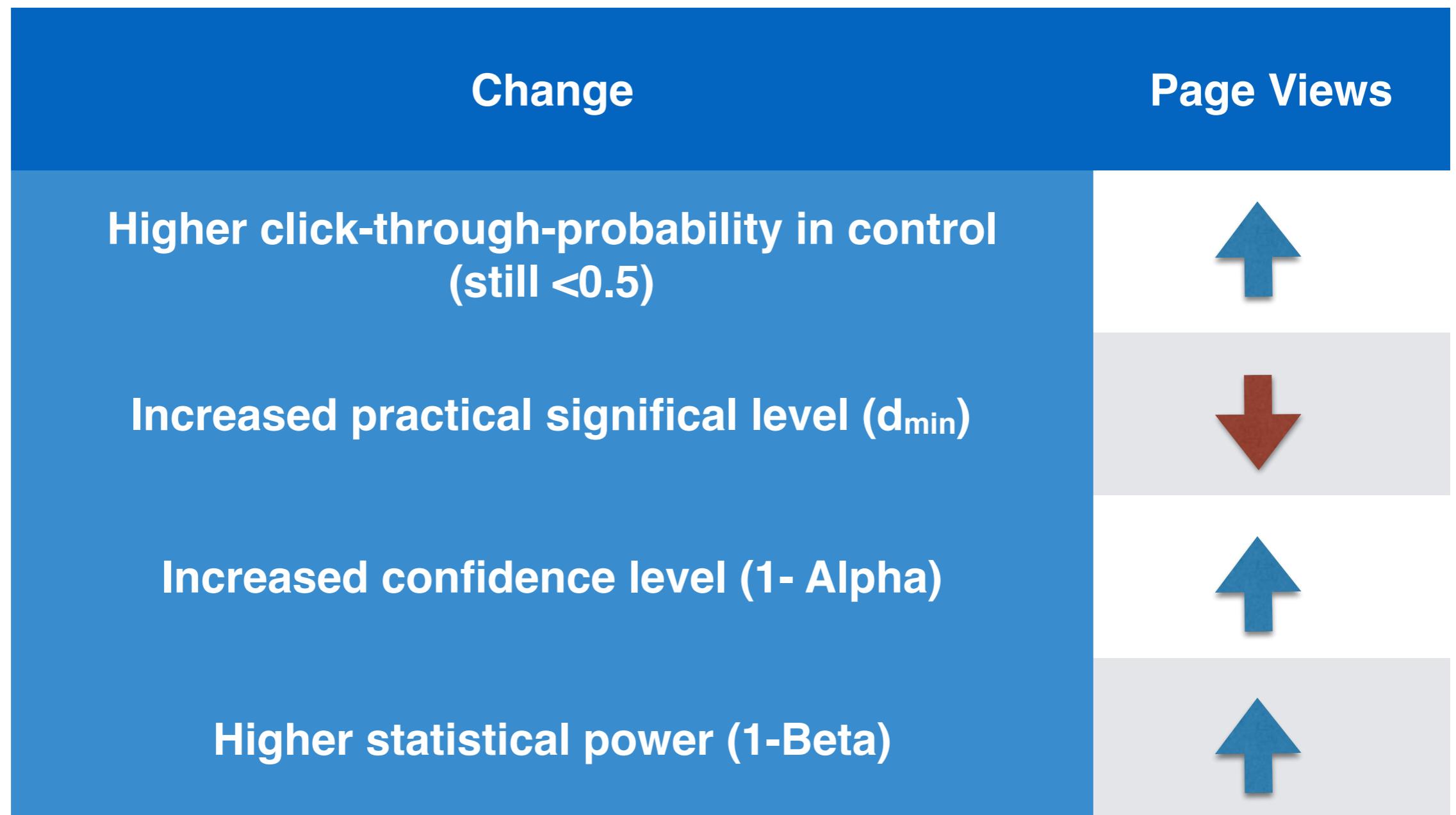
NOTE: n is number in \*each\* group

# How sample size varies

Change	Page Views
Higher click-through-probability in control (still <0.5)	
Increased practical significal level ( $d_{min}$ )	
Increased confidence level (1- Alpha)	
Higher statistical power (1-Beta)	

# How sample size varies

keeping everything else the same



# Analyze Results

$$N_{cont} = 10072$$

$$N_{exp} = 9886$$

$$d_{min} = 0.02$$

$$x_{cont} = 974$$

$$x_{exp} = 1242$$

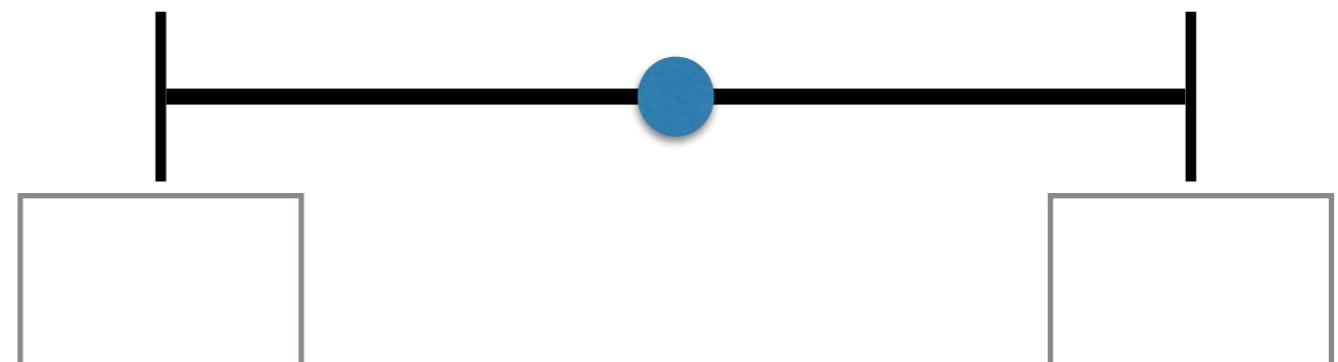
$$1 - \alpha = 95\%$$

$$\hat{p}_{pool} = \frac{974 + 1242}{10072 + 9886} = 0.111$$

$$SE_{pool} = \sqrt{0.111 * (1 - 0.111) * \left(\frac{1}{10072} + \frac{1}{9886}\right)} = 0.00445$$

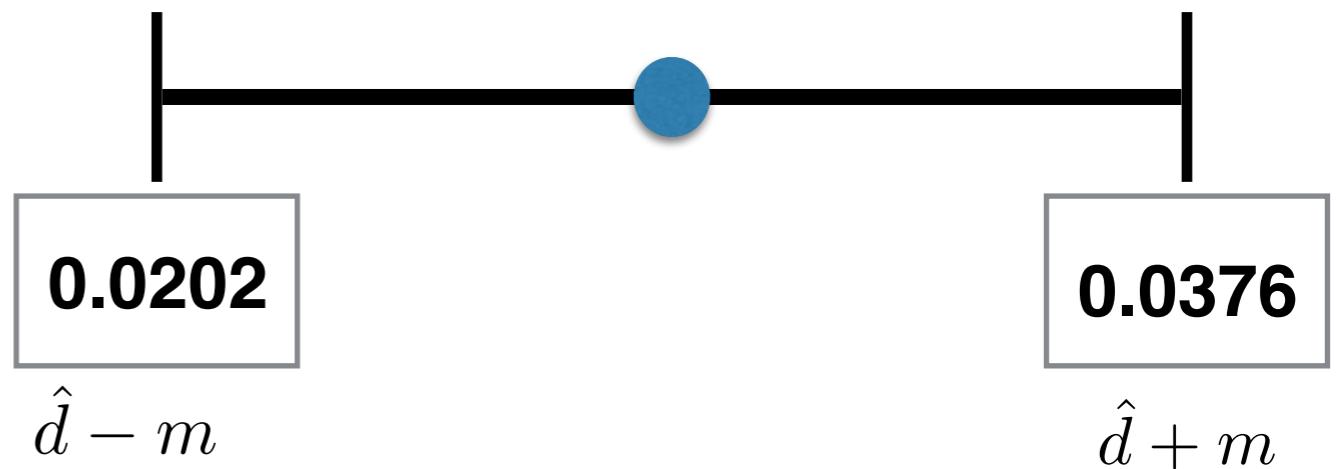
$$\hat{d} = \boxed{\phantom{000}}$$

$$m = \boxed{\phantom{000}}$$



# Calculating the results

$$\hat{d} = \boxed{0.0289} \quad m = \boxed{0.0087}$$

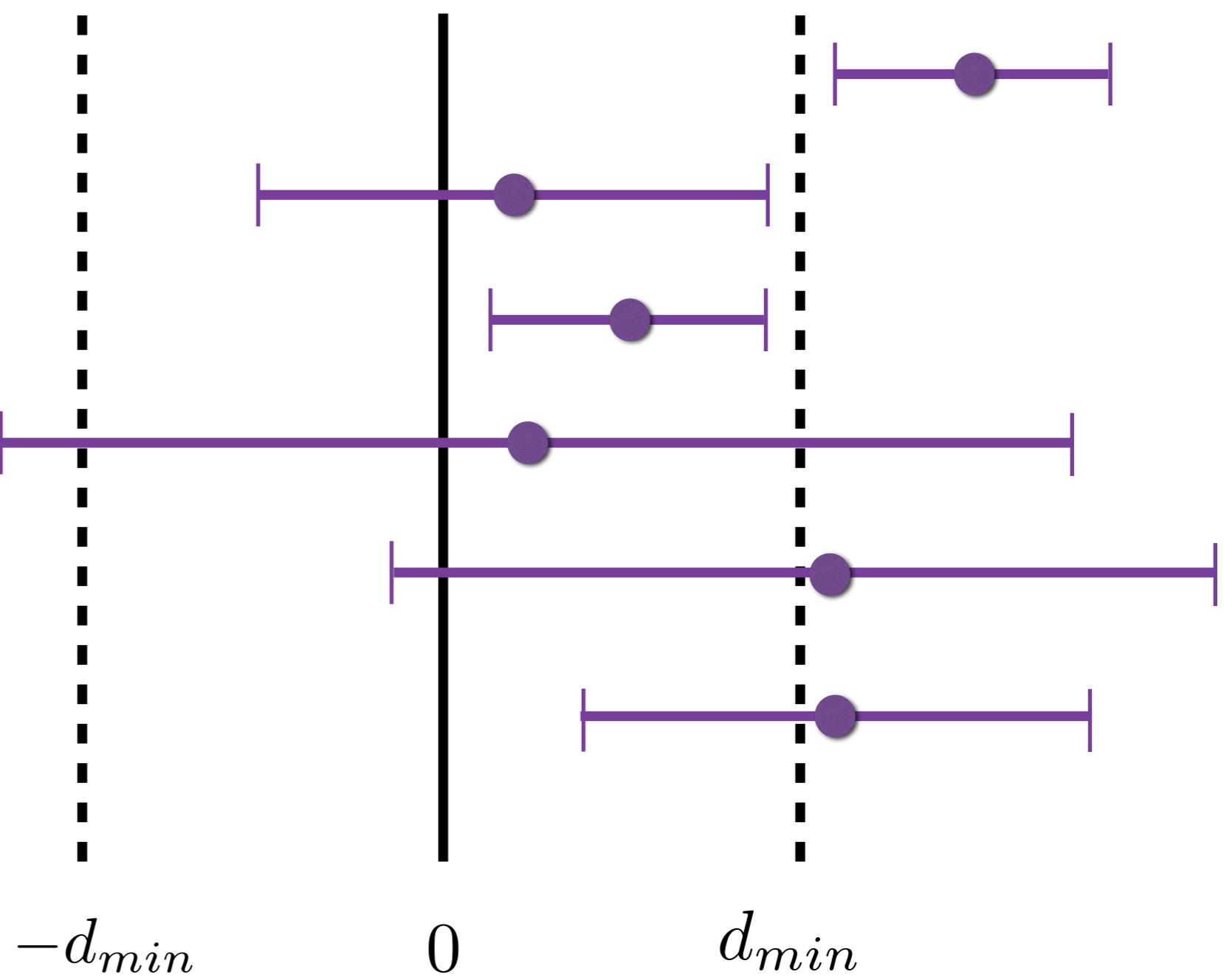
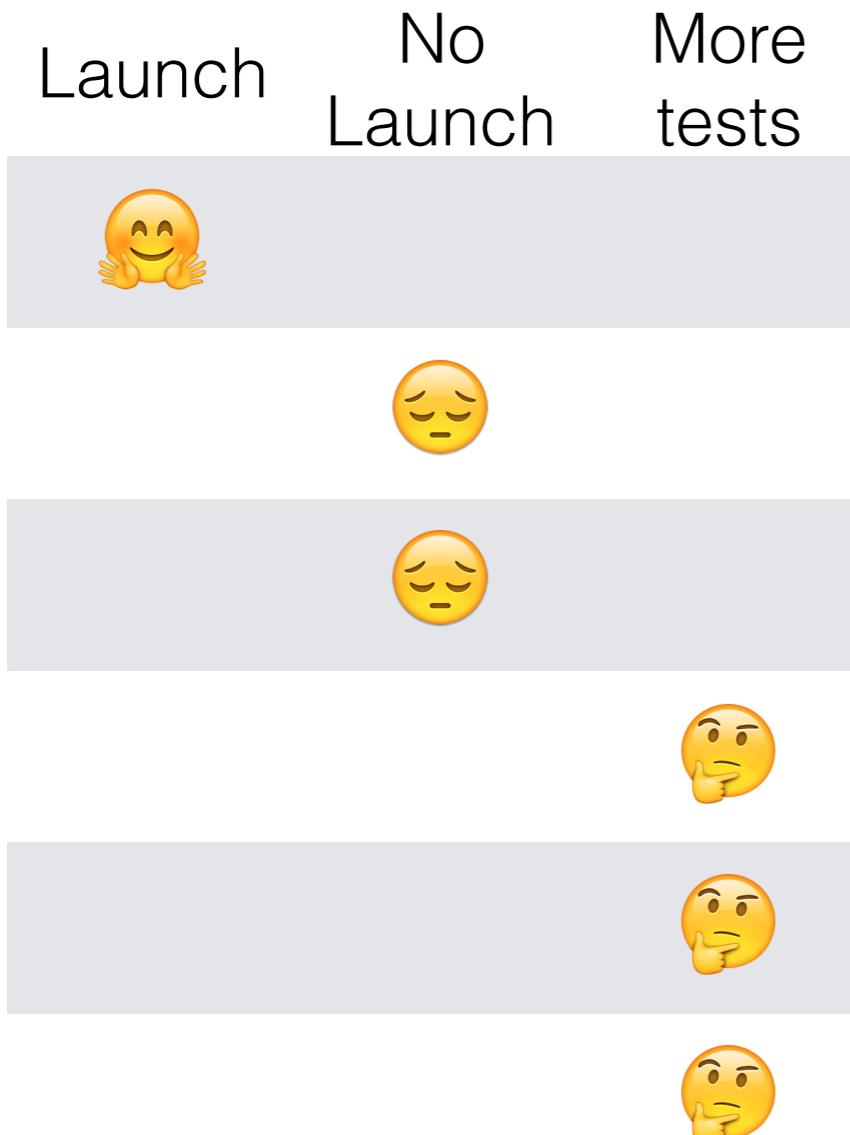


$$\hat{d} = \hat{p}_{exp} - \hat{p}_{cont} = \frac{x_{exp}}{N_{exp}} - \frac{x_{cont}}{N_{cont}} = 0.0289$$

$$m = SE_{pool} * 1.96 = 0.0087$$

Would you  
launch ?

# Confidence Interval Cases



# Uncertain Data

- Communicate clearly
- Know exactly when you need to make a judgement and take a risk.
- Use other factors, like strategic business issues or factors besides the data.

# The importance of metrics

- Analysts and managers work the majority of the time actually coming up with, validating, and choosing metrics to actually use in evaluation
  - Define
  - Build Intuition
  - Characterize

# Define metrics

**Think about how you are going to use the metric:**

- **Invariant checking (sanity check):**  
*Should not change across groups*
  - Are the populations the same?
  - Is the distribution the same?
- **Evaluation**
  - High level business metrics
    - Revenues, market share, # users
  - Detailed metrics
    - User experience with the product

## Invariant Checking

- Before we dive into comparing the click-through probability, we need to do some sanity checks to make sure the experiment is actually run properly
  - Something might have gone wrong in the experiment diversion, are your control and experiment groups still comparable?
  - Did the data capture the events you were looking for?
- We need to check if the experiment population and the control populations are actually comparable
- The invariants shouldn't change when you run your experiment
- **Q:** Say the metric we want to evaluate is the click-through rate, what metrics can we use for invariant checking?

**Q:** If we observed a total of 8294 pageviews in the control group and 8095 pageviews in the experiment group, how do we use it for invariant checking?

```
In [4]: # We can perform a Binomial/proportion test on the invariant  
prop.test(8294, 8294 + 8095)
```

```
1-sample proportions test with continuity correction  
  
data: 8294 out of 8294 + 8095, null probability 0.5  
X-squared = 2.3921, df = 1, p-value = 0.122  
alternative hypothesis: true p is not equal to 0.5  
95 percent confidence interval:  
 0.4983857 0.5137537  
sample estimates:  
      p  
0.5060711
```

# Definition: How?

Several steps

1. High level concept for a metric
2. Details about the high level concept
3. Summarize your data measurements into a metric
  - sum, count, percentile, rate, probability, ratio

# Evaluation Metric(s)?

- Cultural
  - One metric
    - Easy for PR and external reporting
    - Teams work towards the same, clear objective
  - Composite Metric
    - Objective function
    - Overall Evaluation Criterion (OEC)
- The metric should be generally applicable

# Other Techniques

## Brainstorming and Evaluating Metrics

- External data
- Gathering your own in-depth data

# External Data

- Comscore and Hitwise (visitors to a site)
- Nielsen, Forrester, and Pew Research (studies)
- eMarketer
- Conferences
  - CHI
  - WWW
  - KDD
  - WSDM

# What for?

- Validate simple business metrics
  - Look at both time series of internally computed metric and externally available one, and see if the trend and seasonal variability line up.
- Provide supporting evidence for your business metrics
  - Paper

# Your own in-depth data

- User Experience Research (UER)
- Focus Group
- Surveys

# User Experience Research (UER)

- In-depth, intensive, and qualitative
- Useful for generating hypotheses
- Few users (tens of users at most)

- **Diary Study**

- Self-documented behavior
- Self-reporting bias





# Focus Group

- Show screenshots, walk the through demos, ask questions (hypothetical)
- Group-think (convergence on a fewer opinions)
- Hundreds of users.



*"This is interesting, 70% of the respondents to our survey said they don't respond to surveys."*

# Surveys

- Multi-choice answers, open-ended questions
- Google's Consumer Surveys
- Thousands of users
- Self-reported data related disadvantages
  - Ensure the population is representative
  - Wording may prime the participants to give specific answers
  - <https://www.nngroup.com/articles/which-ux-research-methods/>
  - <https://www.usability.gov/what-and-why/user-research.html>

# Retrospective/Observational Analysis

- Running analyses on an existing set of observational data without an experiment structure
- Generating ideas for A/B tests
- Validating metrics

# Long-term prospective experiments

- Come up with or validate metrics to use in the long term
  - Run A/B tests and measure a change in the metric in the long-term
  - Build models to determine which short-term metrics best predict the long-term effects

# Human Evaluation

## “Crowd-sourcing”

- Ranking-oriented systems
- You pay people “raters” to complete a specific task.
  - good for getting a lot of labeled data (simple)
- Mechanical Turk, MicroWorkers
- Search Evaluation at Google

# Sensitivity & Robustness

Picks up the changes you care about but does not move against changes that you don't care about.

How do we measure them?

- Experiments
- A/A tests
- Retrospective analysis of your logs

# Analyzing Results

## 1. Sanity checks

- Experiment and Control groups comparable
- Checking invariants

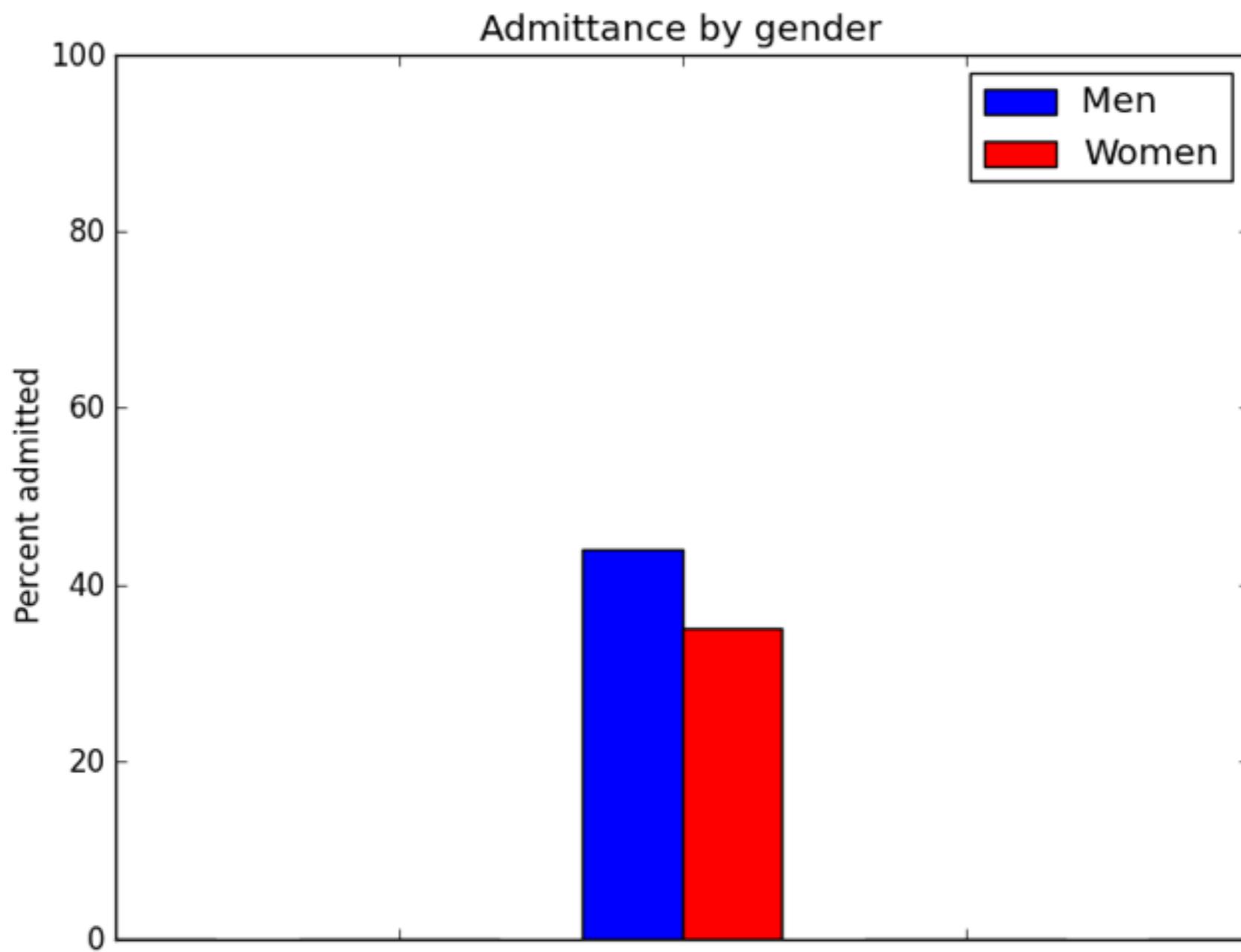
If any of your sanity checks fails? DO NOT PROCEED

- Debug the set up of the experiment working with your engineers
- Retrospective analysis
- Pre and post-period tests

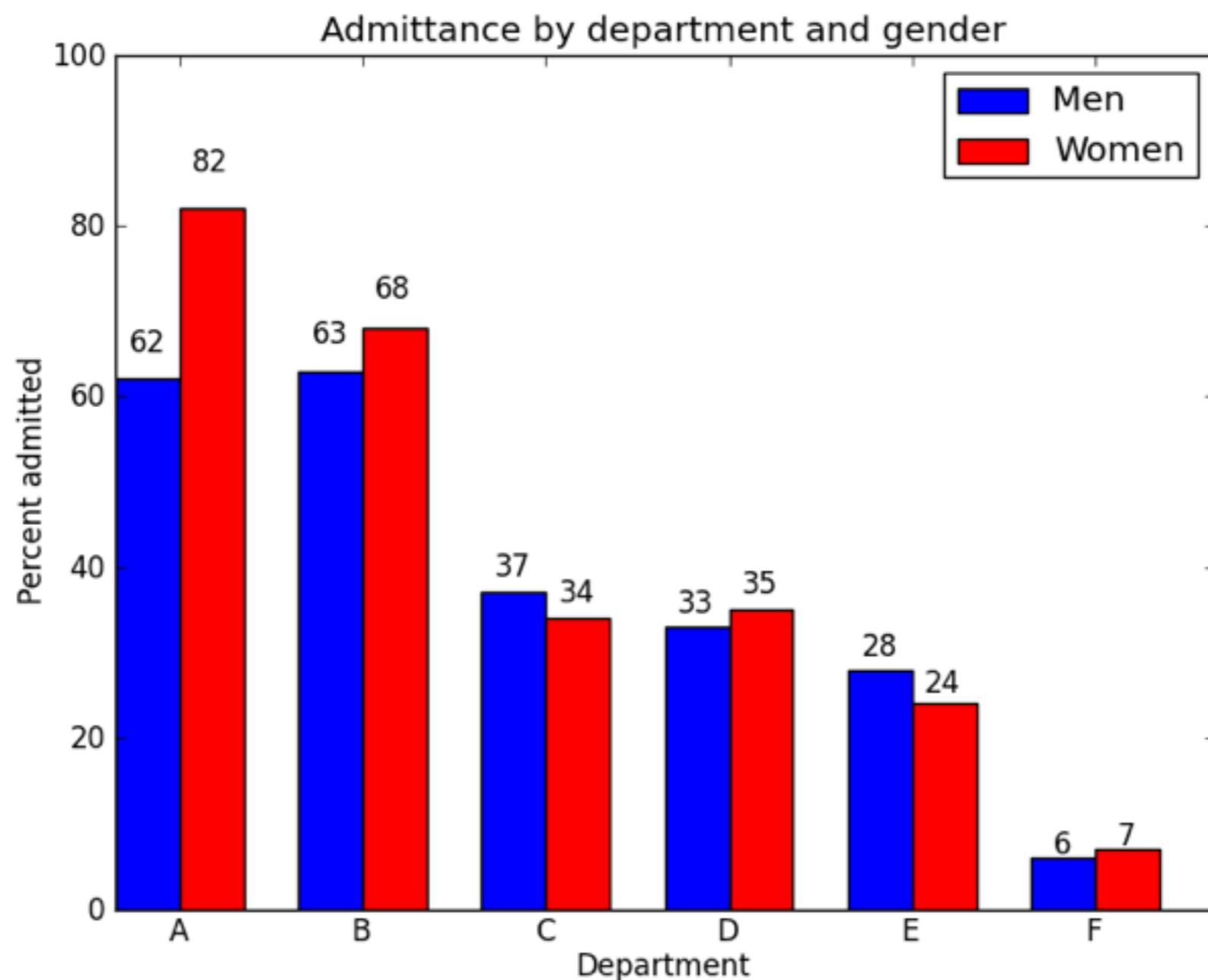
# Simpson's Paradox

Simpson's paradox, or the Yule–Simpson effect, is a paradox in probability and statistics, in which **a trend appears in different groups of data but disappears or reverses when these groups are combined**. It is sometimes given the impersonal title reversal paradox or amalgamation paradox.

# Berkeley Admissions



# Berkeley Admissions



# Multiple Metrics

- As you test more metrics, it becomes more likely that one of them will show a statistically significant result by chance.
- Multiple comparisons technique

# Tracking Multiple Metrics

**Experiment:** Prompt customers to complete their purchase when there is something added to their shopping cart

## Metrics:

1. Probability that customer completes their purchase at any point in the next month
2. How early customers complete their purchase
3. Average price paid by customer

If Iconic tracks all three metrics and does three separate significant tests ( $\alpha=0.05$ ), then what is the probability that at least one metric will show a significant difference if there is no true difference?

# Tracking Multiple Metrics

**Experiment:** Prompt customers to complete their purchase when there is something added to their cart

**For 3 metrics, what is the chance of at least 1 false positive?**

$$P(FP = 0) = 0.95 * 0.95 * 0.95 = 0.857$$

$$P(FP > 0) = 1 - 0.857 = 0.143$$

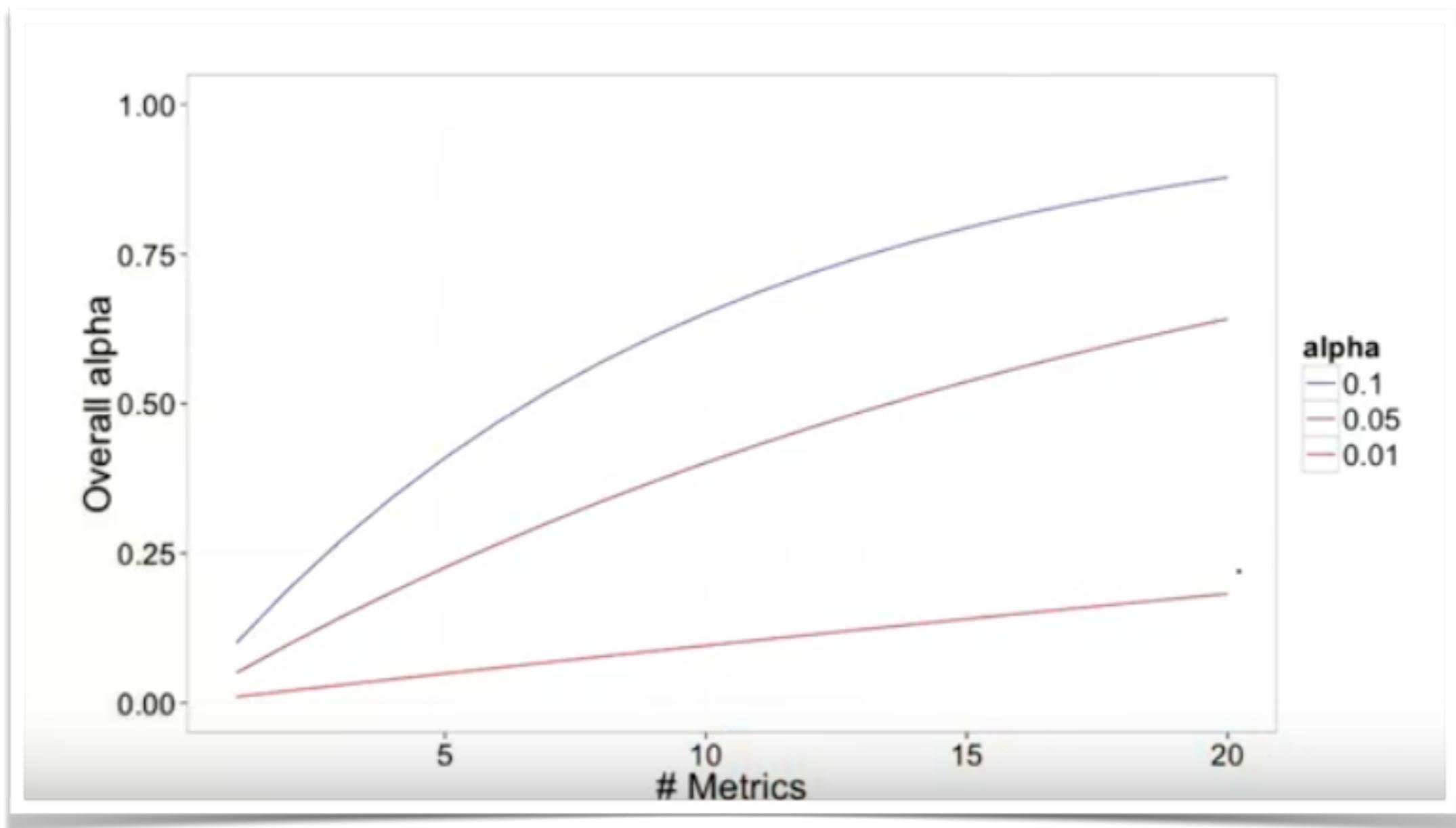
**The overall probability of at least one false positive is:**

$$\alpha_{overall} = 1 - (1 - \alpha_{individual})^n$$

10 metrics and 95% confidence interval => 0.401

10 metrics and 99% confidence interval => 0.096

# Overall Alpha



# Tracking Multiple Metrics

**Problem:** The probability of any false positive increases as you increase number of metrics.

**Solution:** Use higher confidence level for each metric

## Method 1: Assume independence

$$\alpha_{overall} = 1 - (1 - \alpha_{individual})^n$$

## Method 1: Bonferroni Correction Simple/No Assumptions/Conservative

$$\alpha_{individual} = \frac{\alpha_{overall}}{n}$$

# Do I launch or not?

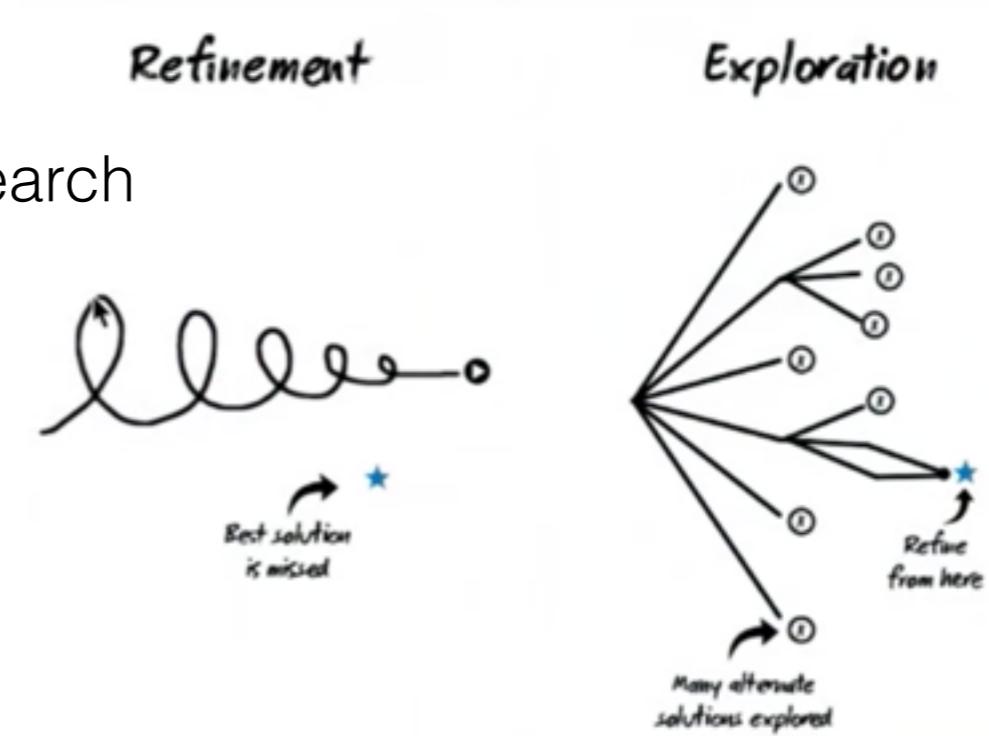
1. Do I have statistically and practically significant results in order to justify the change?
2. Do I understand what that change has actually done with regard to the user experience?
3. Is it worth it?

# Takeaways

- Define quantifiable success metrics

- Explore before you refine

- Breath-first search before a depth-first search



# Case Study

- Data and instructions are available [here](#).
  - Odd # teams (e.g., Team 1, Team 3, etc.,) will present to a non-technical audience
  - Even # teams (e.g., Team 2, Team 4, etc.,) will present to a technical audience.

You should start brainstorming with your teammates today and start working on the case tomorrow after class.

Presentations are due on Thursday, February 9th by 2pm. No late submissions will be accepted.

- Upload presentation slides and ipython notebook to your team's folder available [here](#).