A/B Testing 3

Mini 4 / Spring 2024

THE INTELLIGENT FUTURE

Carnegie Mellon University
Tepper School of Business

Guest Speakers

- Mon, April 8
  - Guest speaker: **Sachal Lakhavani (AI Product Manager @ Meta; Tepper MBA alum)**
  - 30 min Q&A about A/B testing & Product management (Section A: 5-5:30pm, Section E: 6:30-7pm)
  - Everyone is expected to prepare for at least one question (I might cold call!)
  - 15 min 1:1 Q&A (Section A: 5:30-5:45pm, Section E: 7-7:15pm)
- Mon, April 15
  - Guest speaker: **Michael Degnan (Senior VP, Head of Enterprise Innovation @ PNC)**
  - 1 hour lecture on New Product Development & Innovation
  - 30 min Q&A about Innovation & Product management

Final Project Presentations

- Schedule
  - Section A: Wed, April 17 (Teams 1-3) & Mon, April 22 (Teams 4, 5, 7)
  - Section E: Mon, April 15 (all teams)
- Deliverables
  - 20 min presentation & slides (no min/max requirement)
  - Submit the slides to Canvas by the beginning of Wed, April 17 (Section A) or Mon, April 15 (Section E)
- Logistics
  - Each team should ask at least one question for other teams' presentations
  - **Attendance is expected for all teams' presentations**

Final Project Presentations

- Product selection
  - Approved on Canvas; Let me know if you are not happy with the current selection
- Required
  - What is the Problem a product manager is facing (WFH or GenAI)
  - Environment analysis (e.g., 3C)
  - Customer needs discovery (e.g., interview, UGC, survey; please define your target segment)
  - **(At least one) Feature proposal** (show us what features you considered; connection with customer needs is essential; please consider the AARRR framework)
  - Design of A/B testing to assess the proposed feature
- Optional
  - Frameworks from the pre-class videos (e.g., RICE)
  - Qualitative/quantitative insights (e.g., from market reports, data, your own experiences)

Final Project Presentations

- Grading
  - Contents:

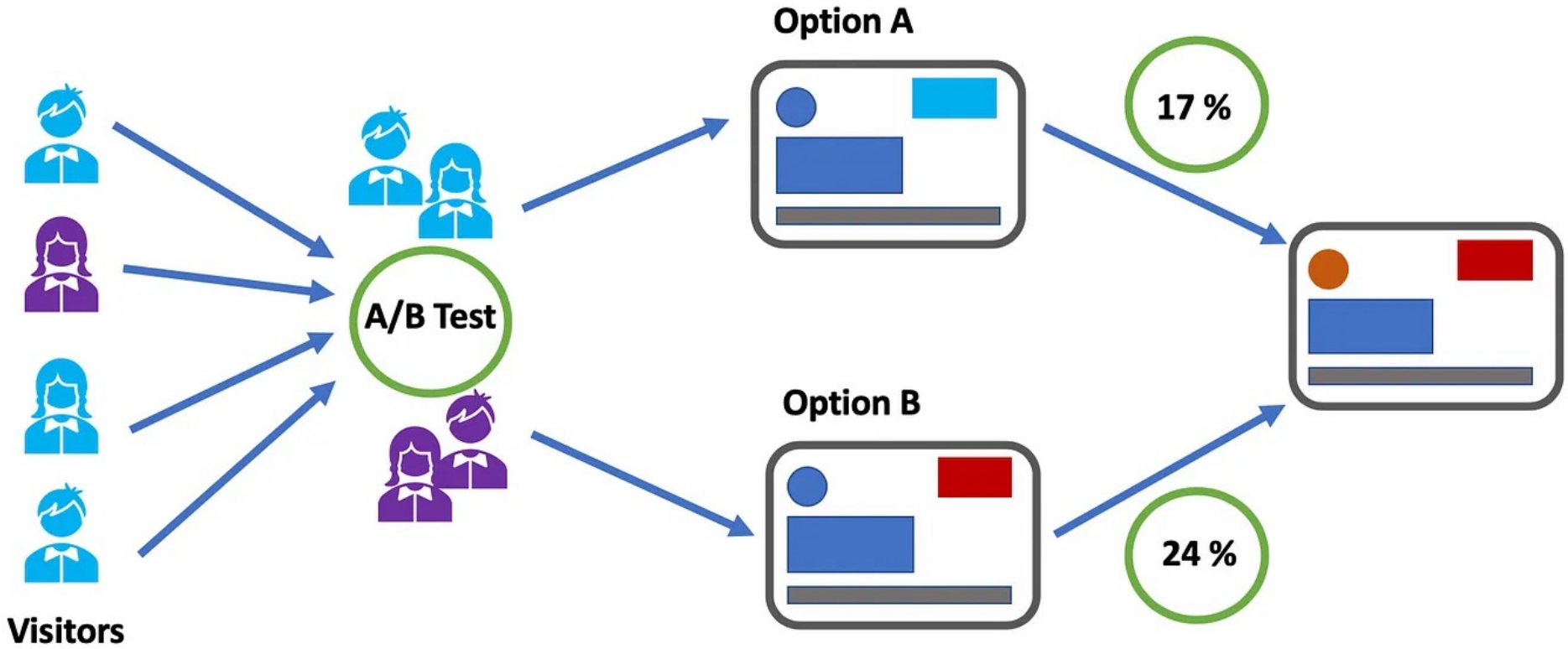
  1) Customer needs discovery
  2) Product feature
  3) A/B testing design (hypothesis, randomization, etc.)

  - Presentation: Clarity, Storytelling, Time limit (20 min)
  - Discussion: Q&A with the class
  - Teamwork: Peer evaluation of team members after the last day of class

- All of you will give feedback to other teams and pick the best team(s) for bonus points (logistics to be elaborated on presentation days)

RECALL: LAST WEEK

- A/B Testing
  - Examples in tech firms: Amazon, Ebay, Netflix, etc.
  - Compare.com case: customer churn
  - How to define hypotheses for A/B testing
  - Uber pool case: how to randomize users into control and treatment conditions
  - Booking.com case: the role of culture and management, challenges in experimentation at scale

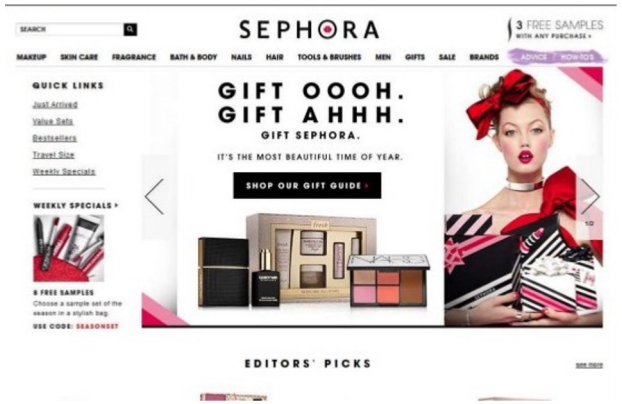

Option A
17 %
A/B Test
Option B
24 %
Visitors

Intuition:

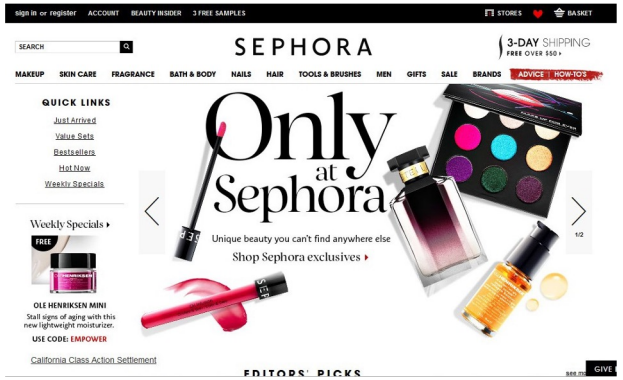
Why A/B Test?

EXAMPLE: IMPROVING SEPHORA'S WEBPAGE (NO A/B TEST)

- Context: Sephora.com launches new landing site by replacing the old landing site for all users and scrapping the old one.
- Metric: "Click-through-rate" of web page.
- Results
  - Old Webpage 10% click-through-rate (before new site)
  - New Webpage 12% click-through-rate (after new site)
- Question: Should we stay with the new webpage? Why or why not?

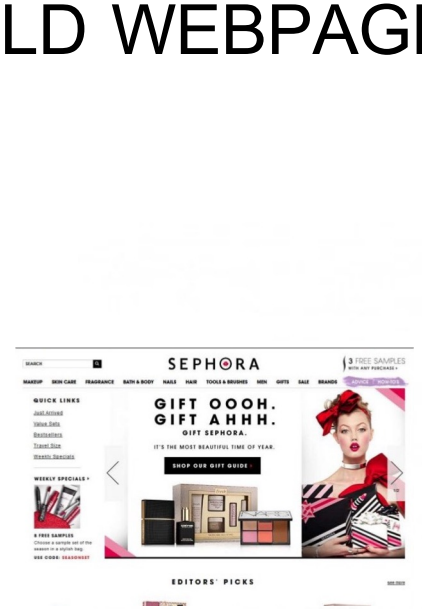


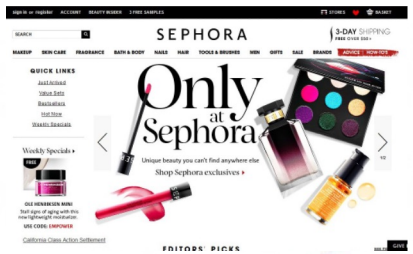
Old Webpage

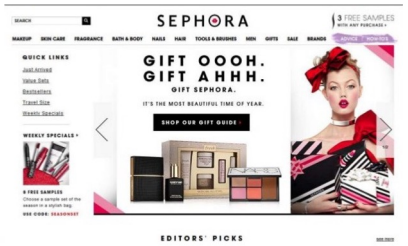


New Webpage

LET'S SAY WE TRIED THE NEW WEBPAGE,THEN BACK TO OLD WEBPAGE, THEN THIS HAPPENED…



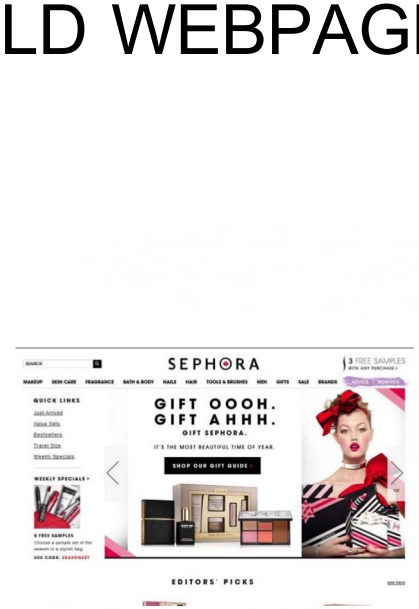


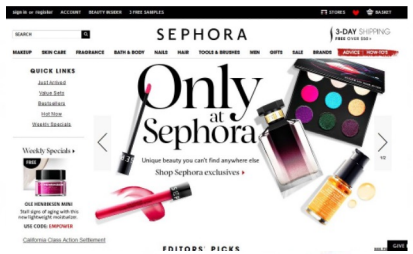


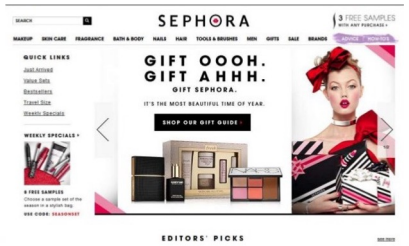
Still going up! Why?

Source: https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7

LET'S SAY WE TRIED THE NEW WEBPAGE,THEN BACK TO OLD WEBPAGE, THEN THIS HAPPENED…







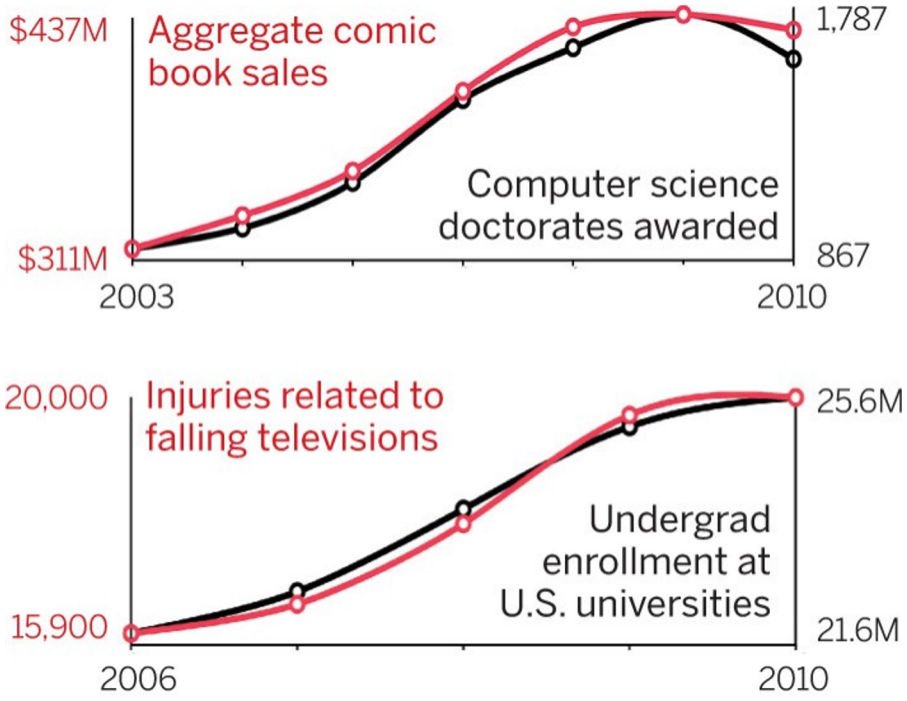
Still going up! Why?

Maybe people are just buying more makeup across all websites…

Source: https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7

CORRELATION IS NOT CAUSATION

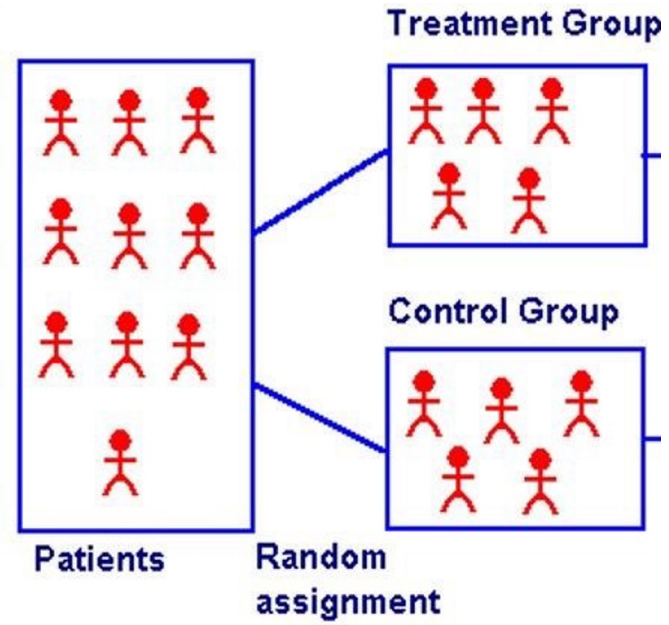
- We can not say the new webpage is better than the old webpage
- The click through rate could simply be higher for any webpage during the time we tested
- We need to control for “unrelated correlation”
- How? Using a ”randomized controlled trial,” also known as an “A/B Test”



KEY IDEA: TREATMENT AND CONTROL GROUPS

- A/B Test is a "**Randomized controlled trial**"
  - Control Group (Group A) - This group sees no change from the current setup.
  - Treatment Group (Group B) - This group is exposed to the new web page

- Goal of A/B Test
  - Compare the click-through-rates of the two groups using statistical inference.
  - Test our Product Feature "Hypothesis"
  - "[Specific repeatable action] will create [expected, measurable result]"

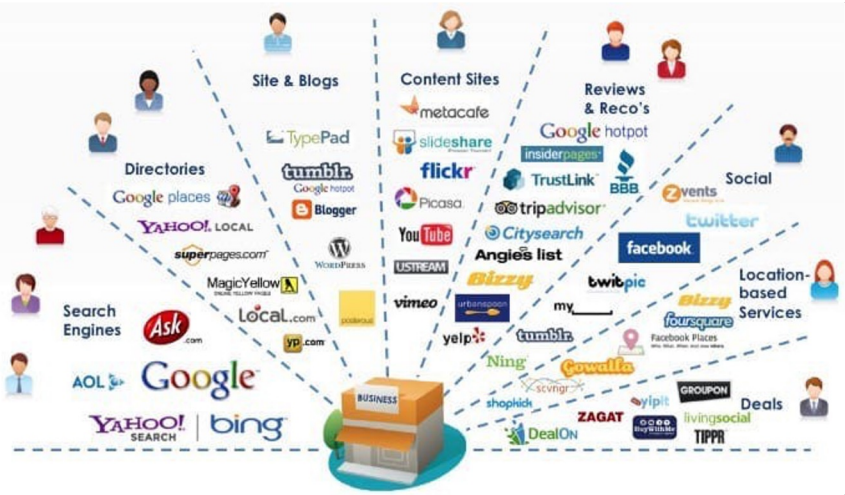


KEY IDEA: WE "RANDOMIZE" USERS INTO CONTROL (A) AND TREATMENT (B) TO MEASURE CAUSATION (OF OUR PRODUCT FEATURE)

- Why? The world is not a vacuum…
  - More is happening than just the experiment (treatment vs control group) and effect.

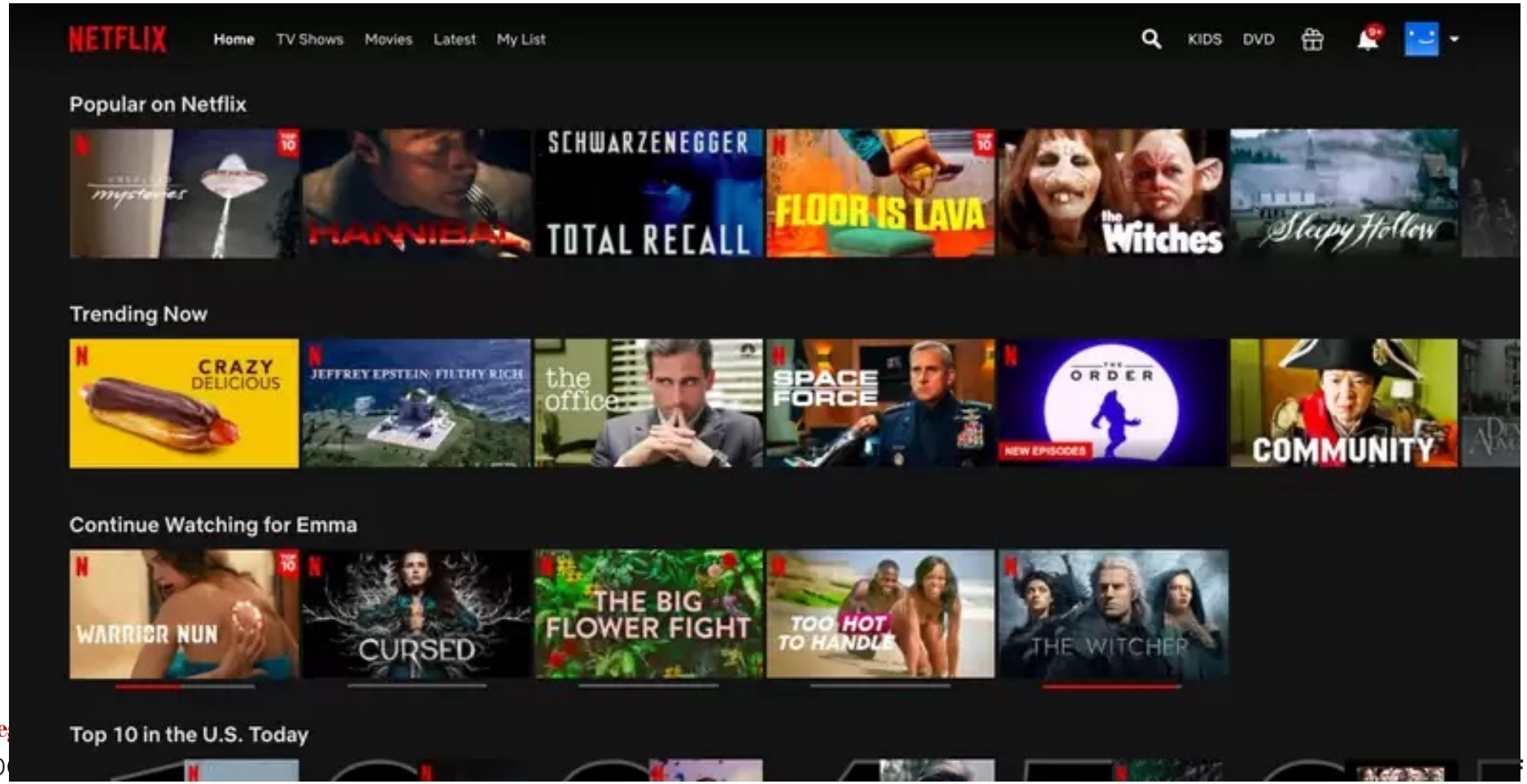
- Many "random" situations lead to "biases"
  - Users overall are changing (e.g., buying more makeup).
  - Users characteristics are changing: Different demographics, new vs returning
  - Users have changing goals and intentions: Browsing around, buying immediately, bored and scrolling.
  - Users are finding our webpage differently than before: email, newsletters, web searches, social media

- <u>Key Point</u>: Randomization of users into treatment (A) and control (B) helps balance these out.



What to consider in A/B Testing?

Let’s suppose you are running a A/B test on Netflix main page layouts.



Let's suppose you are running a A/B test on Netflix main page layouts.

WHY IS IT BETTER THAN HISTORICAL DATA ANALYSIS?
> WE JUST DISCUSSED IT.

QUESTION 1: WHAT IS THE CHALLENGE IN TESTING AT SCALE (TESTING MANY DIMENSIONS AT THE SAME TIME)?

QUESTION 2: WHAT ARE THE PROS/CONS OF RUNNING TESTS FOR ALL CUSTOMERS VS. A SUBSET OF CUSTOMERS (WHO WOULD YOU TEST)?

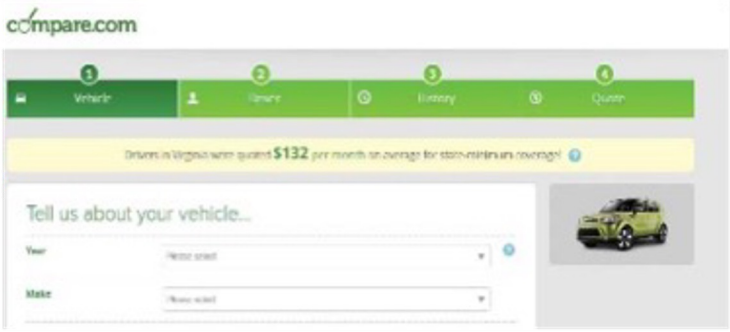
QUESTION 3: WHAT METRICS YOU WOULD MEASURE AS AN OUTCOME?

CAITLIN SMALLWOOD (VP OF SCIENCE @ NETFLIX) IN 2021

From 10:46

3 STEPS FOR A/B TESTING

- Step 1) Hypothesis Definition
  - How do we define success or failure (of our product feature)?
  - What metrics do we need to track and measure to test our hypothesis?



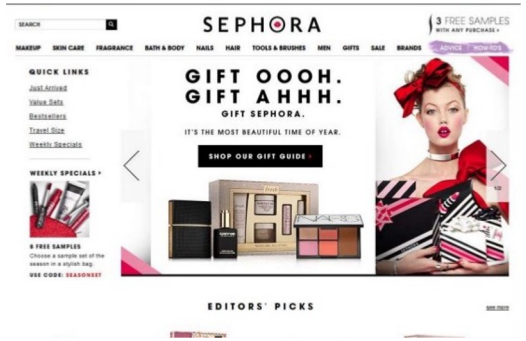
- **Step 2) Experimental Design and Launch A/B Test**
  - Who do we ask / who should we get data from?
  - How do we ensure we are not getting "biased" results?
  - How many samples do I need? How long to run A/B test?

"Drivers in Connecticut were quoted $_____"

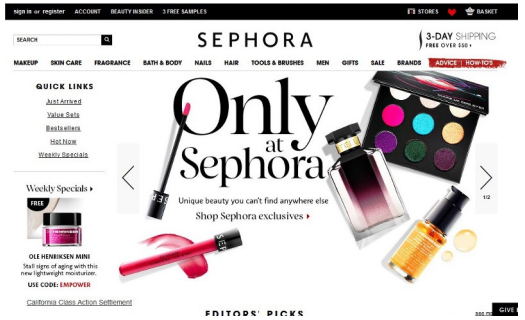
- Step 3) Hypothesis Testing and Interpretation
  - Note: A/B Test = Hypothesis Test
  - What is statistically significant for our test of our hypothesis?
  - Do we go with product feature A or B?

Back to Sephora: Webpage A vs Webpage B (now with A/B test)

- Now Sephora ran Webpages A and B
  - Users randomly assigned A or B
- Outcome
  - Webpage A: 10% click-through-rate
  - Webpage B: 12% click-through-rate



Old Webpage A



New Webpage B

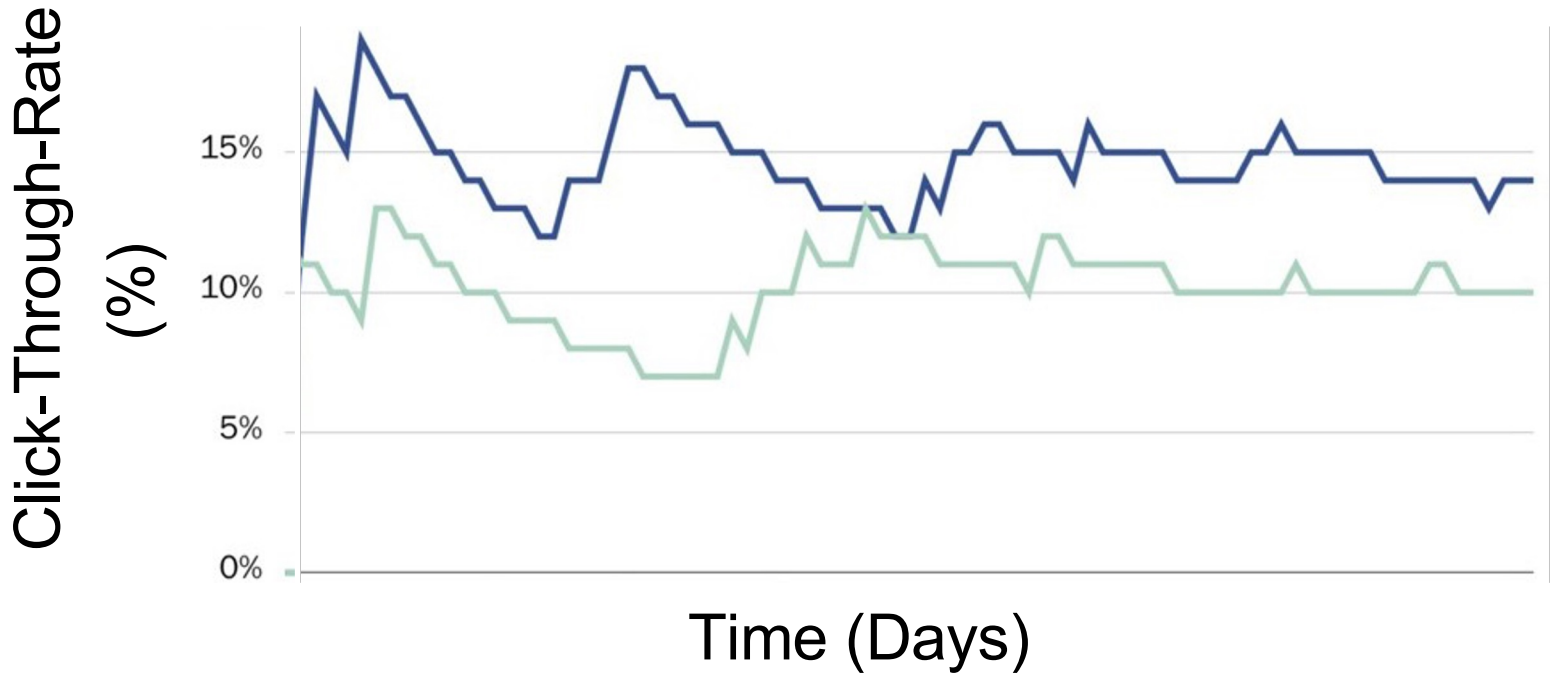
- Question: Do we implement Webpage B? Why or why not?
- Answer: Not yet! We don't know if this is statistically significant.

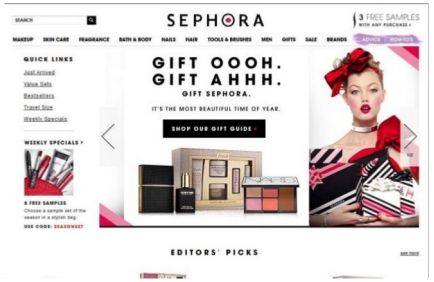
Webpage	Number of Samples	Click-Through-Rate
A	1000	10%
B	1000	12%

WHY HYPOTHESIS TEST?

ANSWER: BECAUSE A AND B HAVE RANDOMNESS.

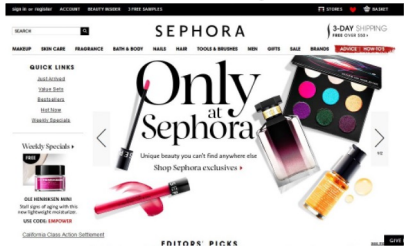
WE NEED TO DETECT SIGNAL OVER NOISE





Webpage A

Webpage B



Notice: a lot of “bouncing around” of metric for A and B.

KEY POINT: METRICS NEARLY ALWAYS HAVE UNEXPLAINABLE RANDOMNESS.

WE NEED (STATISTICAL) HYPOTHESIS TESTING TO CHECK SIGNIFICANCE.

- Why randomness?
  - We do not observe all reasons why people are bouncing at our webpage
  - Example: Economy bouncing back, Valentines Day sales for make up
- Common mistake
  - This is not "measurement error," our measurements are fine
- Statistical Randomness of Sample vs Population
  - The data we collected are only samples of overall population.
  - We don't know if we got samples that randomly had more click-through-rate in general.
  - We need statistics to help account for this.

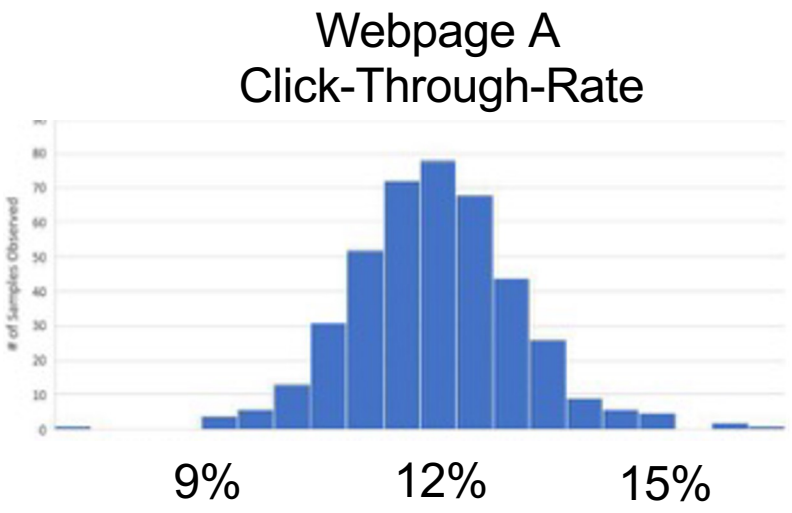
Webpage	Number of Samples	Click-Through-Rate
A	1000	10%
B	1000	12%

Click-Through-Rate (CTR) over Time

12% +/- 3% Click Through Rate

HYPOTHESIS TESTING CONCEPTS

- *Null hypothesis*
  - *The hypothesis, often referred to as H_0, that A and B are not different and observed differences during experiment are due to random fluctuations.*
- *P-value*
- *Confidence level.*
  - *The probability of failing to reject (i.e., retaining) the null hypothesis when it is true.*
  - *Confidence level.* Commonly set to 95%, this implies that 5% of the time we will incorrectly conclude that there is a difference when there is none (Type I error). All else being equal, increasing level reduces our statistical power.
- *Alpha or Significance Level*
  - 1 – Confidence Level
  - There is a 95% chance of the new feature beating the original feature
- *Statistical Power*
  - *The probability of correctly rejecting the null hypothesis, H0, when it is false. Power measures our ability to detect a difference when it indeed exists.*
- *Standard error*
  - SD of the sampling distribution; The smaller the Std-Err, the more powerful the test.



Reality	Do not reject H_0	Reject H_0
H_0 is true	Correct decision ☺	Type I False Positive (α)
H_0 is false	Type II False Negative (β)	Correct decision ☺

NULL HYPOTHESIS: EXAMPLE OF RESTAURANT TECH PRODUCT FEATURE

Hypothesis definition for product feature

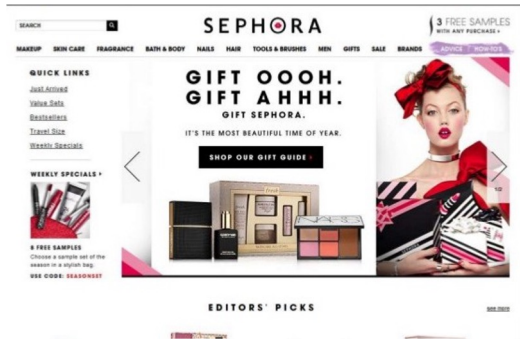
"[Specific repeatable action] will create [expected, measurable result]"

Example: "Restaurants that add the food pickup feature will increase their overall number of orders per day"

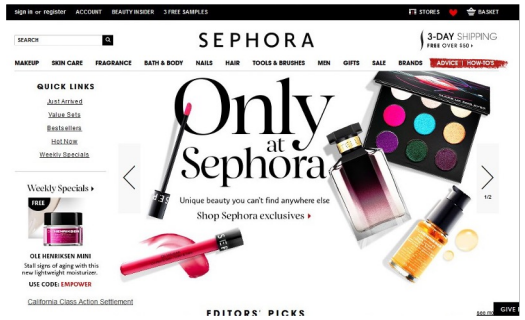
- H_0 *(null)*: Restaurants that add the food pickup feature will neither increase or decrease total number of orders per day
- H_1 *(ours)*: Restaurants that add the food pickup feature will increase their overall number of orders per day

QUESTION: WHAT IS NULL HYPOTHESIS AND HYPOTHESIS TO TEST?

Context: Testing two version of website – A (old) and B (new)
Metric: Click-through-rate (CT) for header banner on website



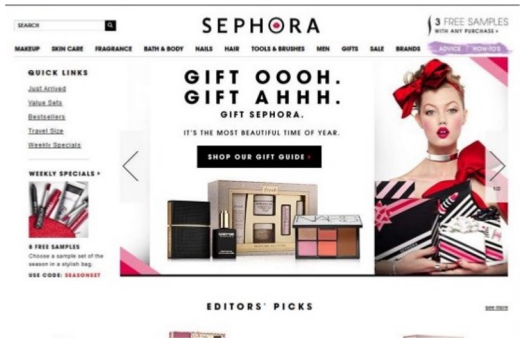
A



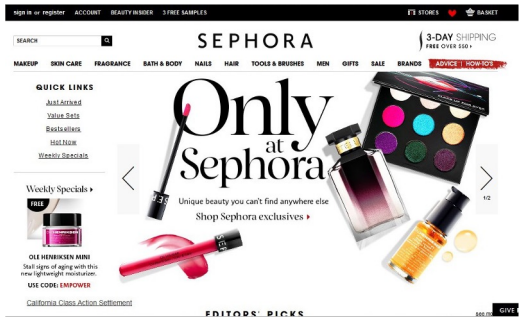
B

- H_0: <null hypothesis>
- H_1: <hypothesis to test>

ANSWER: WHAT IS NULL HYPOTHESIS AND OUR HYPOTHESIS?



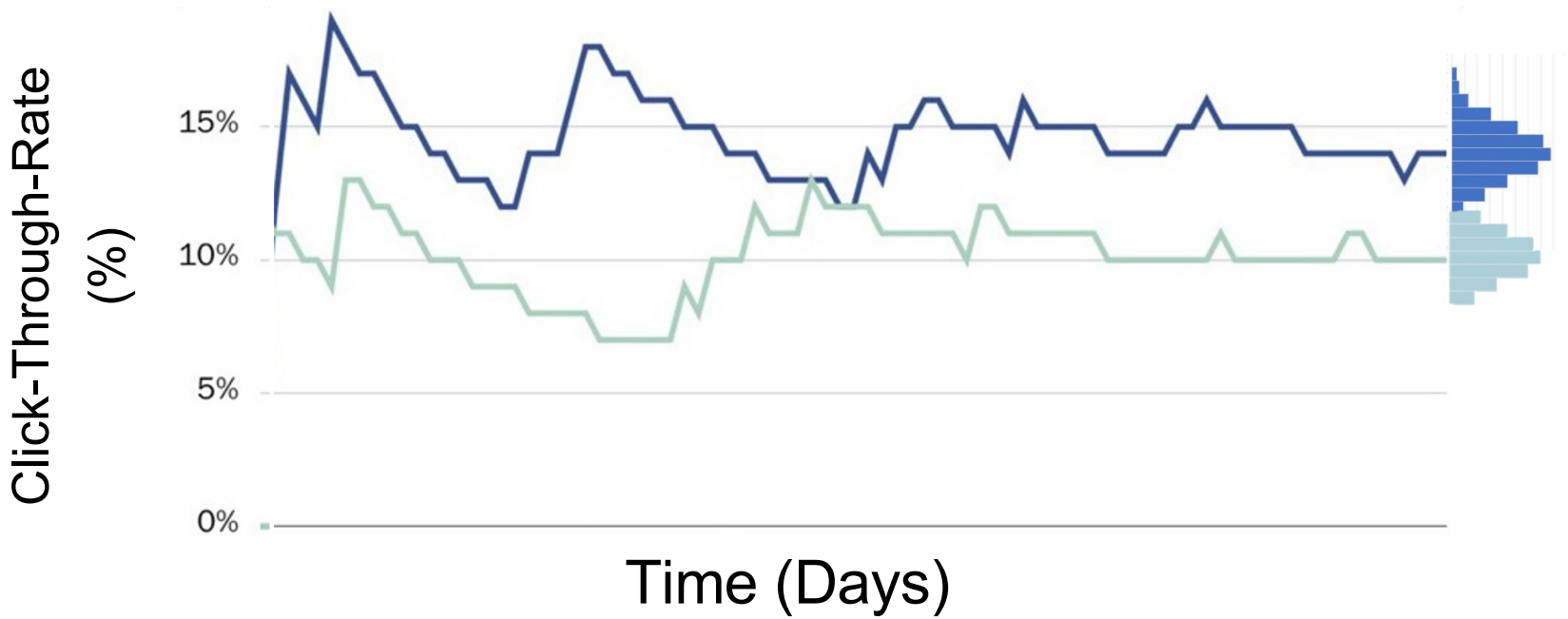
A

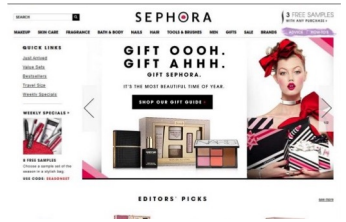


B

- H_0: the click through rate is the same for A and B.
- H_1: the click-through-rate is higher (or lower) for webpage A than webpage B. There is an "effect."

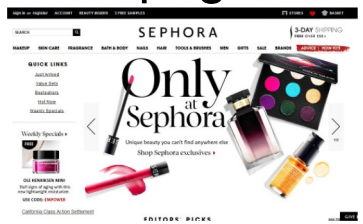
INTUITION: "IS B BETTER THAN A?" AND THEIR DISTRIBUTIONS





Webpage A

Webpage B



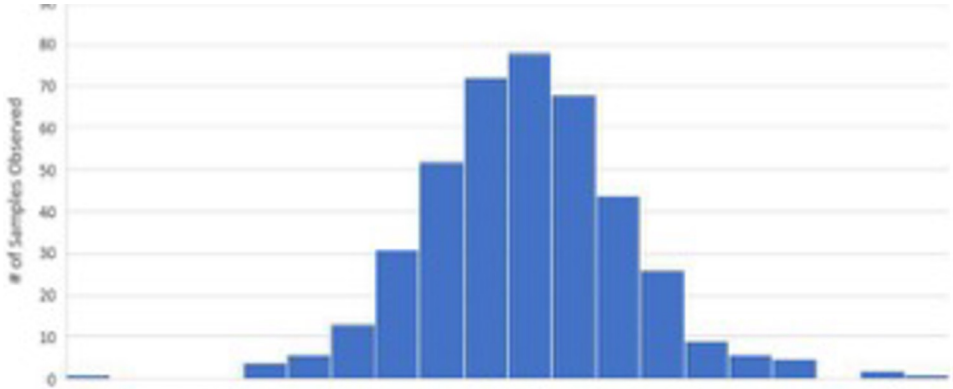
Notice: There is a lot of "bouncing around" of metric for A and B.

Key Point: This randomness leads to a distribution of our tracked metric for A and B (blue and teal distributions).

Intuition: The "normal" distribution of A and B

- "Normal" is the "statistical model" we will be using for click-through-rates (and A/B testing)
  - A.k.a., the "Gaussian" Distribution
- Technical Note:
  - We actually need "Binomial Distribution" - the number of Heads in a sequence of Bernoulli Trials with replacement
  - But this is approximated with a Gaussian

Webpage A
Click-Through-Rate

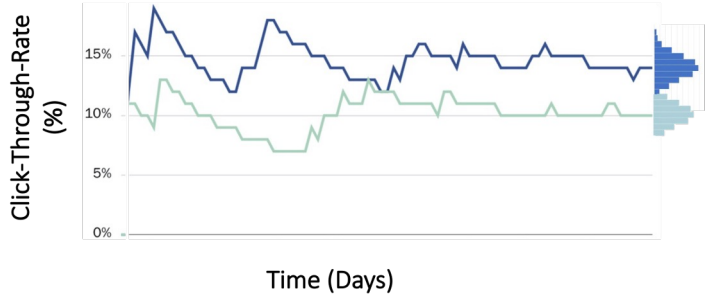


9% 12% 15%

Intuition for hypothesis test: overlapping distributions A and B

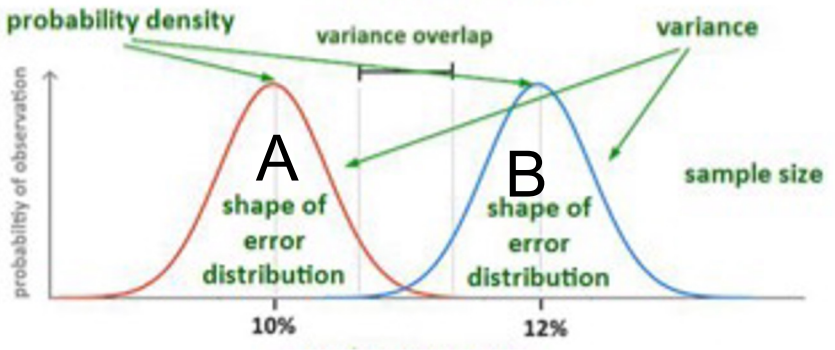
Overlap of Two Distributions

- Webpage A and Webpage B



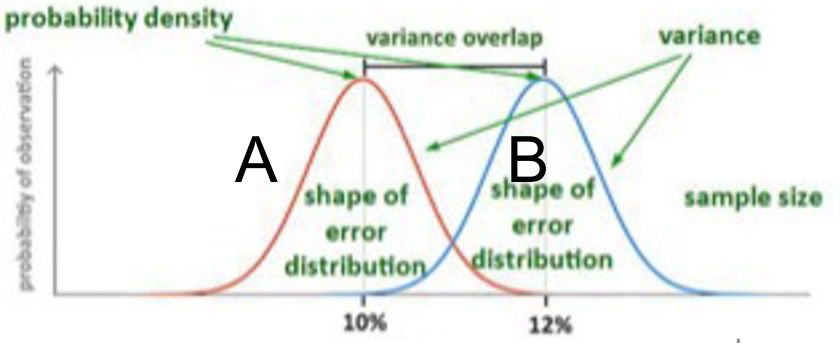
Intuition: Less overlap of A and B

- <u>More likely</u> an “effect” of Webpage B versus A
- “Webpage B increases click-through-rate (CTR)”



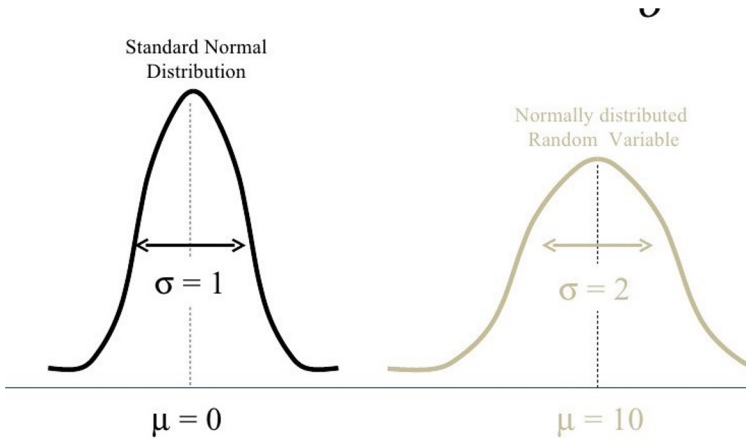
Intuition: More overlap of A and B

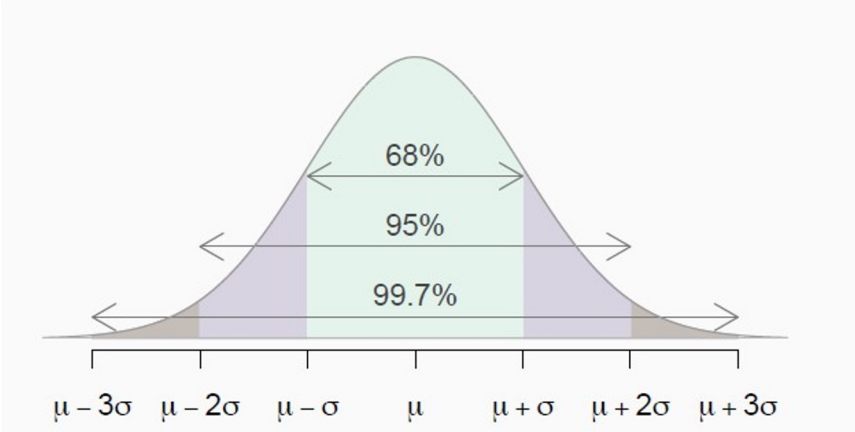
- <u>Less likely</u> there is an “effect” of B vs A
- “Webpage B doesn’t change CTR”



How to hypothesis test: converting data from A and B

- Data on A and B
  - Users that saw either Webpage A or Webpage B
  - "Distribution" of user data for each webpage
- Take the difference between webpage A and B distributions.
  - This creates a new distribution which we can convert to a "standard" Normal distribution
- Statistical Test
  - This allows us to use a "Z-Test" to test our hypotheses

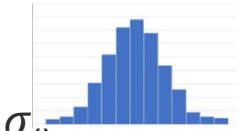




How to hypotheses test:

Mean and standard deviation of collected data

- Webpage A:
  - Mean μ_A
  - Standard Deviation: σ_A
  - Data sample i of CTR: a_i
  - Sample Size: n_A



- Webpage B:
  - Mean μ_B
  - Standard Deviation: σ_B
  - Data sample i of CTR: b_i
  - Sample Size: n_B

- Z-Statistic
  - Difference of Means
  - Scale by Standard Deviations

Mean and Standard Deviation of Webpage A and Webpage B

$$\hat{\mu}_A = \frac{1}{n_A}\sum_i a_i \text{ and } \hat{\sigma}_A = \frac{1}{n_A}\sum_i (\hat{\mu}_A - a_i)^2$$

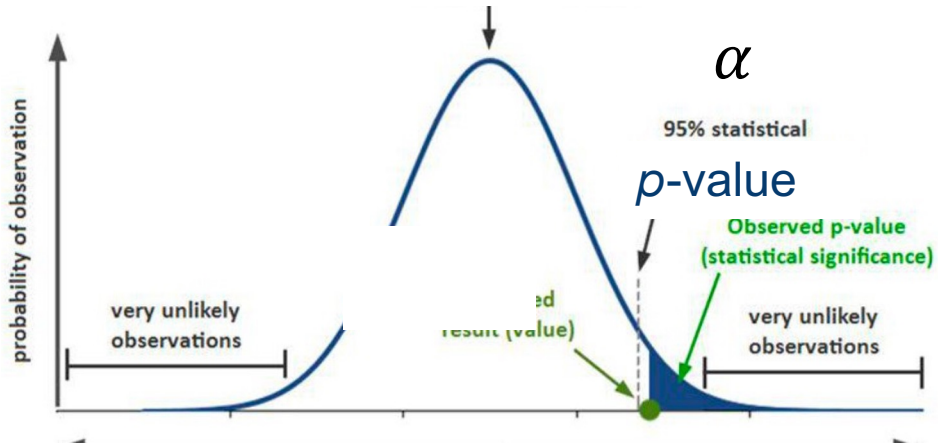
$$\hat{\mu}_B = \frac{1}{n_B}\sum_i b_i \text{ and } \hat{\sigma}_B = \frac{1}{n_B}\sum_i (\hat{\mu}_B - b_i)^2$$

"Z-Transformation of our Data"
(Difference between Webpage A and B)

$$z_{AB} = \frac{\hat{\mu}_B - \hat{\mu}_A}{\sqrt{\frac{\hat{\sigma}_B^2}{n_B} + \frac{\hat{\sigma}_A^2}{n_A}}}$$

STATISTICAL SIGNIFICANCE

- How?
  - Choose significance level α for rejecting the null hypothesis H_0 that webpage A and B have same CTR



- Reject H_0 if our *p*-value is **<= *a***
  - This means we find H1 true
  - Or more accurately, H1 is more consistent with the data

- H0: “the click through rate is the same for the two webpages”
- H1: “the click-through-rate is higher (or lower) for webpage B than A

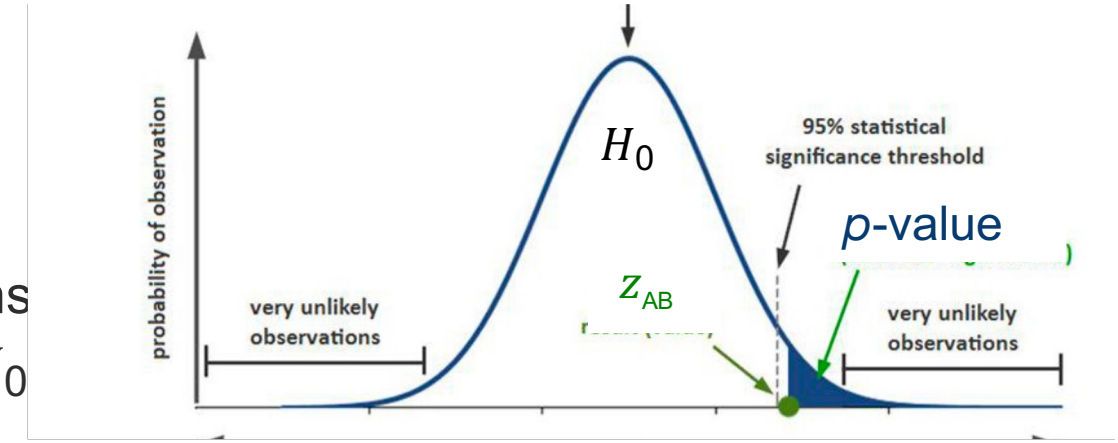
P-VALUE

- *p*-value – Measure of the strength of evidence the sample data provides against the null hypothesis H_0
  - We get *p*-value from z-statistic z_{AB}

$$p\text{-value} = P(Z \geq z_{AB}; H_0)$$

- *p*-value is the probability of the null hypothesis H_0 giving a more extreme difference between webpage A and B than the observed difference z_{AB}

- It's the amount of "tail" in the difference of means distribution, the standard Normal specified for H_0

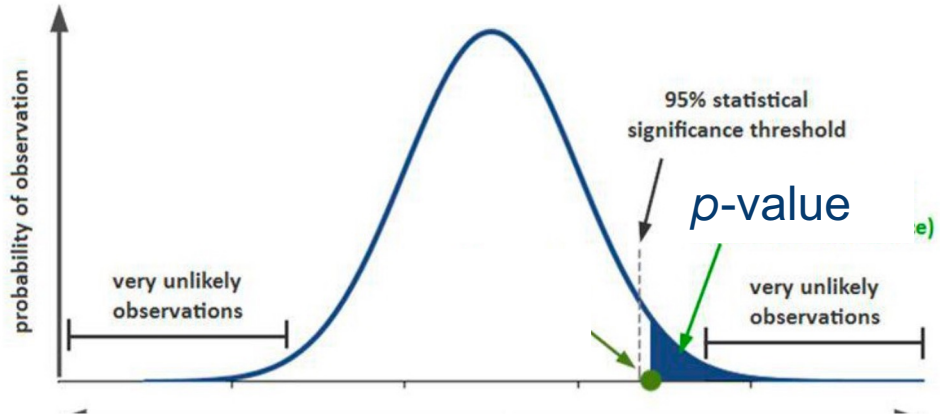


STATISTICAL SIGNIFICANCE: COMPARING *P*-VALUE WITH ALPHA

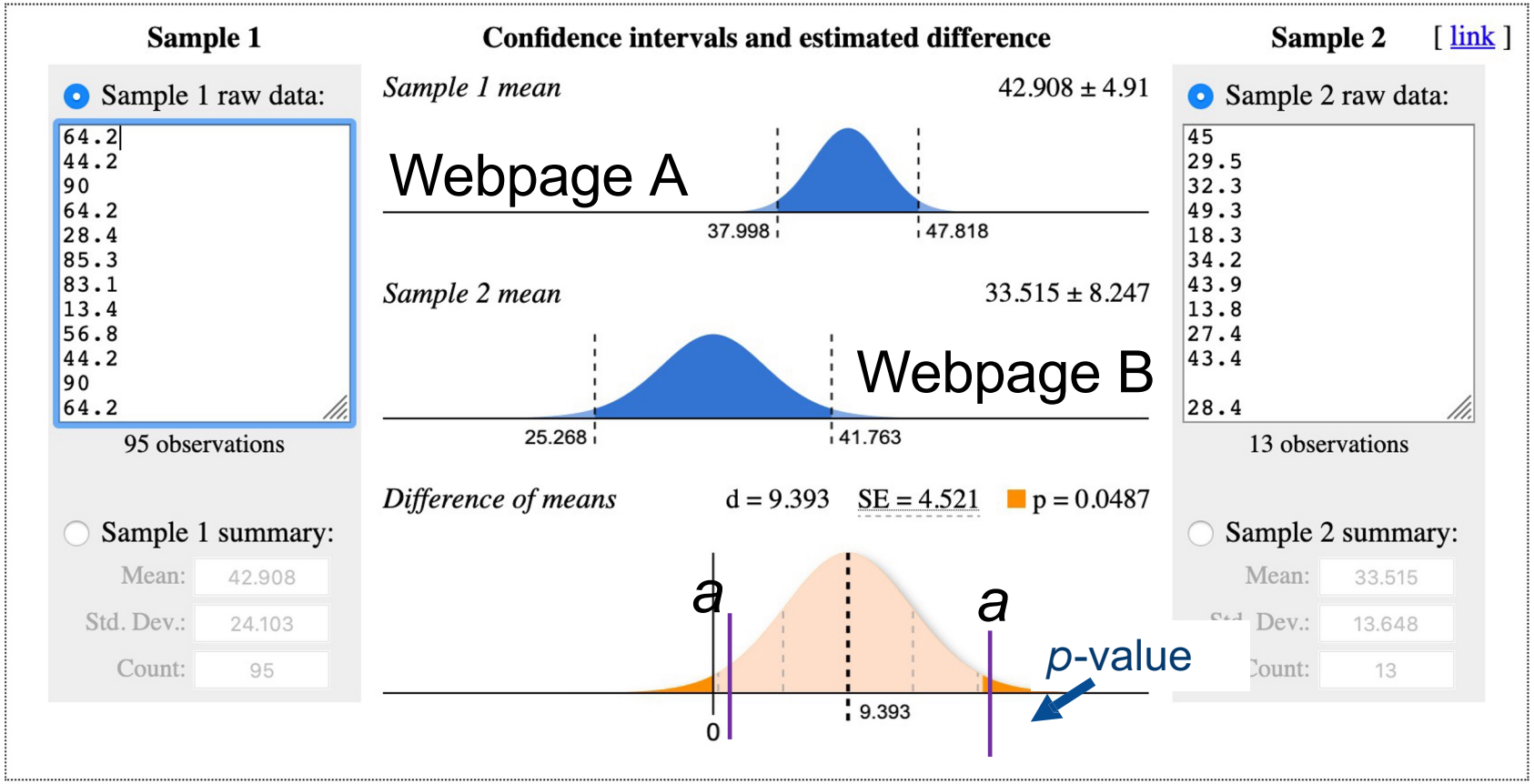
- We saw comparing p-value to alpha performs hypothesis test. But how do we interpret?

- *p*-value: We do not choose this. This is calculated from data.

- Alpha or "Significance Level"
  - We choose this. Default a is often 5%
  - You may have heard, "$p < 0.05$"
  - There is a 5% or less chance of "false positive'
  - There is a 95% chance of the new webpage A having higher CTR than the old webpage B. This interpretation is called the "confidence level"



STATISTICAL SIGNIFICANCE: OVERLAPPING DISTRIBUTIONS



Verdict: Sample 1 mean is greater

Webpage A is statistically significant in its CTR from Webpage B

QUESTION: DOES P-VALUE < 0.05 ALWAYS MEAN STATISTICALLY SIGNIFICANT?

- We choose the alpha. It just happens that many academic fields have adopted alpha of 0.05
  - This was introduced in 1925
  - It is an arbitrary choice.
  - Physicists for example used a p-value of 0.0000003 for the Higgs Boson

- p < 0.05 is very abused
  - With "internet scale" data, we can often "find" statistical significance.

Title: Redefine Statistical Significance

Authors: Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7,10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12,13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], James Holland Jones[34,40], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P.A. Ioannidis[37], Minjeong Jeon[38], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don Moore[49], Stephen L. Morgan[50], Marcus Munafò[51,52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71], Valen E. Johnson[72]*

Affiliations:

[1]Center for Economic and Social Research and Department of Economics, University of Southern California, Los Angeles, CA 90089-3332, USA.

[2]Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA.

[3]Department of Economics, Stockholm School of Economics, SE-113 83 Stockholm, Sweden.

[4]University of Virginia, Charlottesville, VA 22908, USA.

[5]Center for Open Science, Charlottesville, VA 22903, USA.

[6]University of Amsterdam, Department of Psychology, 1018 VZ Amsterdam, The Netherlands.

[7]University of Pennsylvania, School of Arts and Sciences and Department of Criminology, Philadelphia, PA 19104-6286, USA.

[8]University of North Carolina Chapel Hill, Department of Psychology and Neuroscience, Department of Sociology, Chapel Hill, NC 27599-3270, USA.

[9]Institute of Zoology - Neurogenetics, Universität Regensburg, Universitätsstrasse 31 93040 Regensburg, Germany.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature human behaviour*, *2*(1), 6-10.

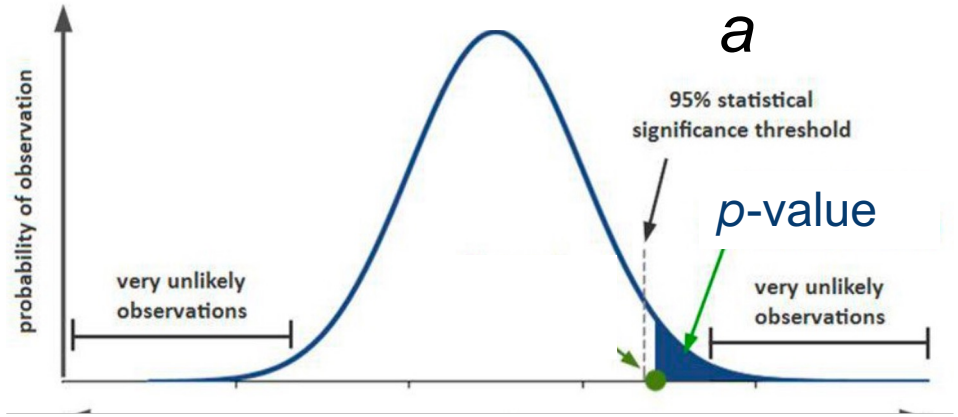
K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philosophical Magazine, vol. 50, no. 5, p. 157–175, 1900.

R. Fischer, Statistical Methods for Research Workers, Edinburgh: Oliver & Boyd, 1925.

QUESTION: WHAT CAN LEAD TO A LOW P-VALUE?

Observing a low *p*-value means either:

1. The null hypothesis is not true.
2. The null hypothesis is true, but we have observed a very rare outcome
3. The statistical model is inadequate so the calculated *p*-value is not an actual *p*-value.

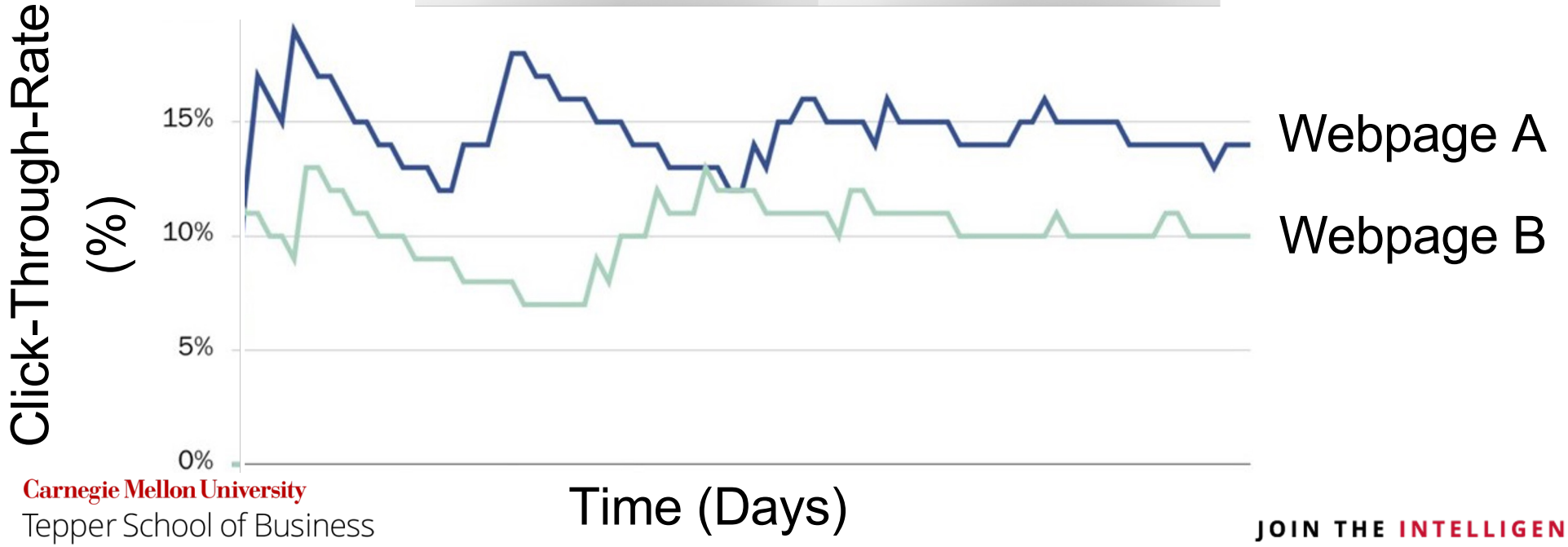


How to get a low p-value

- The larger the variance of the data, the larger the p-value, all else being equal.
- The larger the sample size, the smaller p-value, all else being equal.
- The larger the observed discrepancy, the smaller the p-value, all else being equal.

Recommended interactive visualization of how sample size and variance affect statistical significance: https://www.evanmiller.org/ab-testing/t-test.html

QUESTION: HOW LONG SHOULD WE RUN A/B TEST?



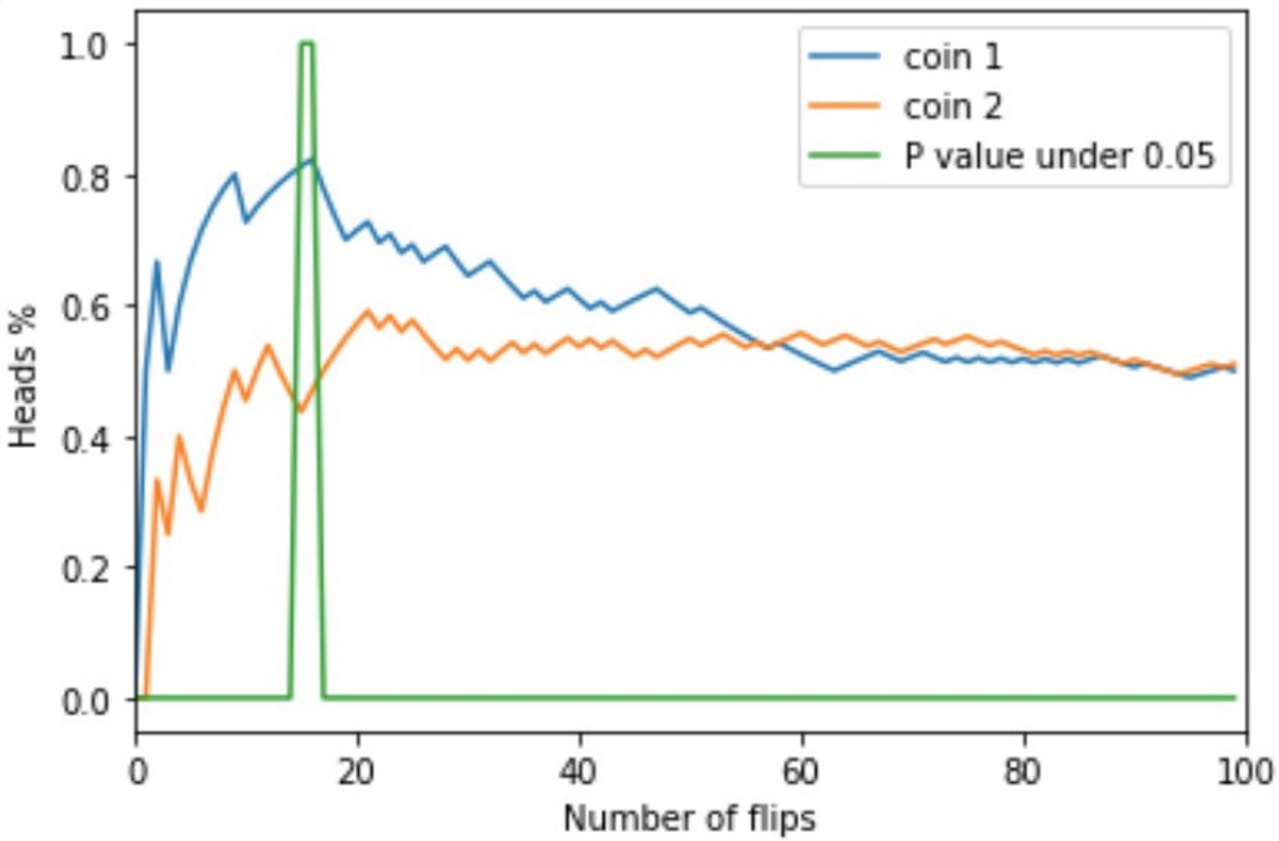
Carnegie Mellon University
Tepper School of Business

JOIN THE INTELLIGENT FUTURE

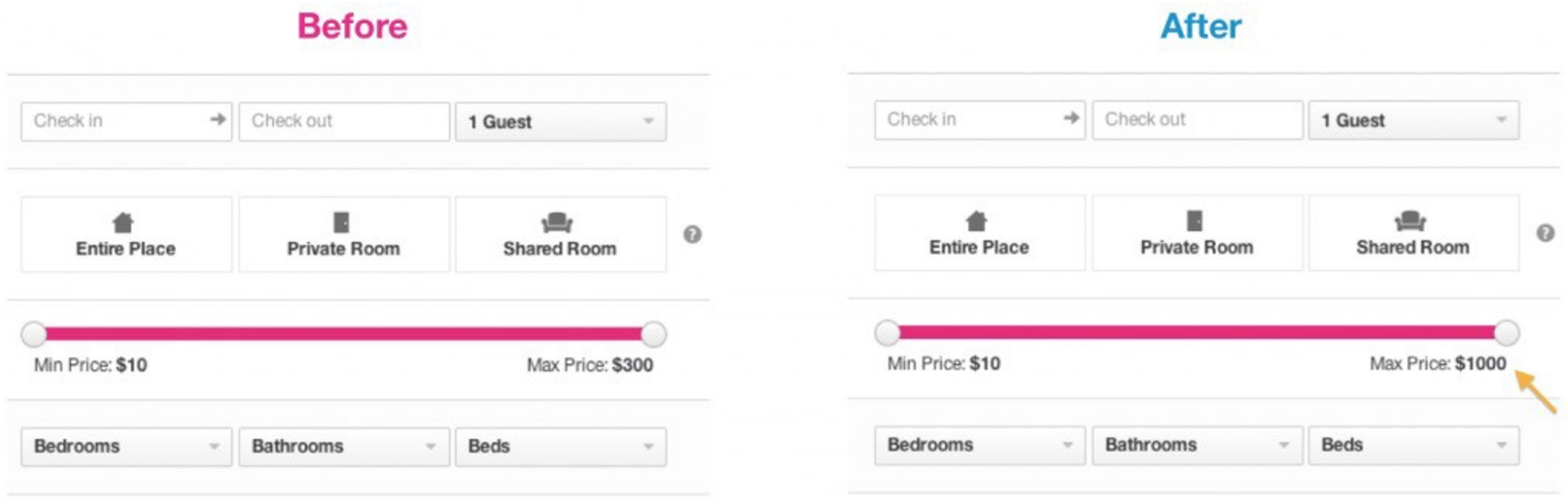
COMMON PITFALL: EARLY STOPPING

- Imagine two coins
  - Coin 1: 50% Heads
  - Coin 2: 50% Heads

- "If we stopped when we saw a "significant effect" we would have said the 2 coins were different



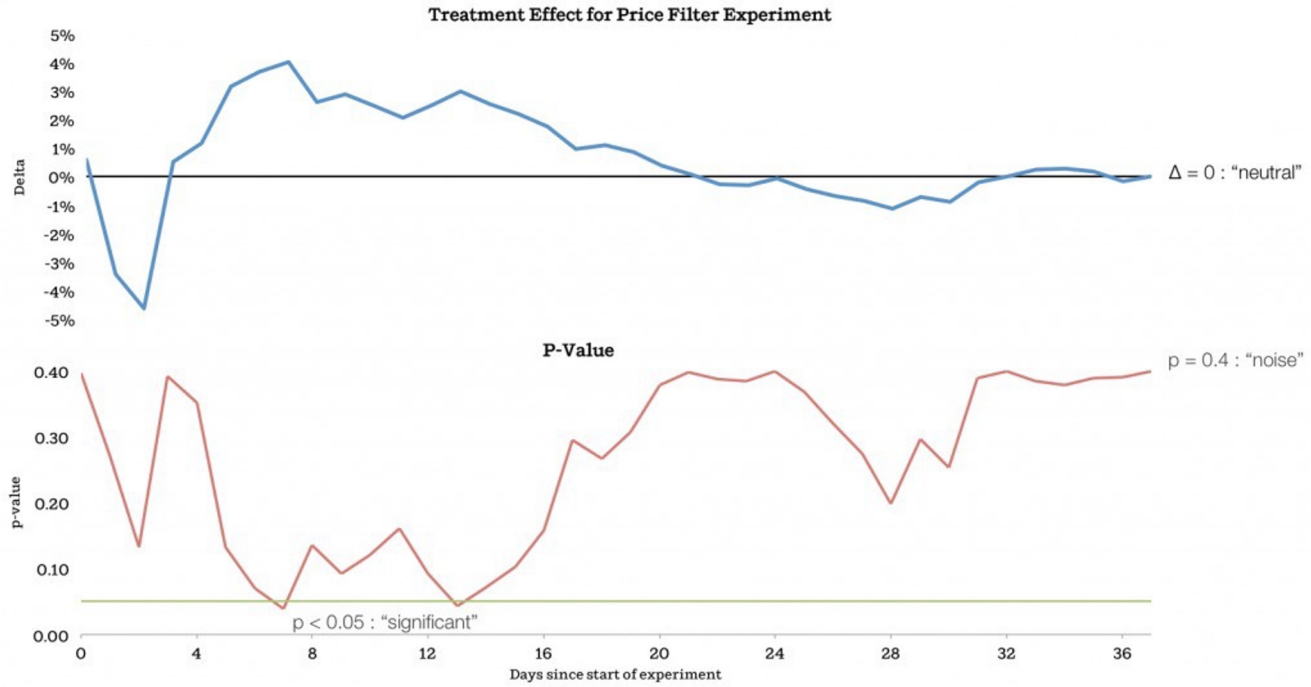
REAL EXAMPLE: AIRBNB



- Changing the Maximum Price Filter on Search Page

Source: https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7

REAL EXAMPLE: AIRBNB



Takeaway: If Airbnb stopped when it was "significant" then this would have been (an incorrect) spurious correlation.

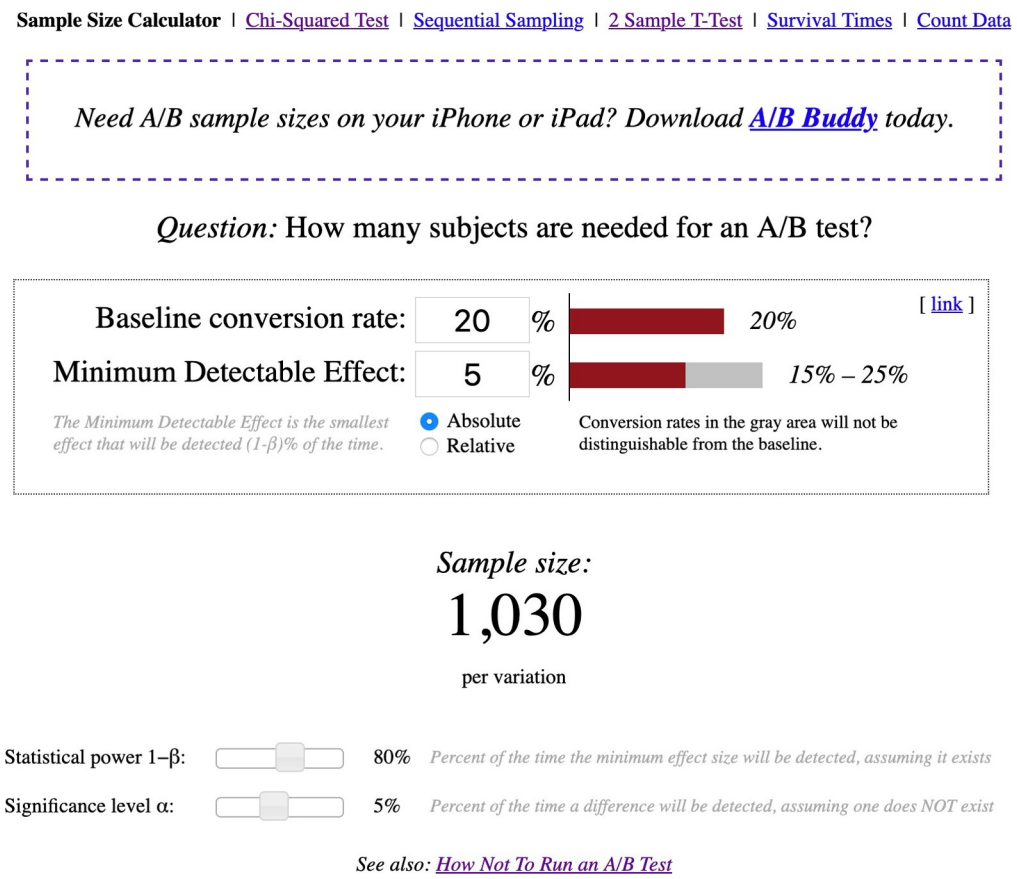
How many samples do I need?

Equation are based on desired detection of effect size (from base rate) and standard deviations of A and B

Use sample size calculators:

- Many online
- Link below is good one

https://www.evanmiller.org/ab-testing/sample-size.html



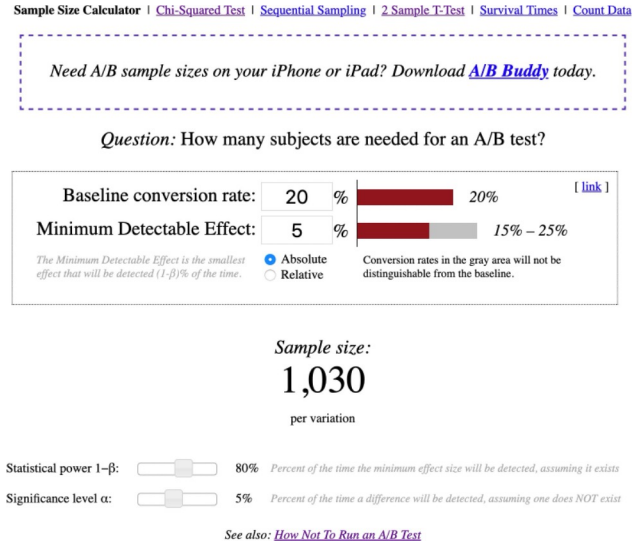
In-class exercise: How many samples would you need and how many days needed?

Go to sample size calculator

- https://www.evanmiller.org/ab-testing/sample-size.html

Question 1: How many samples needed for?

- Significance Level = 5%
- False Error Rate = 20% (Power = 80%)
- Baseline Metric =40%
- Minimum Effect = 6% (Relative)
- Average Daily Volume (Unique Users) = 1,500
- % of unique users eligible for test = 10%



Question 2: How many days will you need for an A/B test with two variations?

Credit: Evan Miller

TODAY'S LEARNING OBJECTIVES

- What and Why A/B Test?
  - Focus on Intuition and Definitions
  - 3 Steps for A/B Testing
- How to A/B Test?
  - Statistical Significance
  - Concepts: p-values, confidence intervals, etc.