A/B Testing in the Wild

How to Get Started

Charin Polpanumas
Lead Data Scientist @ Central

Which Sunflower Seed Packet Will Sell Better?

Gut feeling - what does your instinct tell you?

Red Pack
(Current)

Grey Pack
(Experimental)

Hamster Inc. is a large food-and-beverage corporation operating in Hamland. One of its flagship products is sunflower seed packets. It is considering changing packaging color from red to grey.

Gut Feeling

Why would it be a bad idea?

Which Sunflower Seed Packet Will Sell Better?

Extrapolation from past experiences

Red Pack
(Current)

Grey Pack
(Experimental)

Hamster Inc. has 21 products. 5 products have red packaging and are selling very well. 3 products have grey packaging; they are not selling as well as the red products.

Extrapolation from Past Experiences

Why would it be a bad idea?

Hamster Inc. has 21 products. 5 products have red packaging and are selling very well. 3 products have grey packaging; they are not selling as well as the red products.

- All 5 red products are flagship products that might do well regardless of packaging color
- All 3 grey products might be experimental products that customers are not familiar with yet
- Some products might be totally different than sunflower seeds thus not comparable
- How long back into the past should we look when comparing red vs grey products

Which Sunflower Seed Packet Will Sell Better?

Proxies - ask potential customers what they want

Red Pack
(Current)

Grey Pack
(Experimental)

Hamster Inc. hired McGuinea and Company to conduct a survey of 2,000 respondents asking if they prefer red or grey sunflower seed packets. 820 respondents prefer red, 750 prefer grey and 430 say they are indifferent.

Proxies

Why would it be a bad idea?

Hamster Inc. hired McGuinea and Company to conduct a survey of 2,000 respondents asking if they prefer red or grey sunflower seed packets. 820 respondents prefer red, 750 prefer grey and 430 say they are indifferent.

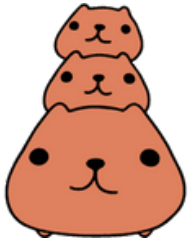
- Answering questionnaires does not have a cost; actually buying the product does
- The Pepsi Challenge Effect; customers might prefer one packet in an experimental setting and the other in real-world settings
- Respondents might not be representative of actual customers
- What is a good-enough margin to declare winner

Which Sunflower Seed Packet Will Sell Better?

Fear of missing out

Red Pack
(Current)

Grey Pack
(Experimental)



Kapibara Syndicate, our rival firm, has debuted their grey sunflower seed packets. The product has done really well in the market and at this rate we might lose market share to them.

Fear of Missing Out

Why would it be a bad idea?

Kapibara Syndicate, our rival firm, has debuted their grey sunflower seed packets. The product has done really well in the market and at this rate we might lose market share to them.

- Kapibara Syndicate sunflower seed packets might do well because of other factors such as taste and price
- Our target customers and Kapibara Syndicate's might be totally different
- We do not know how much of the sales are a result of marketing campaigns to promote their new products, not a result of the grey packaging itself

Which Sunflower Seed Packet Will Sell Better?

Highest Paid Person's Opinion

Red Pack
(Current)

Grey Pack
(Experimental)

Dr. Hamtaro Kamado, CEO of Hamster Inc. and Forbes 100 Most Influential Rodents, came up with the grey packet idea. He has pioneered all of Hamster Inc. successful products so far.

Highest Paid Person's Opinion

Why would it be a bad idea?

Dr. Hamtaro Kamado, CEO of Hamster Inc. and Forbes 100 Most Influential Rodents, came up with the grey packet idea. He has pioneered all of Hamster Inc. successful products so far.

- The brightest minds are not always right. Even Steve Jobs has made some mistakes like Apple III and Lisa
- HiPPOs might be paid highly because of their abilities irrelevant to packet design; case-in-point Henry Kissinger on Theranos' board
- HiPPOs might have their own incentives

How Business Decisions Are Made

- Realistic assumptions
- Scientific methods

- Gut feelings
- Extrapolation from past experiences
- Proxies
- Fear of missing out (FOMO)
- Highest Paid Person's Opinion (HiPPO)

Randomness in Business Decisions

Customer preferences, trends and seasonality, competitors

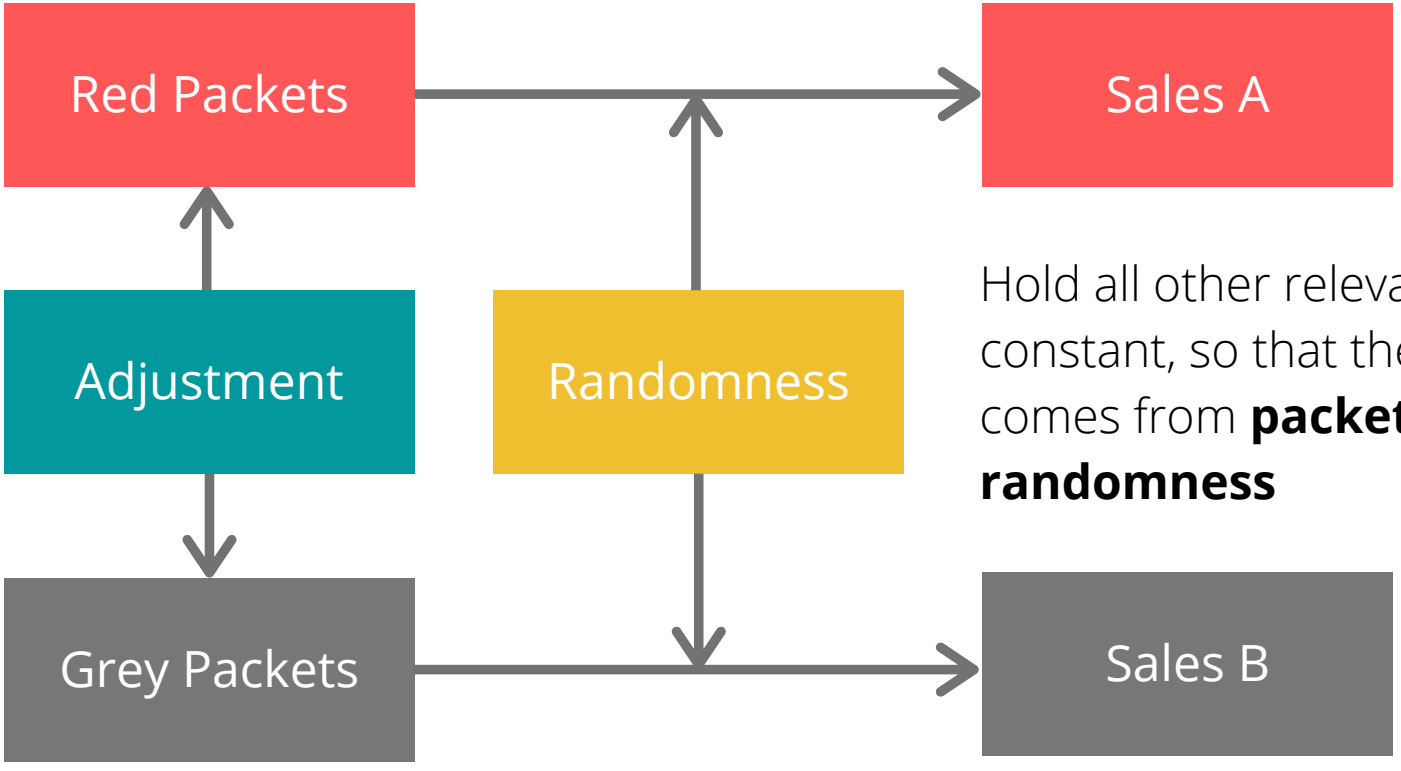
- Chess
- Go
- Checkers
- Santorini

- UI/UX improvements
- Marketing campaigns
- Product development
- Clinical trials

※ We will define this properly later

How to Deal with Randomness

Everything else being equal, what is the effect of actions on outcomes



Hold all other relevant factors constant, so that the only difference comes from **packet color** and **randomness**

A/B Test Use Case Montage

Who have done it? Was it worth it?

Bing Link Color Changes (2014)

Bing made 10M USD more from ads with a different link color

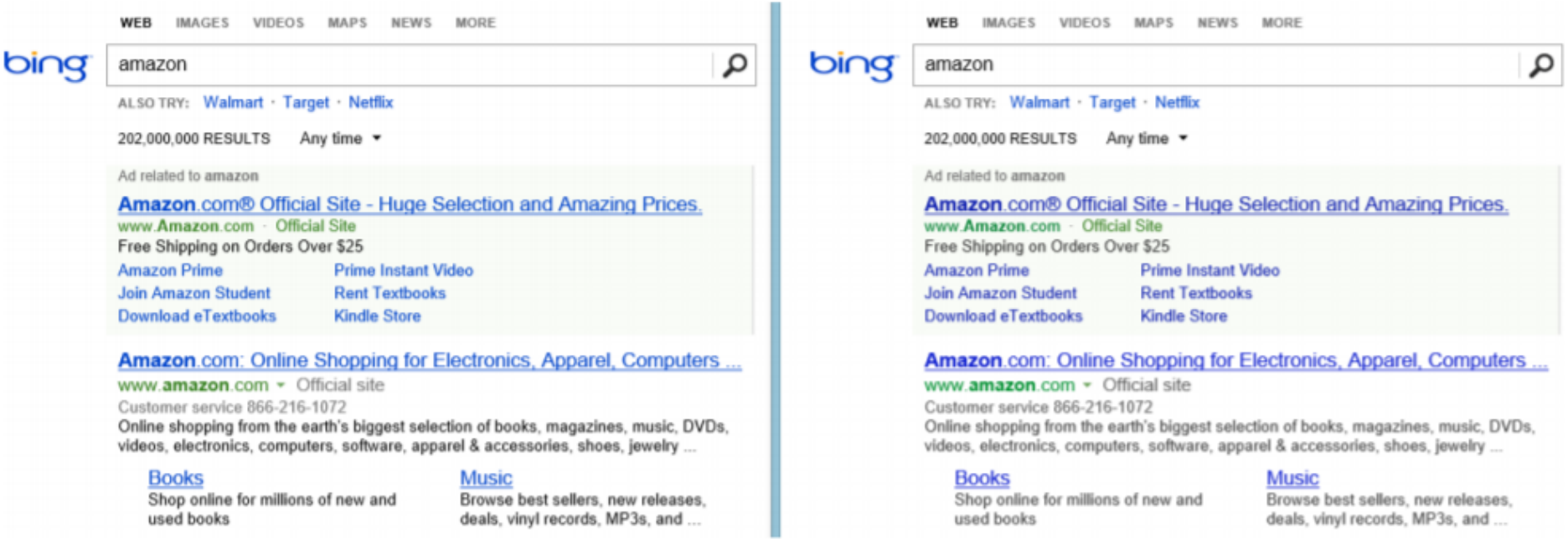


Figure 1: Font color experiment. Can you tell the difference?

Speed Matters for Google Web Search (2009)

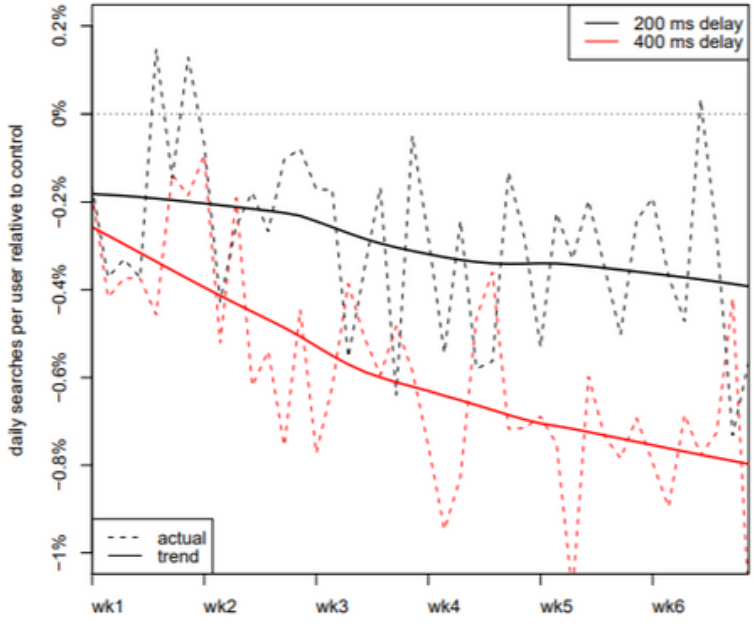
Longer website load time impacts daily usage per user

Table 1: Experiment Impact on Daily Searches Per User

Type of Delay	Magnitude	Duration	Impact
Pre-header	50 ms	4 weeks	—
Pre-header	100 ms	4 weeks	−0.20%
Post-header	200 ms	6 weeks	−0.29%
Post-header	400 ms	6 weeks	−0.59%
Post-ads	200 ms	4 weeks	−0.30%

Average impact over 4 or 6 weeks hides any trend over time. By focusing on the subset of users who were part of the experiment (or control group) from the beginning (as identified by a browser cookie), one can determine if there is such a trend. Figure 2 illustrates the trend for the two 6 week experiments.

Figure 2: Impact of Post-header Delays Over Time



A/B Shark Doo Doo Doo Doo (2018)

In all, we have probably 30 or 40 different variations of Baby Shark

They have also re-recorded the song with different themed beats, including R&B and EDM versions. “In all, we have probably 30 or 40 different variations of Baby Shark,”

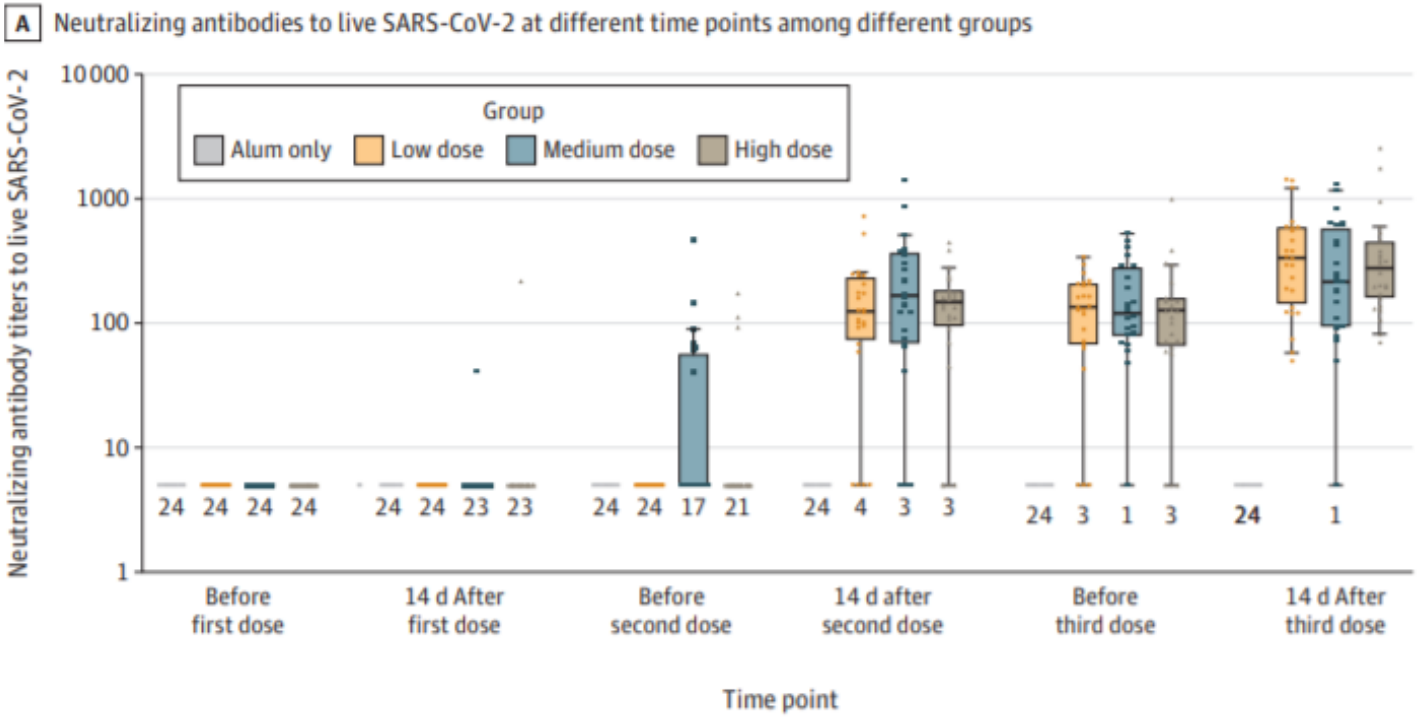
Jamie Oh
Pinkfong’s global marketing director

SARS-CoV-2 Vaccine Clinical Trials (2020)

Randomized controlled trials are A/B tests

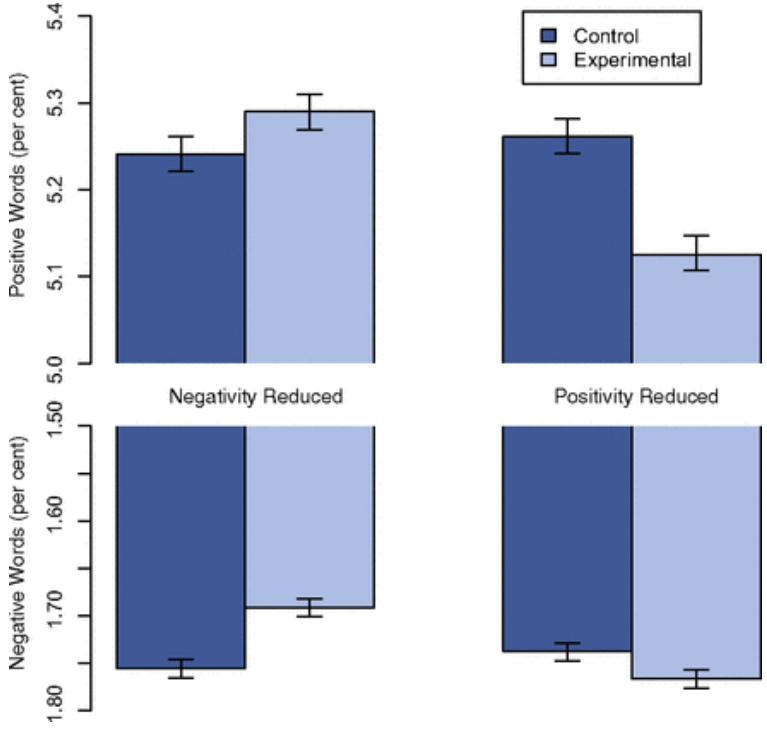
INTERVENTIONS In the phase 1 trial, 96 participants were assigned to 1 of the 3 dose groups (2.5, 5, and 10 μg/dose) and an aluminum hydroxide (alum) adjuvant-only group (n = 24 in each group), and received 3 intramuscular injections at days 0, 28, and 56. In the phase 2 trial, 224 adults were randomized to 5 μg/dose in 2 schedule groups (injections on days 0 and 14 [n = 84] vs alum only [n = 28], and days 0 and 21 [n = 84] vs alum only [n = 28]).

Figure 2. Antibody Responses at Different Time Points in the Phase 1 Trial



Facebook Emotional Contagion Experiment (2014)

People will say more positive or negative things according to what they read



The experiment manipulated the extent to which people (N = 689,003) were exposed to emotional expressions in their News Feed. This tested whether exposure to emotions led people to change their own posting behaviors, in particular whether exposure to emotional content led people to post content that was consistent with the exposure.

Step-by-Step Guide to A/B Test in the Wild

1. Define relevant metrics
2. Split samples into comparable groups
3. Choose statistical tests and validate their assumptions
4. Decide on stopping criteria
5. Run and monitor the experiment
6. Analyze results and suggest actions

Which Sunflower Seed Packet Will Sell Better?

Why don't we A/B test it

เมล็ดทานตะวันอบรสพุทราจีน เม็ดใหญ่ สดใหม่ เคี้ยวเพลิน ขนาด 500 กรัม Sunflower Seed ตราอีสือเจีย

★★★★★ 28 Ratings

Brand: No Brand | More Chocolate, Snacks & Sweets from No Brand

฿50.00

~~฿70.00~~ -29%

Quantity 1

Buy Now | Add to Cart

www.hamzada.com/sunflower-seeds?variation=A

เมล็ดทานตะวันอบรสพุทราจีน เม็ดใหญ่ สดใหม่ เคี้ยวเพลิน ขนาด 500 กรัม Sunflower Seed ตราอีสือเจีย

★★★★★ 28 Ratings

Brand: No Brand | More Chocolate, Snacks & Sweets from No Brand

฿50.00

~~฿70.00~~ -29%

Quantity 1

Buy Now | Add to Cart

www.hamzada.com/sunflower-seeds?variation=B

Dr. Hamtaro Kamado decided to A/B test the packaging on hamzada.com—an e-commerce website. The goal is to find out which packet color customers **will more likely buy**.

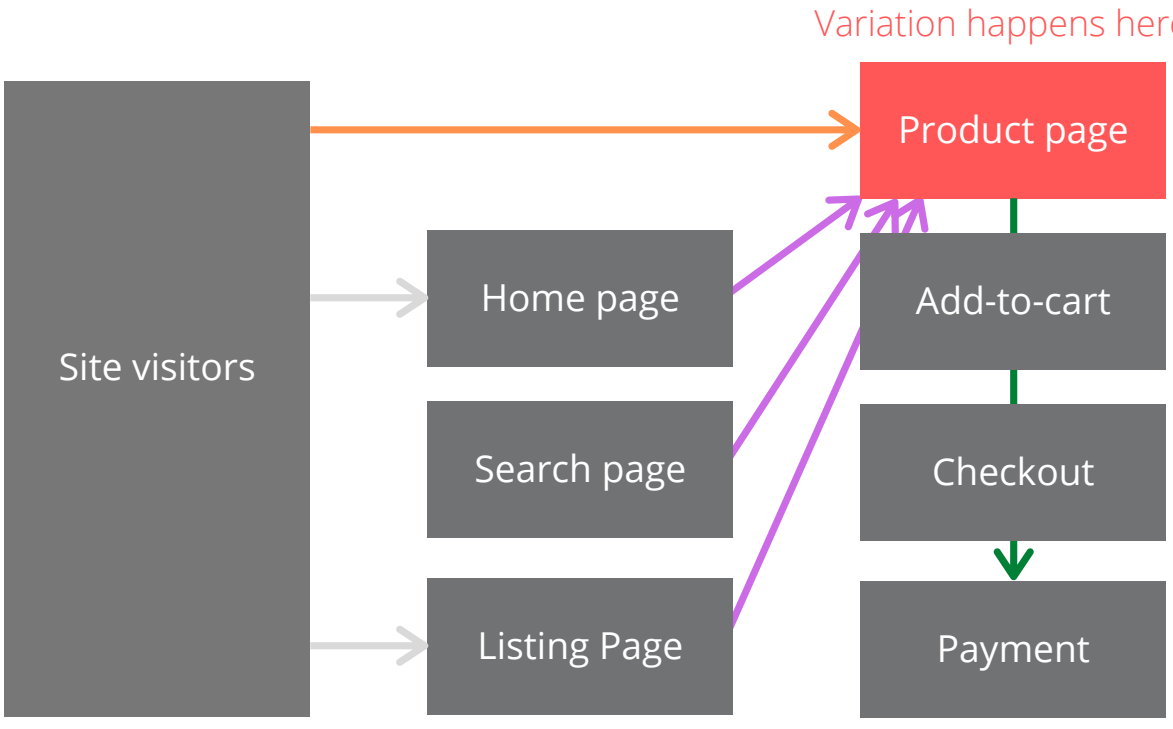
Define Relevant Metrics

Build intuition about your system

1. Relevance vs vanity
2. Granularity
3. Attribution periods
4. Assumed distributions
5. Robustness vs sensitivity
6. Features vs targets

Relevance vs Vanity

Understand the context of the problem



Relevant metrics

1. conversions / conversion rates
2. add-to-carts / add-to-cart rates
3. checkouts / checkout rates

Vanity metrics

1. clicks / click through rates
2. product page views / visitors

Granularity

Most common metrics in A/B tests are event probabilities

Possible denominators

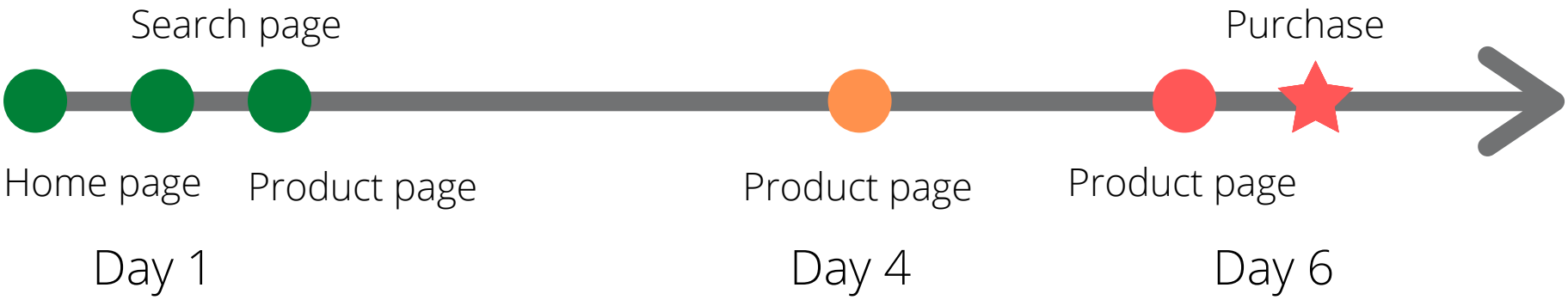
1. **Hits/Impressions** - number of times a page is viewed
2. **Sessions** - number of times the website if visited; timeout after 15-30 minutes
3. **Cookies** - number of cookies used to visit the website/page
4. **User Ids/Reach** - number of registered users who visited the website/page
5. **Device Ids** - number of devices who visited the website/page
6. **People** - number of actual people who visited the website/page

Possible event probabilities

1. **# checkout events / # hits** - double-count on page refreshes
2. **# checkout events/ # sessions** - double-count on inactive visits; good to see which products get bought within fewer visits
3. **# checkout events / # cookies on product page** - "users" as denominator; includes both logins and non-logins; different browsers/devices double-counts
4. **# payment events / # cookies on product page** - captures successful purchases
5. **# payment events / # user ids** - non-logins count as one user id
6. **# payment events / # people** - who are people?

Attribution Periods

Customer journey as denominator

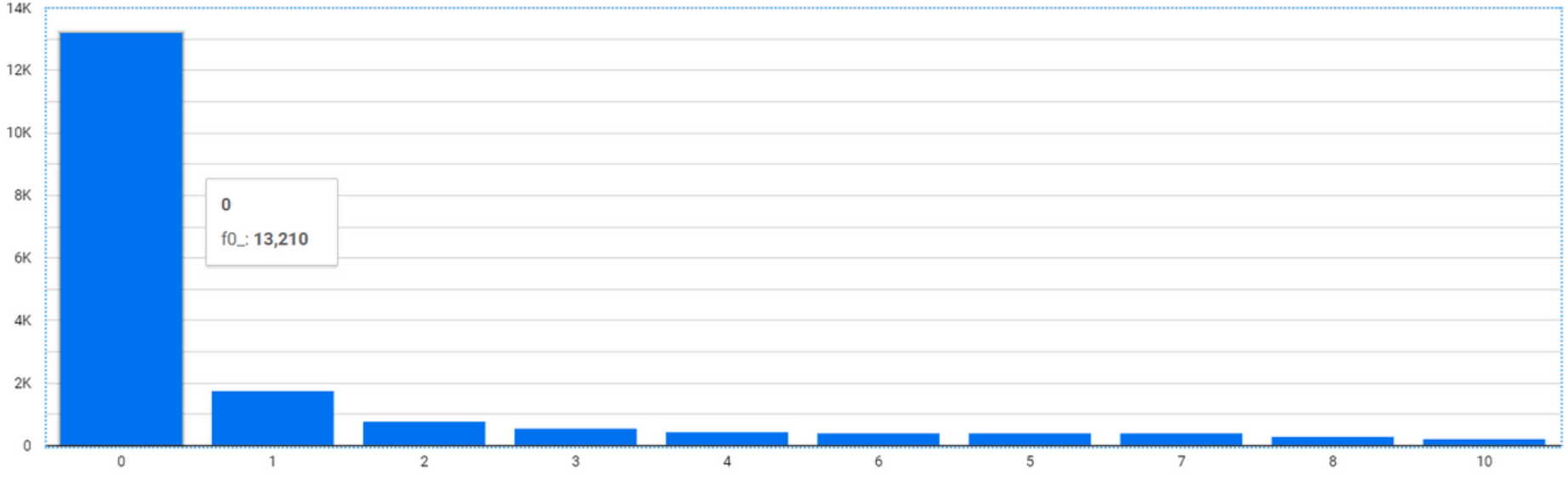


payment events / # cookies per X day attribution

One-day attribution - 3 journeys; 1 success
Three-day attribution - 2 journeys; 1 success
One-week attribution - 1 journey; 1 success

Attribution Periods

How long is one customer journey? Plot it out!



Most people (cookies?) who visit e-commerce websites might end their journey in one day, whereas more complex sales might take weeks or months

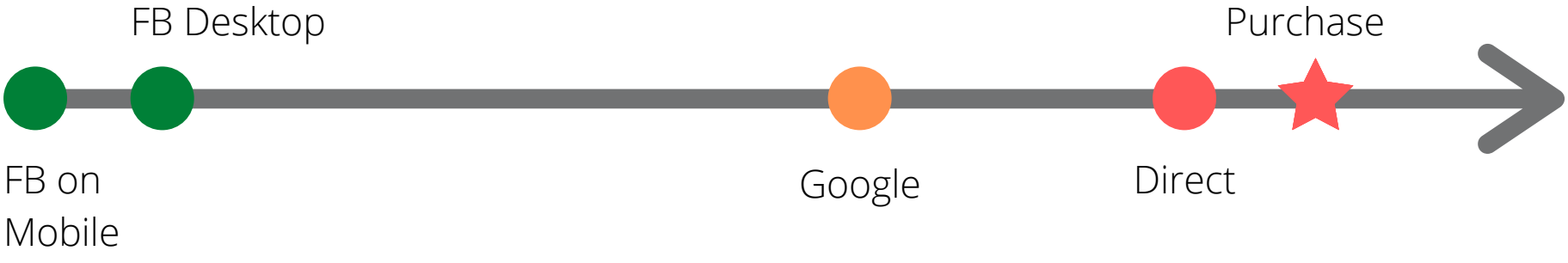
Attribution Periods

Event-based vs cohort-based

Conversion rate of August
conversions within August / number of users that visited in August

Conversion rate of August cohort
conversions within X days / number of users that visited in August

Example: Ads attribution



Assumed Distributions

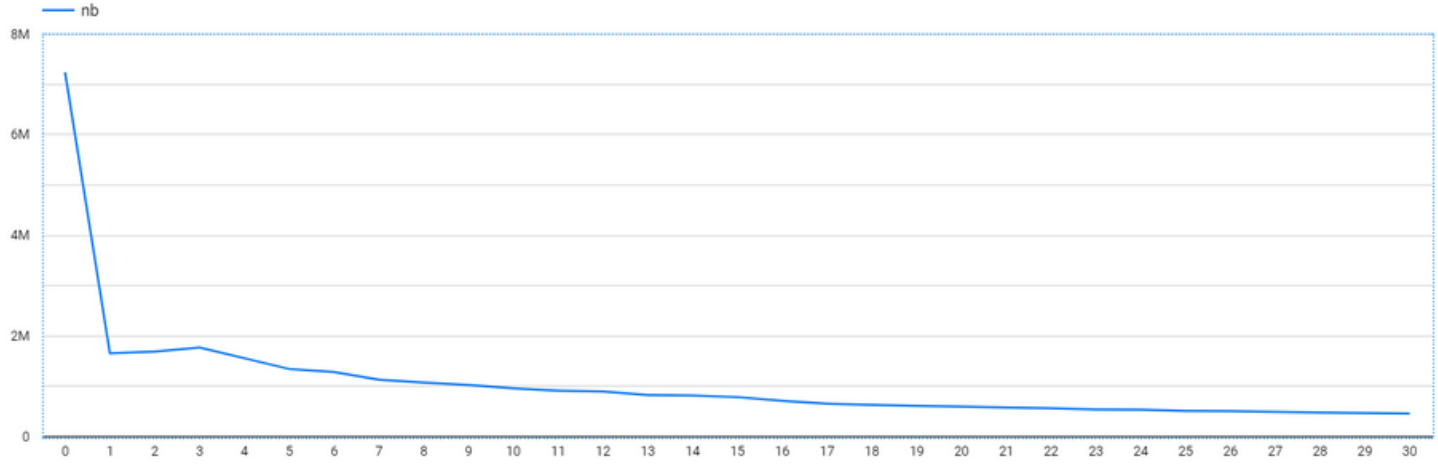
Put the right assumptions on the right data

Frequentist hypothesis testing requires us to know the variance of the data.
We typically assume the following distributions for the following data:

1. **Bernoulli** - event probability
2. **Exponential** - average revenue per user, average position clicked
3. **Normal/Student t** - height, weight, test scores, measurement errors

Also they ~~must~~ should be independent and identically distributed

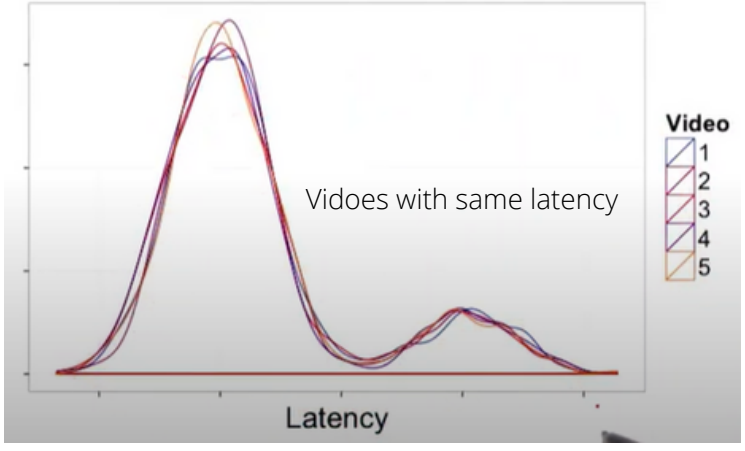
How can we be sure? Plot it out!
Example: average position clicked from listing page

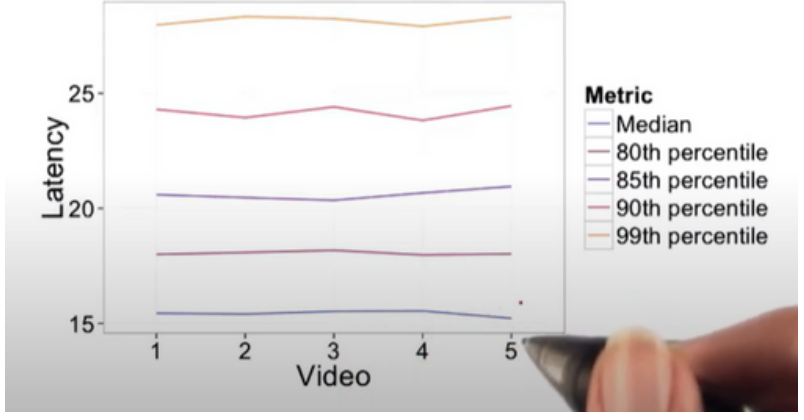


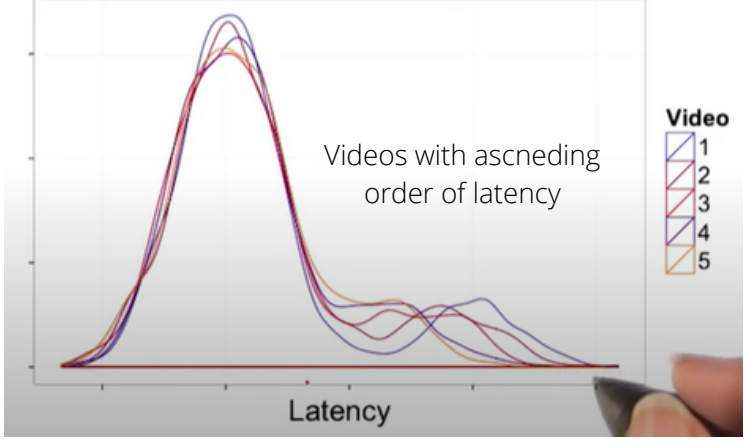
Robustness vs Sensitivity

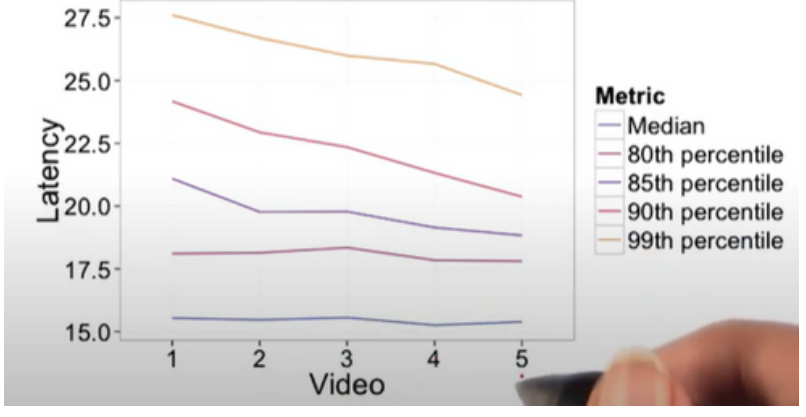
When you need to choose summary statistics to test

Example: summary statistics to represent YouTube video latency









Features vs Targets

Sometimes you train on one thing and optimize for the other

Problem: search re-ranking models that are trained on product clicks (since clicks are less sparse than conversions) make conversion rate worse

variant	#session	CTR	CR
original	3,322	43.86%	10.84%
model	3,281	43.68%	10.12%

Solution: makes product clicks count as 1 and conversions count as 2

variant	#session	CTR	CR
original	3,653	45.63%	13.09%
model	3,685	46.54%	14.11%

Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)

Split Samples into Comparable Groups

Build intuition about your system

1. Split independently
2. Control for X
3. Frame the experiment fairly
4. Multiple experiments

Split Independently

Hypothesis tests rely on independent and identically distributed data

- If you are splitting on a larger granularity than what you are measuring, for example splitting on cookies while measuring at hit level, you will need more samples for the test to have the same power
  - **Hits** - measure usability; for example, does UI change make people more likely to click; might change on refresh
  - **Sessions** - measure usability but doesn't change within session
  - **Cookies** - same "users" always get the same results; for example, testing 2 recommendation models; changes only when browsers/devices change
  - **User Ids** - more strict than 3. but does not work on non-logins
- Hits and sessions are likely to be NOT independent since the same users can be recorded for both control and test groups
- Avoid making your groups interact with one another; for example, if you are testing out a new function on social media, avoid doing it with users who can interact with one another.

Split Independently

How adventurous is your team

50/50
Shorter time
Higher risks

80/20
Longer time
Lower risks

Control for X

Simpson's paradox - control for confounders

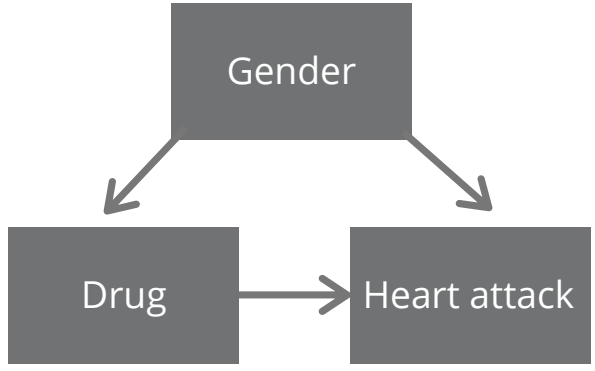
Story 1: "The Bad / Bad / Good Drug"

	Control Group (no drug)		Treatment Group (took drug)		
	Heart Attack	No HA	Heart Attack	No HA	
Female	1	19	3	37	60
Male	12	28	8	12	60
Total	13	47	11	49	

5% (1/20) women in **control** group had heart attack
7.5 % (3/40) women in **treatment** group had heart attack
→Drug is bad for women

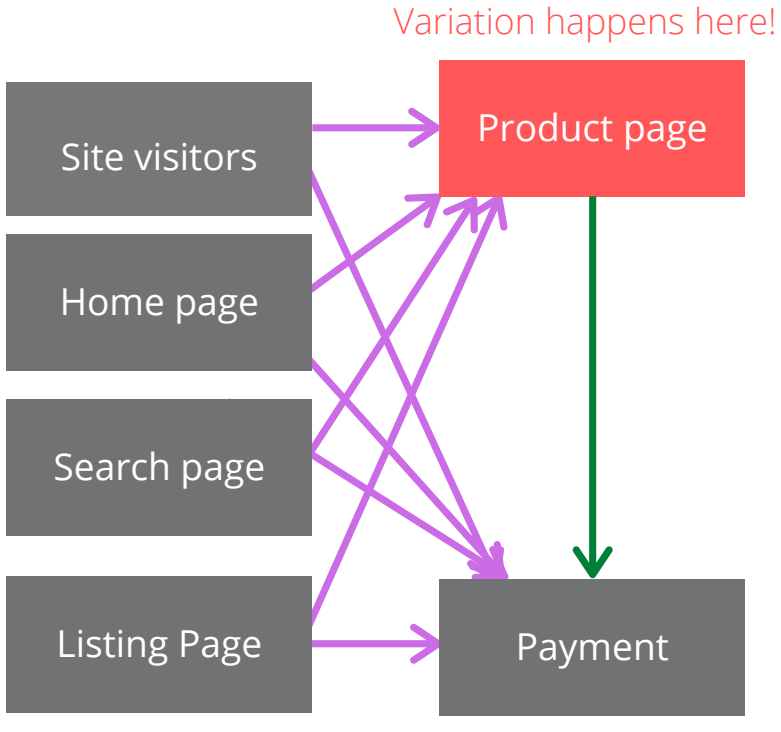
30% (12/40) men in **control** group had heart attack
40% (8/20) men in **treatment** group had heart attack
→Drug is bad for men

22% (13/60) **controls** had heart attack
18% (11/60) **treatment** group had heart attack
→Drug is good for people



Control for X

Simpson's paradox - control for confounders



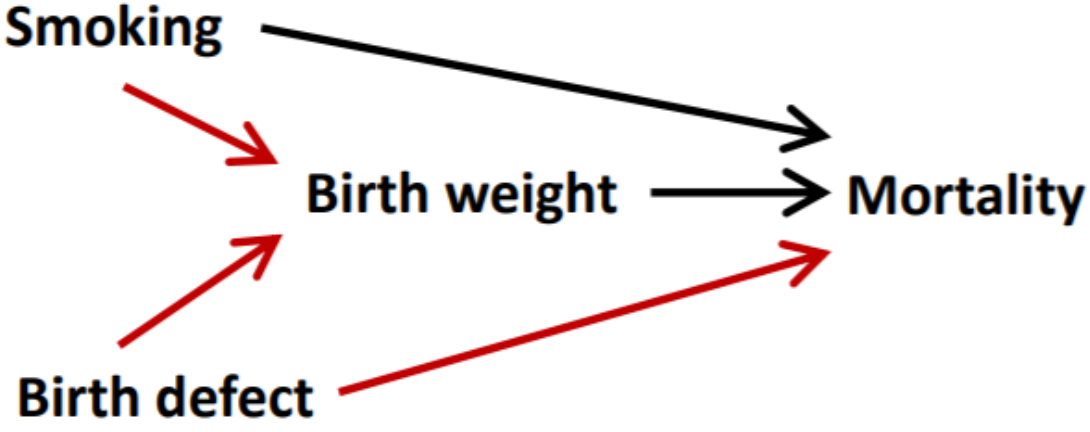
We can consider how the users come to product page as a confounder:

1. Search users might convert more easily because they are specifically looking for the sunflower seeds
2. Direct site visitors might convert less since they probably came from ads

(Do Not) Control for X

Smoking mothers - DO NOT control for colliders

Mid 1960s study by Yerushalmy on 15,000 children found that low birthweight babies of smoking mothers had better survival than low birthweight babies of nonsmoking mothers.



Intuition-based explanation is that babies with low birthweight from non-smoking mothers are likely ones who have some serious birth defects that are more serious than having a smoking mother.

Frame the Experiment Fairly

Variation/Treatment/Intervention can impact how groups behave

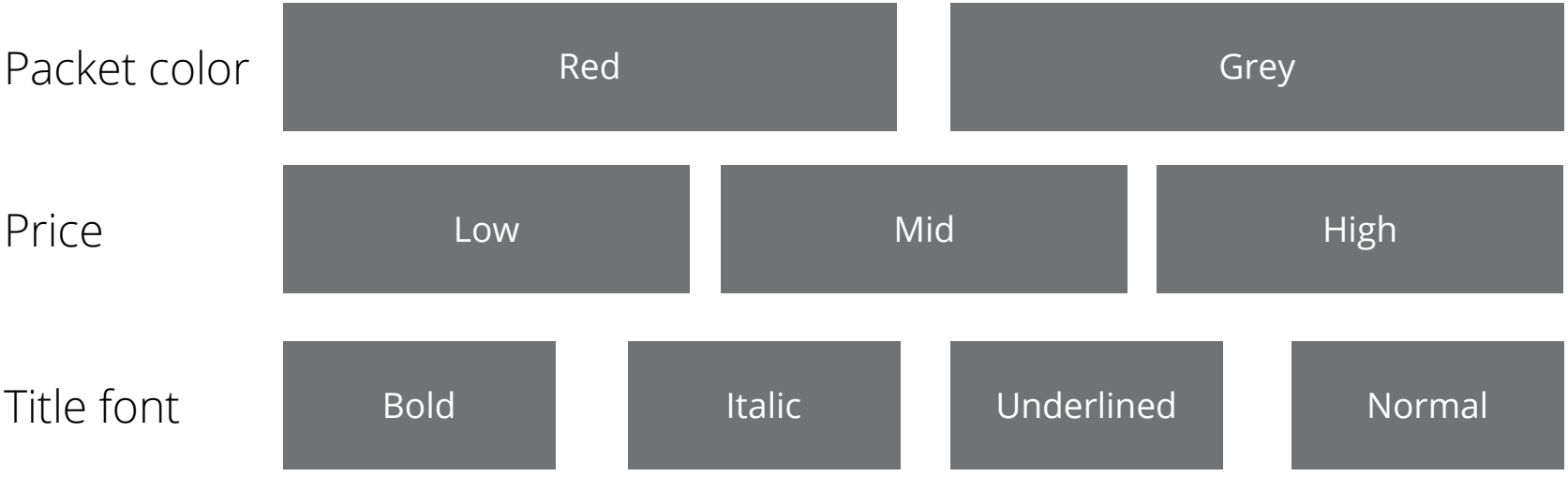
Control: comment at bottom

Test: comment on top

Multiple Experiments

The more you slice and dice, the more you need to be sure they are random

If you are running multiple experiments at the same time, you need to make sure than each other experiment is controlled for. For instance, if red variation got more low price, it might convert better but not because it was red but because the product was cheaper.



Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

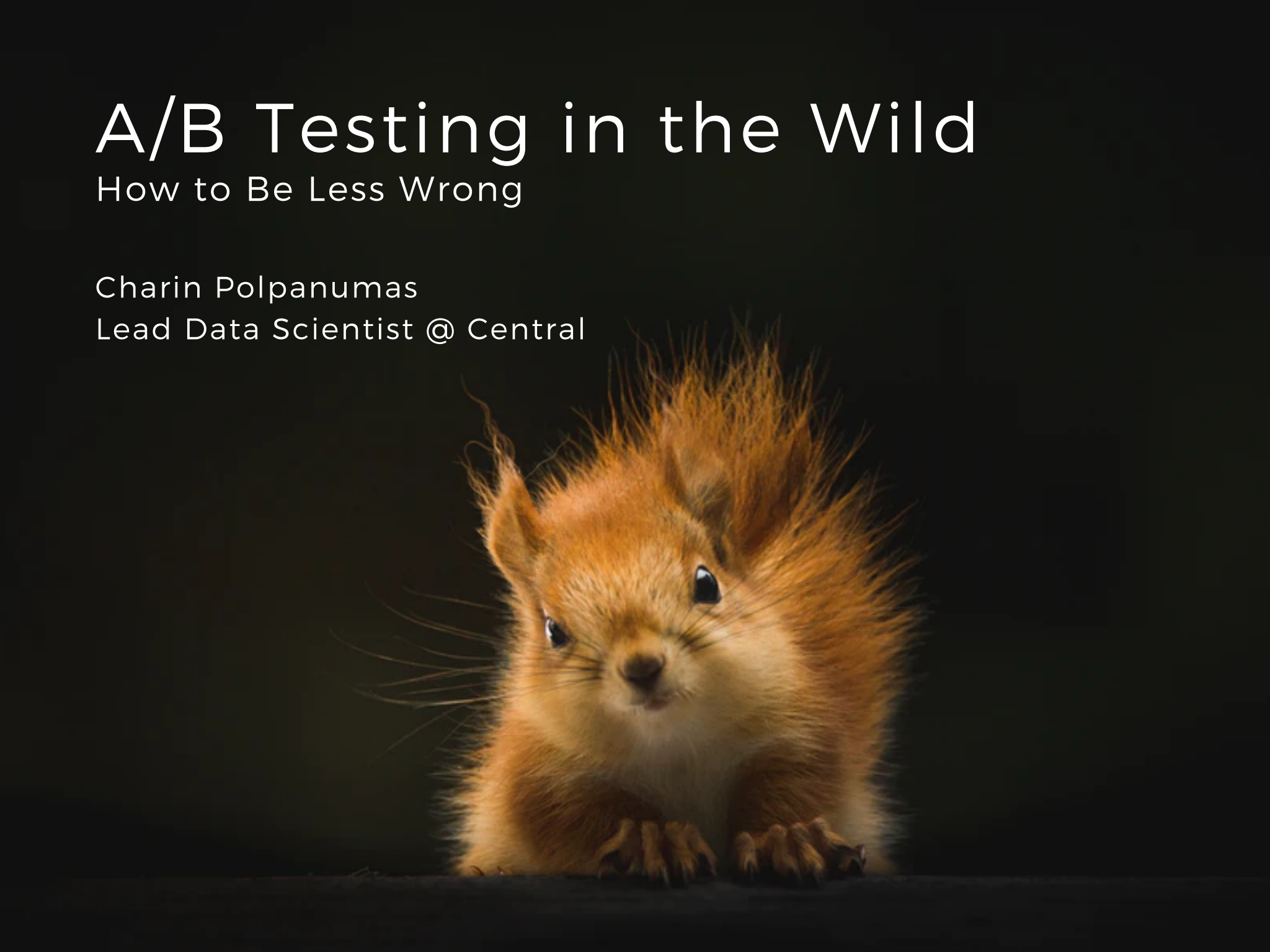
www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct

A/B Testing in the Wild

How to Be Less Wrong

Charin Polpanumas
Lead Data Scientist @ Central



Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct

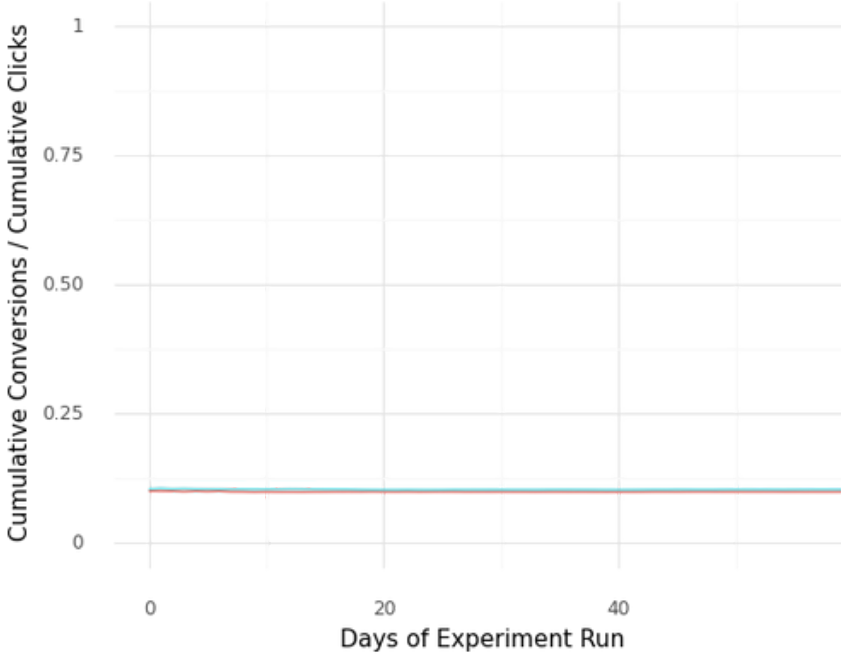
Choose statistical tests and validate their assumptions

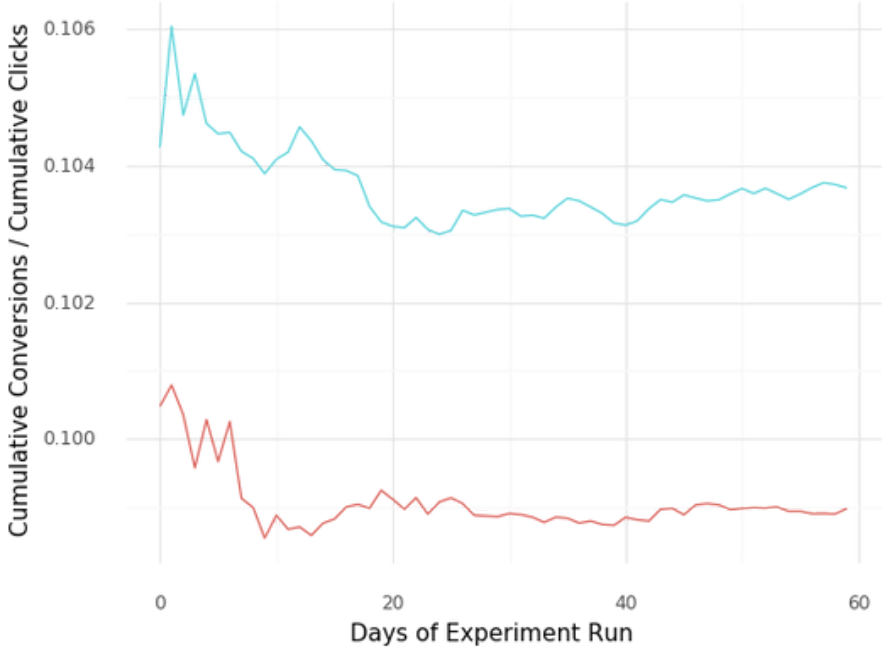
See frequentist.ipynb

1. Law of Large Numbers
2. Central Limit Theorem
3. Hypothesis tests
4. What frequentist hypothesis tests are saying
5. What frequentist hypothesis tests are NOT saying
6. Confidence intervals

God Mode

Imagine we already know the answer

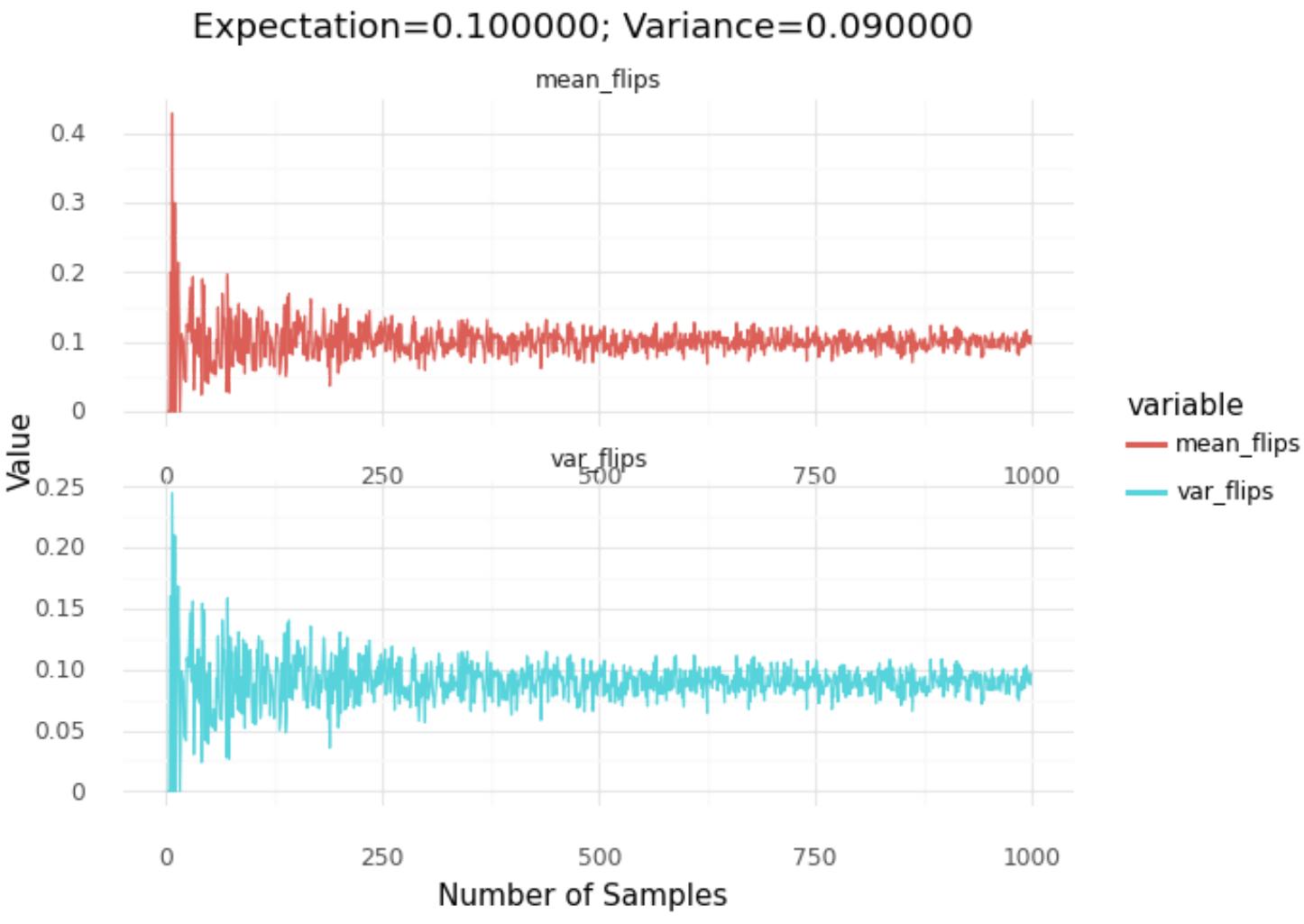




- Assume that our assumption of conversion process as Bernoulli trials is correct
- Red packets have population mean conversion rate of 10%
- Grey packets have population mean conversion rate of 10.5%
- How do we decide if this is enough to say one is better than the other?

Law of Large Numbers (LLN)

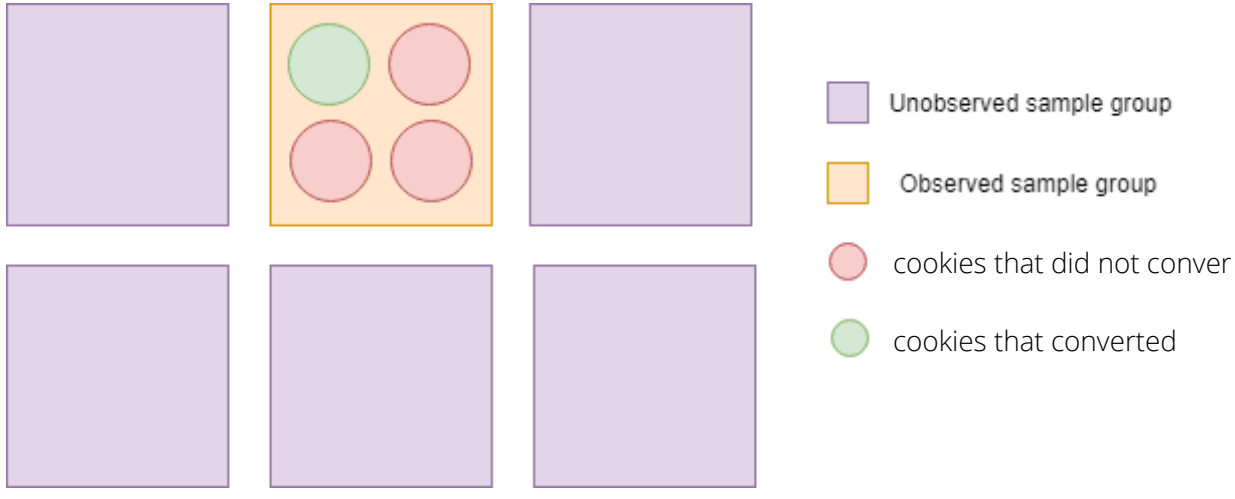
Sample mean will converge to expectation as sample size grows



Central Limit Theorem (CLT)

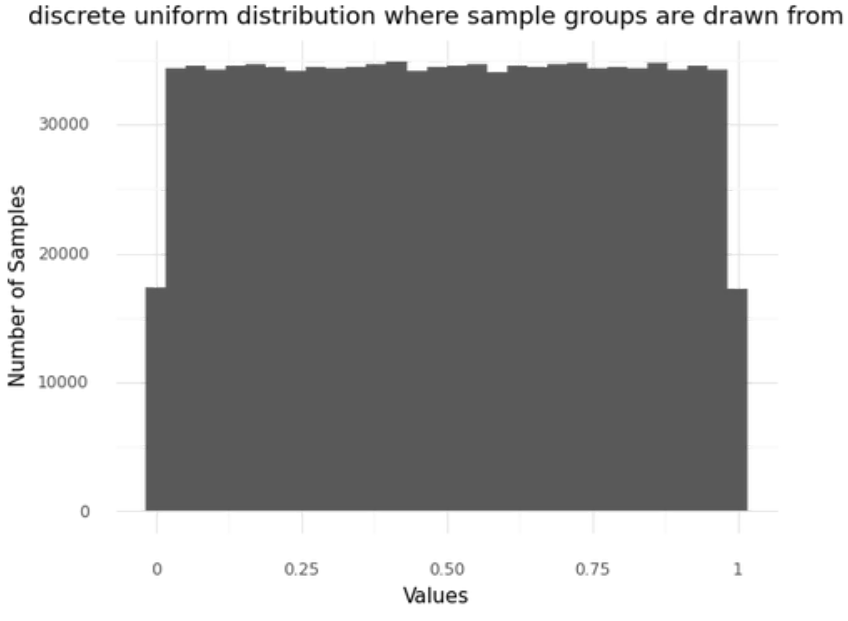
The foundation of all frequentist hypothesis tests

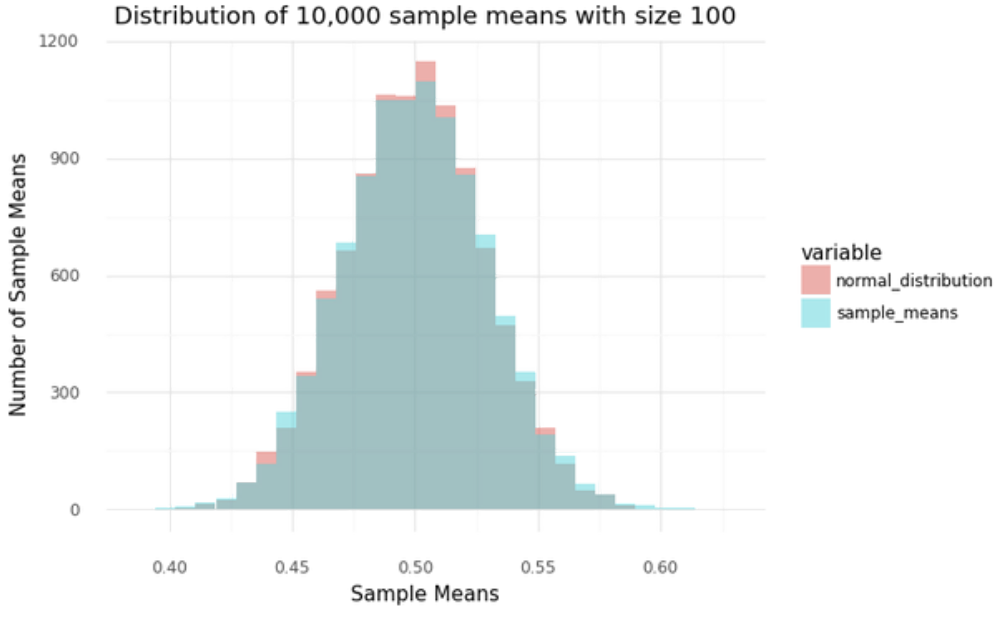
- X_i is independent and identically distributed of ANY distribution with mean μ and variance σ^2
- A sample group of X_i contains n samples of X_i
- The means of such sample groups will follow a normal distribution with mean μ and variance σ^2/n



Central Limit Theorem (CLT)

The foundation of all frequentist hypothesis tests





Hypothesis Tests

Frame the difference between groups as a mean of a sample group

1. Assume conversion rates follow normalized binomial distribution; we know it to be true in our god mode example
2. With CLT, we can treat the **conversion rates** and subsequently **difference in conversion rates** as following a normal distribution with expectation μ and variance σ^2/n
3. With LLN, we uses sample mean and sample variance in place of expectation and variance

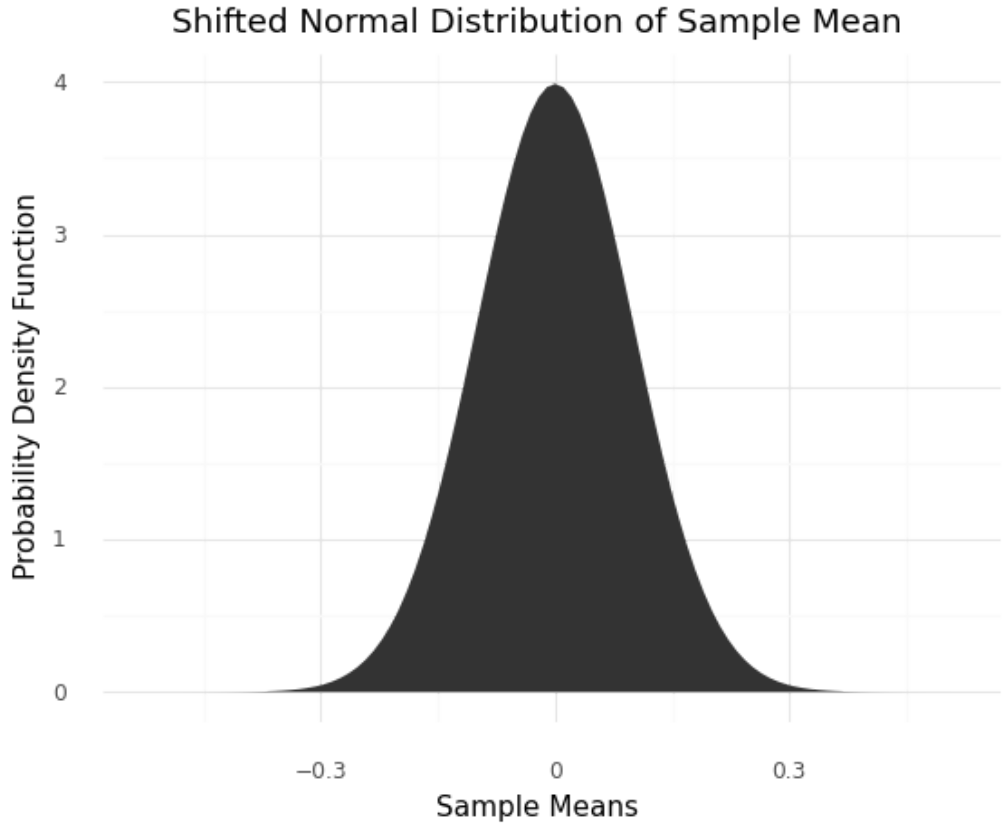
	campaign_id	cookies	conv_cnt	conv_per
0	A	59504	5890	0.098985
1	B	58944	6111	0.103675

Hypothesis Tests

One-sample Z-Test - transform distribution of sample means to N(0,1)

H0: conversion rate = 0.1

H1: not H0



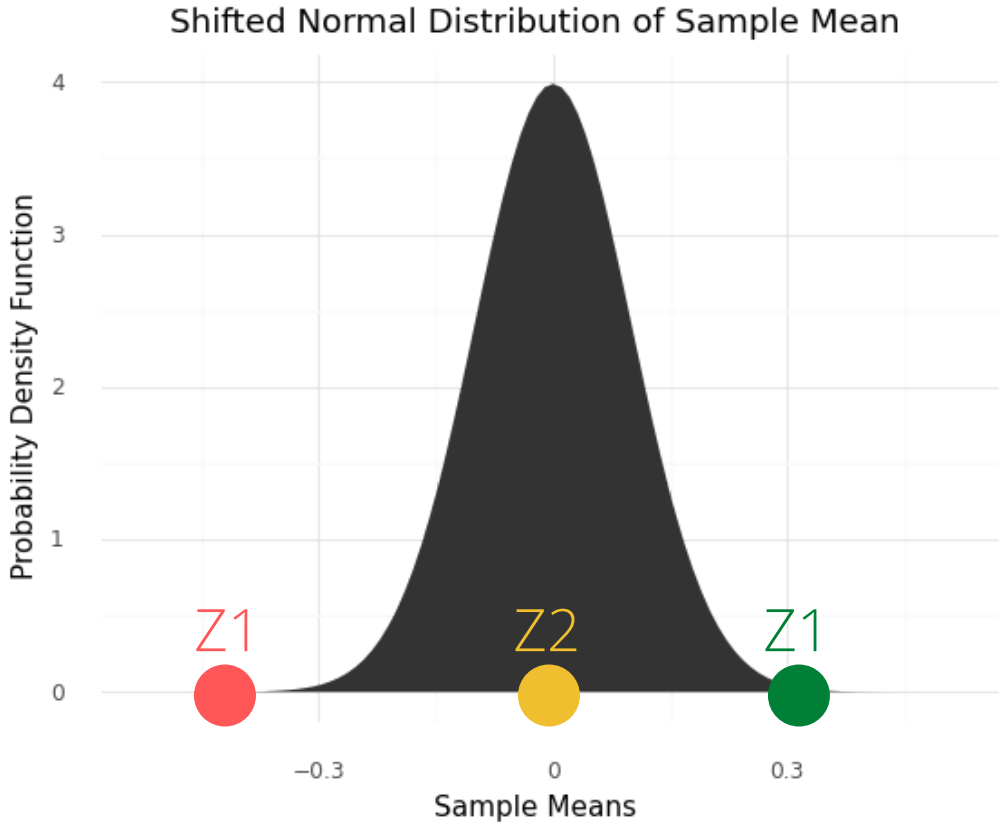
$$
\begin{aligned}
E[\bar{X}_j - \mu] &= E[\bar{X}_j] - \mu \\
&= \mu - \mu \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
Var(\frac{\bar{X}_j}{\sqrt{\sigma^2/n}}) &= \frac{1}{\sigma^2/n} Var(\bar{X}_j) \\
&= \frac{\sigma^2/n}{\sigma^2/n} \\
&= 1
\end{aligned}
$$

$$
\bar{Z}_j = \frac{\bar{X}_j - \mu}{\sigma/\sqrt{n}}
$$

Hypothesis Tests
Two-sample Z-Test

H0: no difference in conversion rates
H1: not H0



$$\bar{Z_\Delta} = \frac{\bar{X_\Delta} - \mu}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

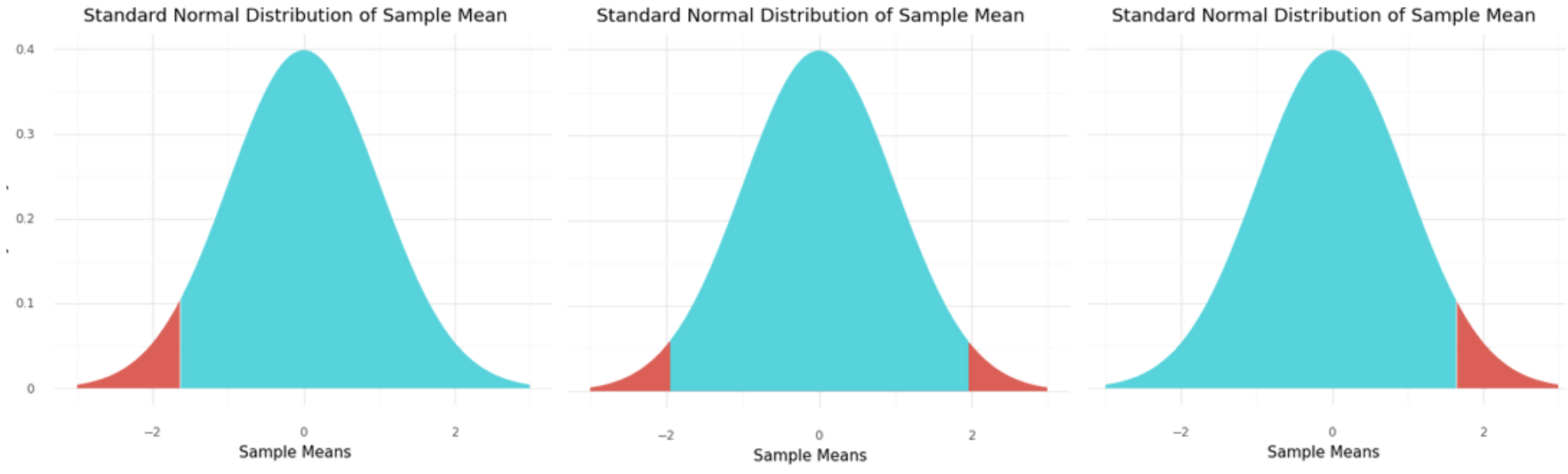
$$= \frac{\bar{X_\Delta} - \mu}{\sqrt{\sigma_{pooled}^2 * (\frac{1}{n_A} + \frac{1}{n_B})}}$$

- Pooled probability is total conversions / total cookies since we assumed in H0 there is no difference
- Variance of sum/difference is sum of variance
- We think of our experiment results as one point in the distribution N(0,1)

Hypothesis Tests

Decide what are statistically significant differences

Assumed that **what we think is true** (no difference in conversion),
what is **the chance** of us seeing sample mean values that are more extreme than **what we are seeing right now (our Z score)**?

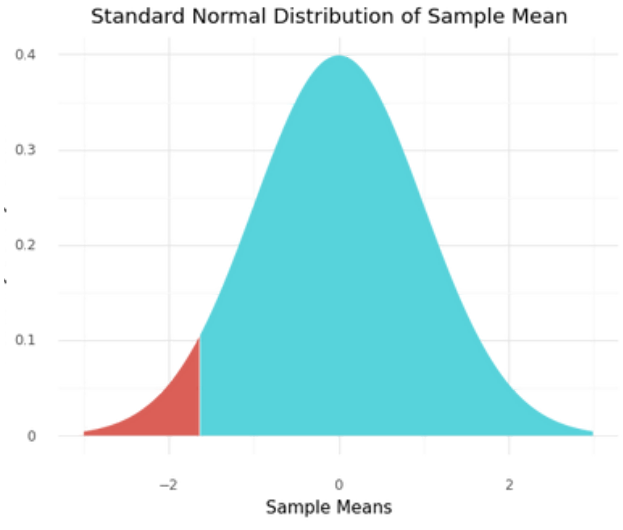


H0: difference is positive H0: difference is zero H0: difference is negative

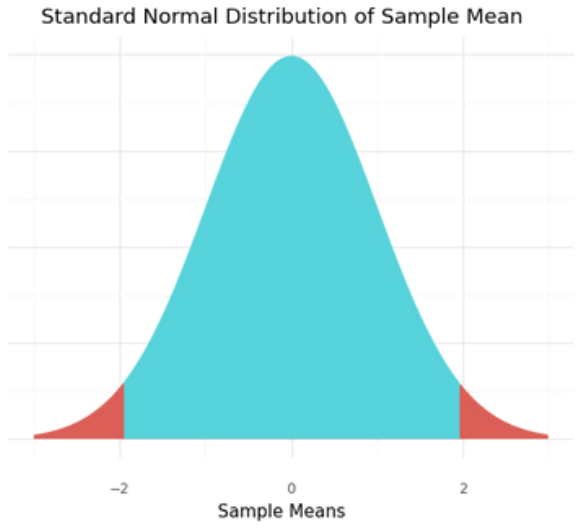
This chance is called p-value; they represent false positive rate of the test.
Frequentist tests require us to assume the level of false positive rate we can tolerate usually 1-10% called **statistical significance level (α)**.

What Frequentist Hypothesis Tests Are Saying

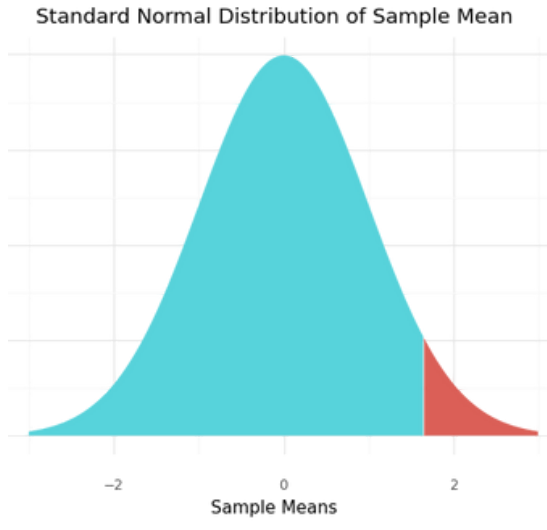
Assumed that **what we think is true** (H0: no difference in conversion), **the chance (p-value)** of us seeing sample mean values that are more extreme than **what we are seeing right now (our Z score)** is rarer than than the threshold of false positive rate we can accept (**statistical significance**; 1%, 5%, 10%)



H0: difference is positive



H0: difference is zero



H0: difference is negative

*Note that more vs less comparisons have Z-score thresholds than are less extreme since they only take care of one tail of the distribution

What Frequentist Hypothesis Tests Are NOT Saying

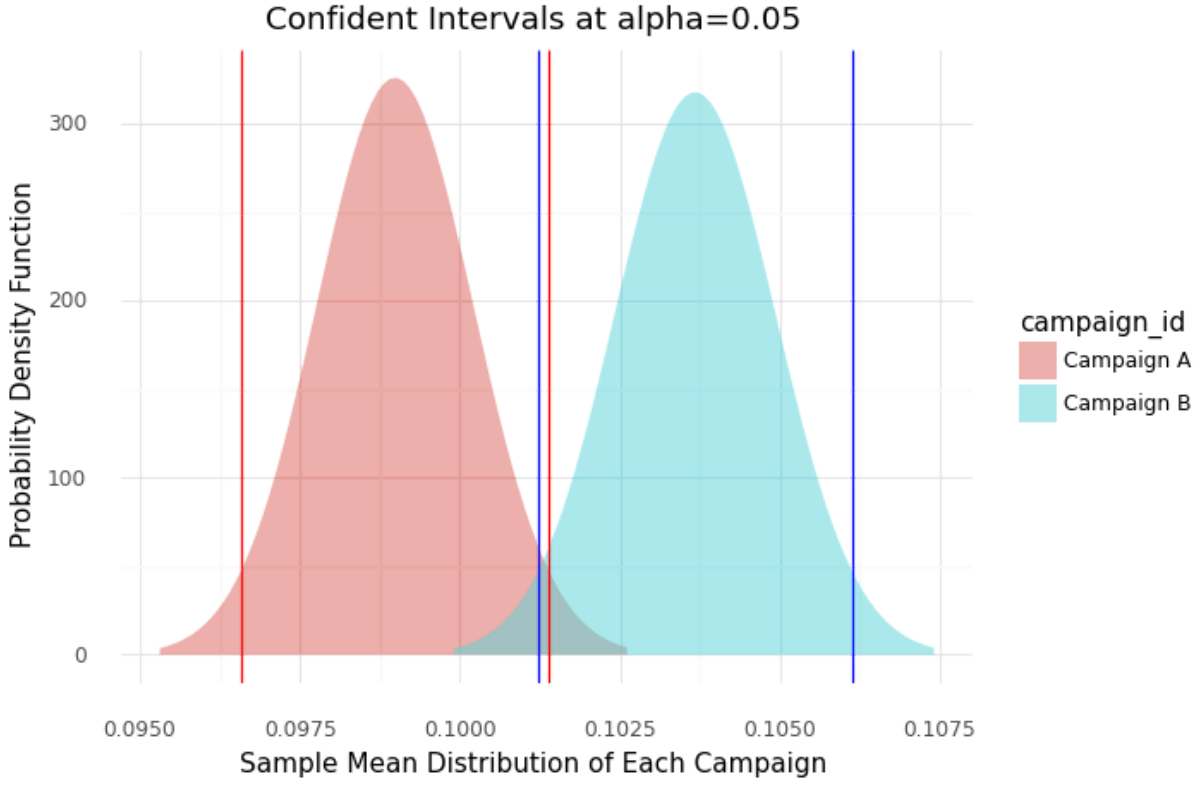
Misuse of p-values

- The p-value is not the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.
- The p-value is not the probability that the observed effects were produced by random chance alone.
- The 0.05 significance level is merely a convention.
- The p-value does not indicate the size or importance of the observed effect.
- The p-value is not the false positive rate of rejecting the null hypothesis; that depends on the prevalence of the data

Confidence Intervals

In the long run, X% of the samples will fall into this range

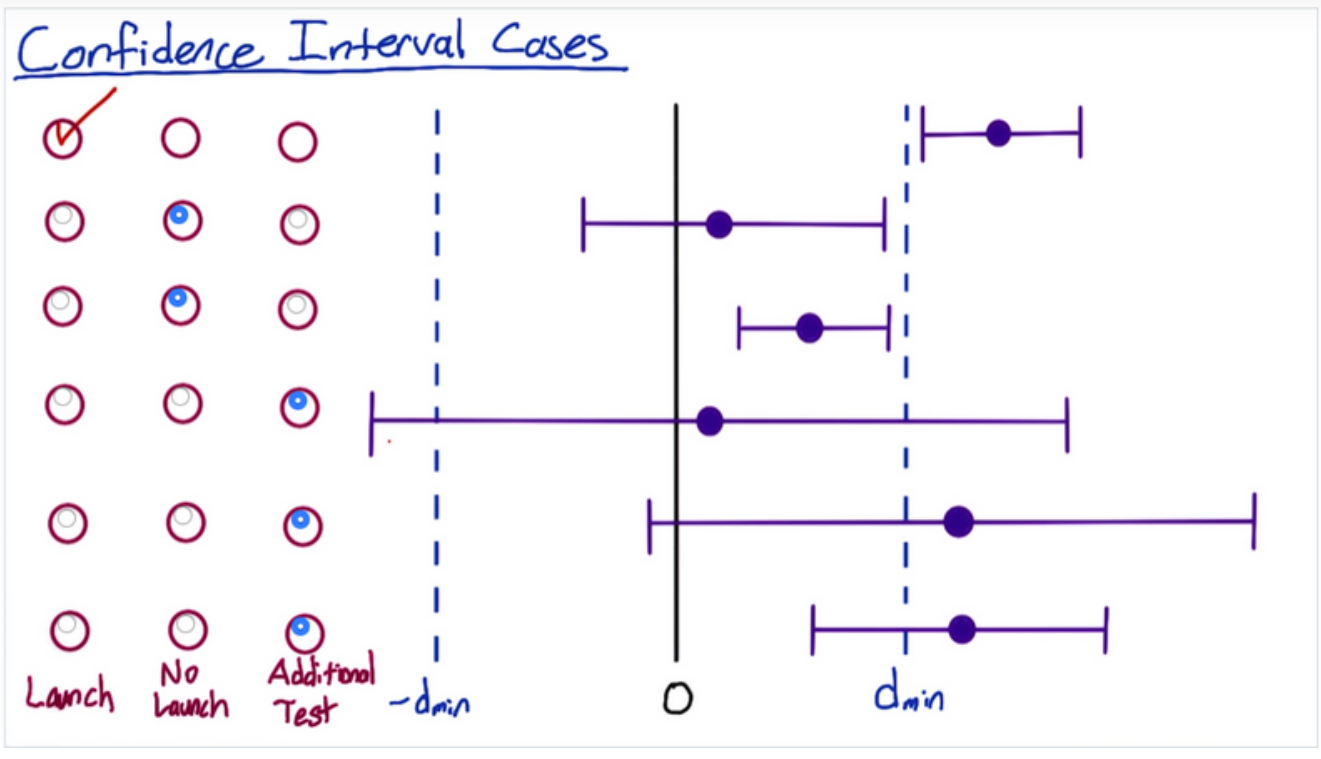
The intervals where there is an arbitrary probability, say 95%, that sample mean of A or B will fall into. We back calculate this from Z-score at 5% significance threshold (-1.96 and 1.96)



Confidence Intervals

Example from Google's Udacity course

You can also see the confidence interval of the difference in conversion rates, and decide if there is an effect or not, or you need to perform more tests. This is more lenient and practical than sticking to a significance threshold.



Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct
3. Implement one-tailed Z test at 5% significance level

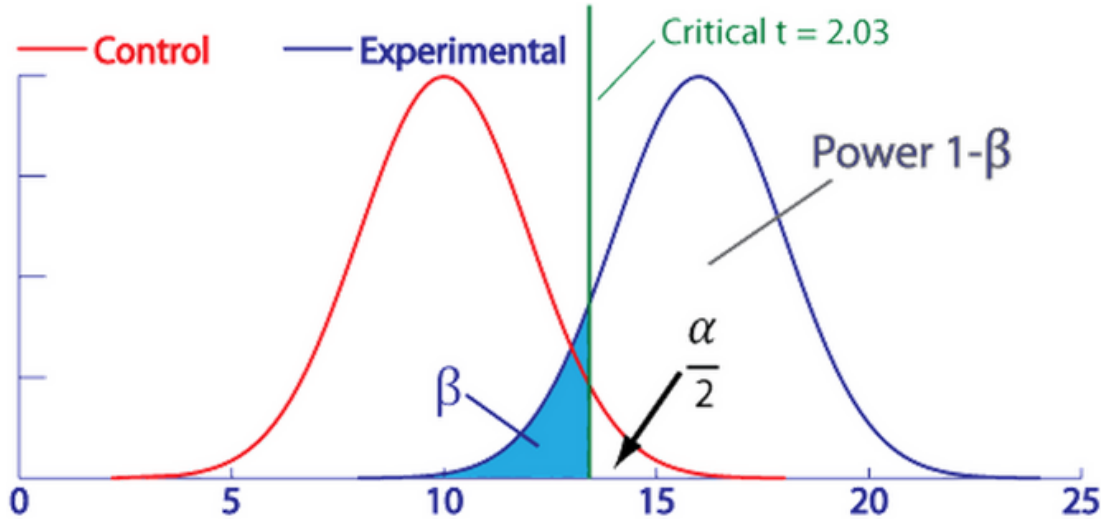
Decide on stopping criteria

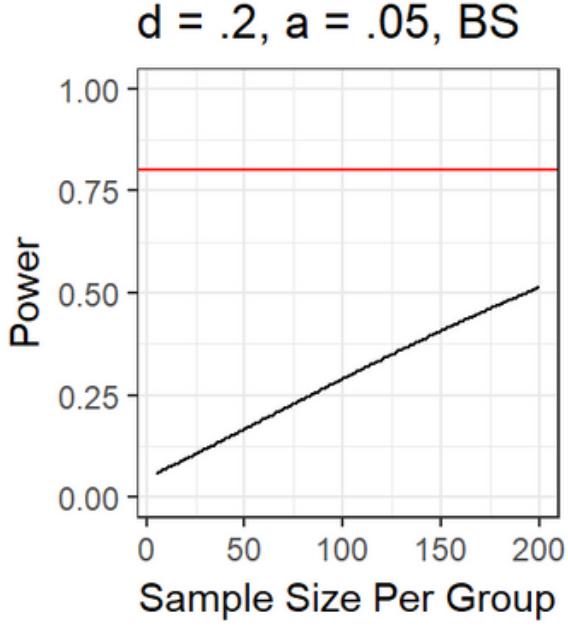
Many frequentist test are significant if you run them long enough

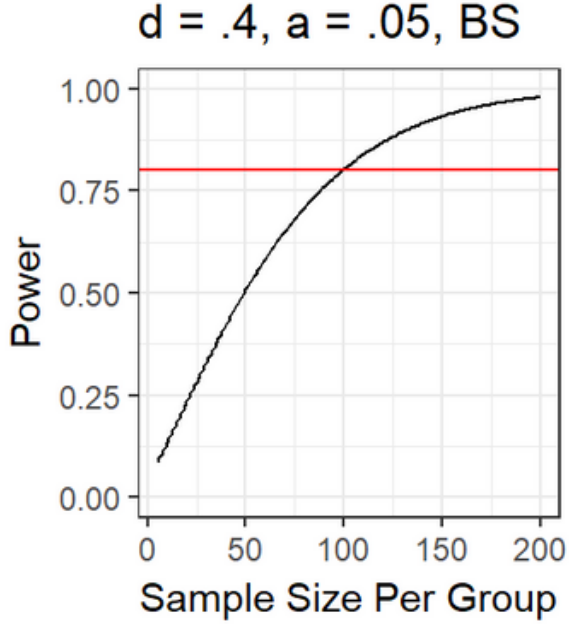
1. Minimum detectable effect
2. A/A tests
3. Trends and seasonality

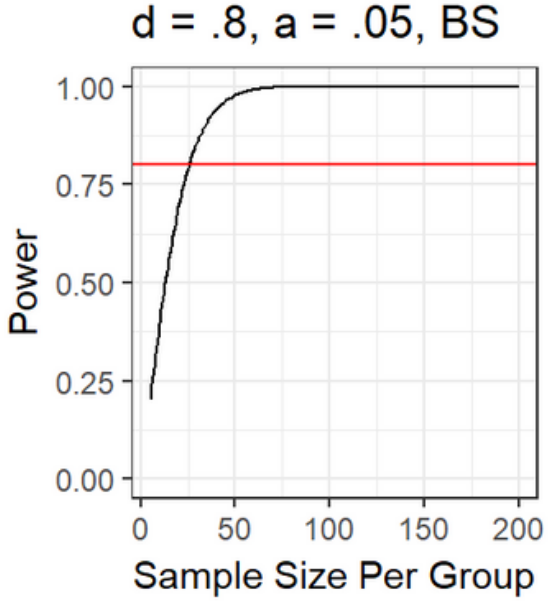
Minimum Detectable Effect

Effect size, power (true positive) and significance level (false positive)



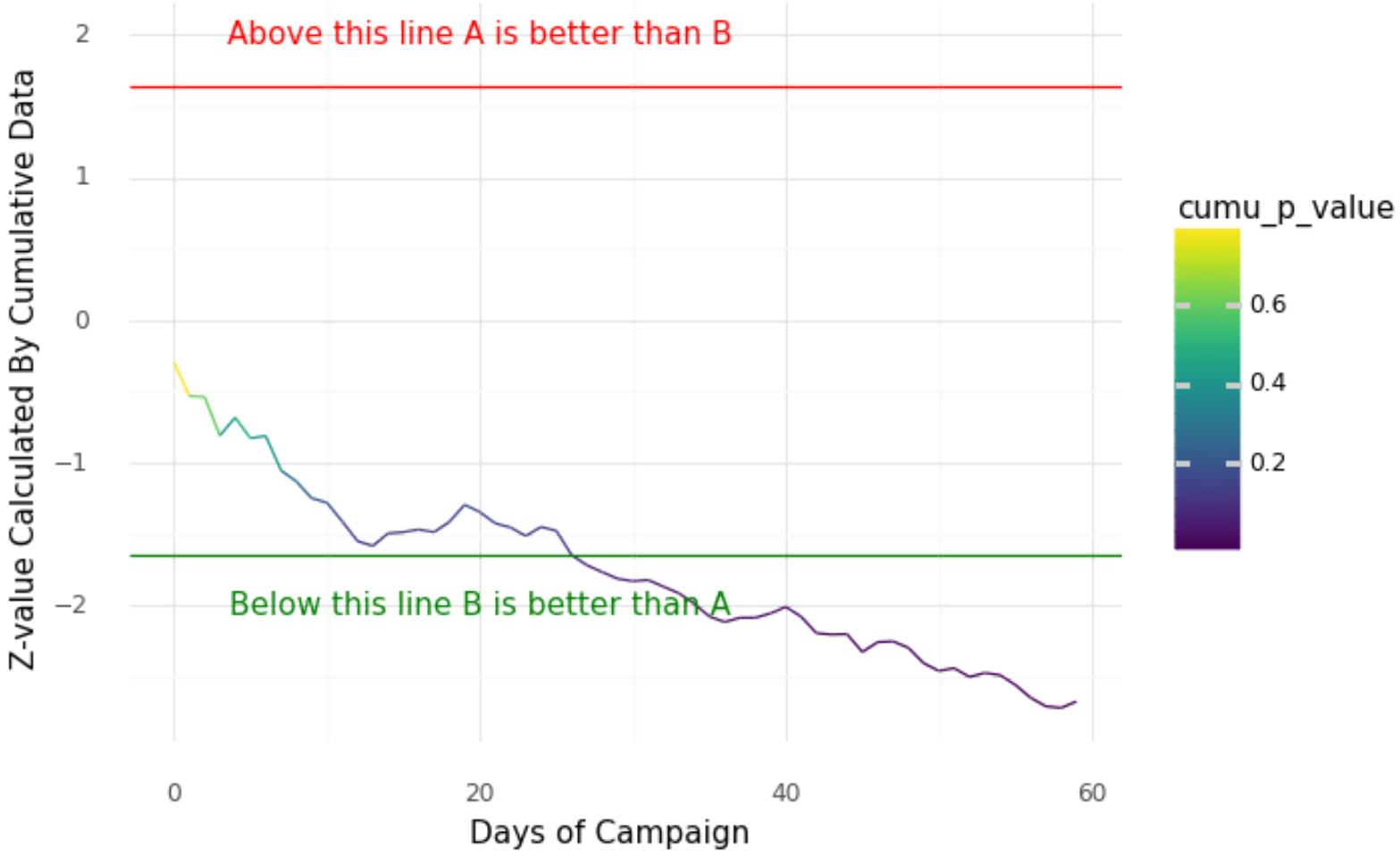






Minimum Detectable Effect

Many frequentist tests are statistically significant if you run them long enough



Minimum Detectable Effect

How we constructed our test statistics

observed difference

$$\bar{Z}_j = \frac{\bar{X}_j - \mu}{\sigma / \sqrt{n}}$$

usually 0

standard deviation of our data

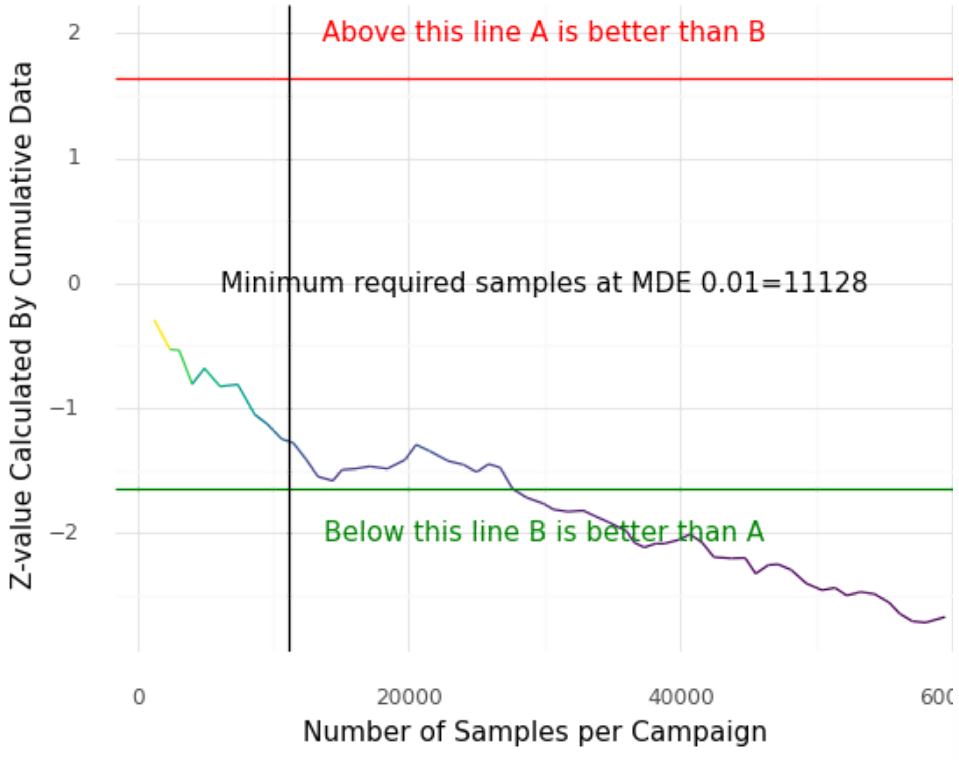
higher variation in data means differences (or lack thereof) will be more difficult to detect

number of samples

higher number of samples will make the differences (or lack thereof) easier to detect

Minimum Detectable Effect

Back calculate required sample size per variant



$$Z_{critical} = \mu_{H0} + Z_\alpha * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}$$

$$Z_{critical} = 0 + Z_\alpha * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}$$

$$Z_{critical} = \mu_{H1} - \mu_{H0} - Z_\beta * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}$$

$$Z_{critical} = \text{MDE} - Z_\beta * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}$$

$$0 + Z_\alpha * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})} = \text{MDE} - Z_\beta * \sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}$$

$$Z_\alpha + Z_\beta = \frac{\text{MDE}}{\sqrt{\sigma^2 * (\frac{1}{n} + \frac{1}{mn})}}$$

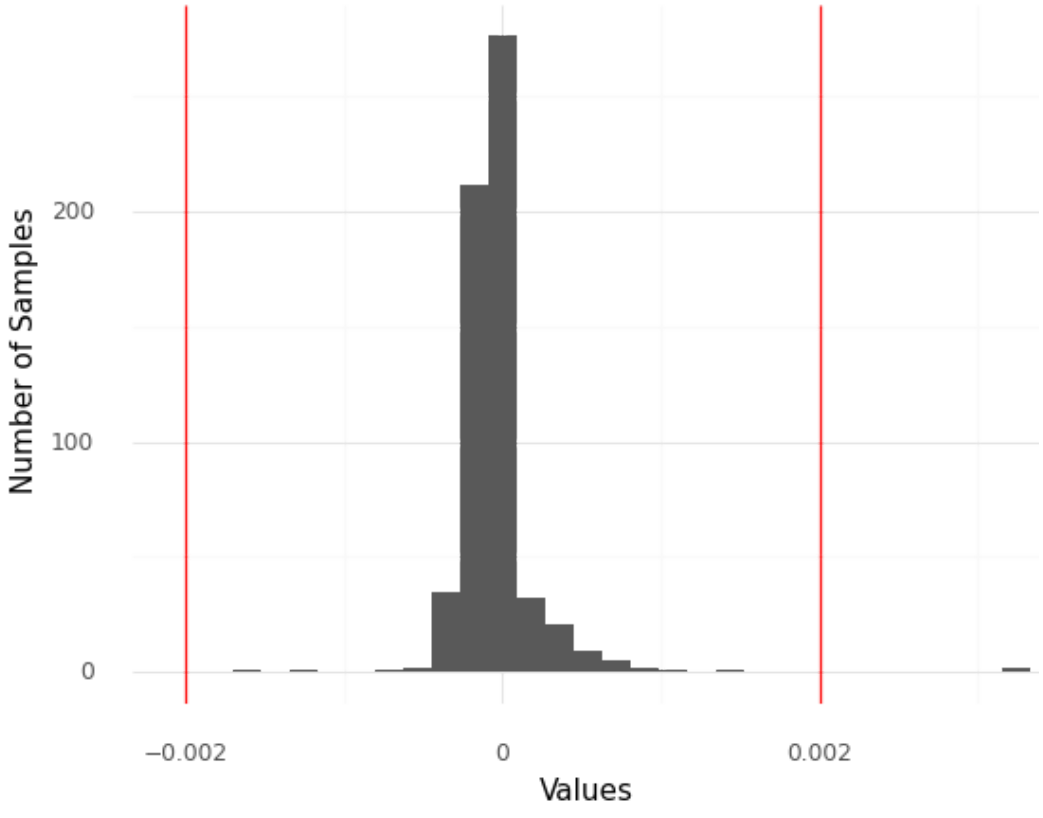
$$\frac{(m+1)\sigma^2}{mn} = (\frac{\text{MDE}}{Z_\alpha + Z_\beta})^2$$

$$n = \frac{m+1}{m}(\frac{(Z_\alpha + Z_\beta)\sigma}{\text{MDE}})^2$$

$$n = 2(\frac{(Z_\alpha + Z_\beta)\sigma}{\text{MDE}})^2; m = 1$$

A/A Tests

A/A tests to see the distribution of differences between groups

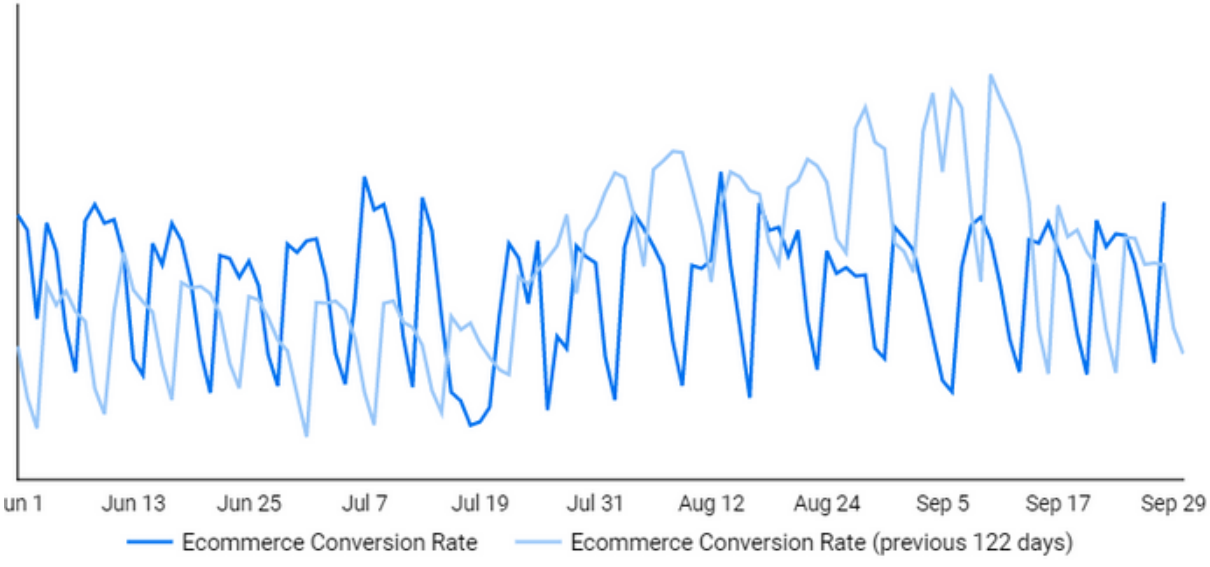


- Intentionally split cookies into control and control groups
- Show both of them the same variation
- Plot the distribution of differences between conversion rates
- A/A test results can also be used to estimate pooled sample variance empirically; Google's Udacity course discusses this extensively

Trends and Seasonality

Run until required samples are met or over all known seasonality

- Run full length of the seasonality such as weekend shoppers
- Normalize by year-on-year data to de-trend



Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct
3. Implement one-tailed Z test at 5% significance level
4. Assume MDE of 1% based on A/A tests and run until 5,000 cookies on each variation and at least one week

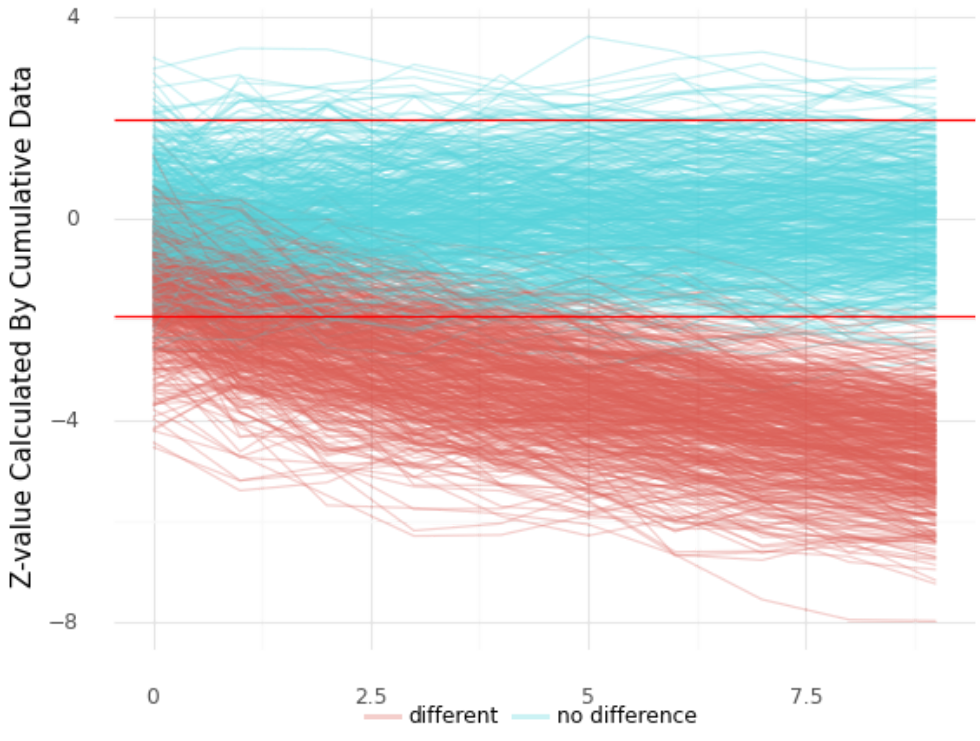
Run and monitor the experiment

Keep your tests in line

1. Do not peek
2. Sanity check
3. Post-test monitoring

Do Not Peek

Peeking results earlier than required samples will break your promise



Given the null hypothesis is true, frequentist tests promise to keep false positive rate lower than the significance threshold. Peeking earlier multiple times will break this promise.

Out of 1,000 simulations per group,
99.6% simulations with real difference reject H0 by peeking (99.6% at the end)
22.4% simulations with no difference reject H0 by peeking (**6.4%** at the end)

Sanity Check

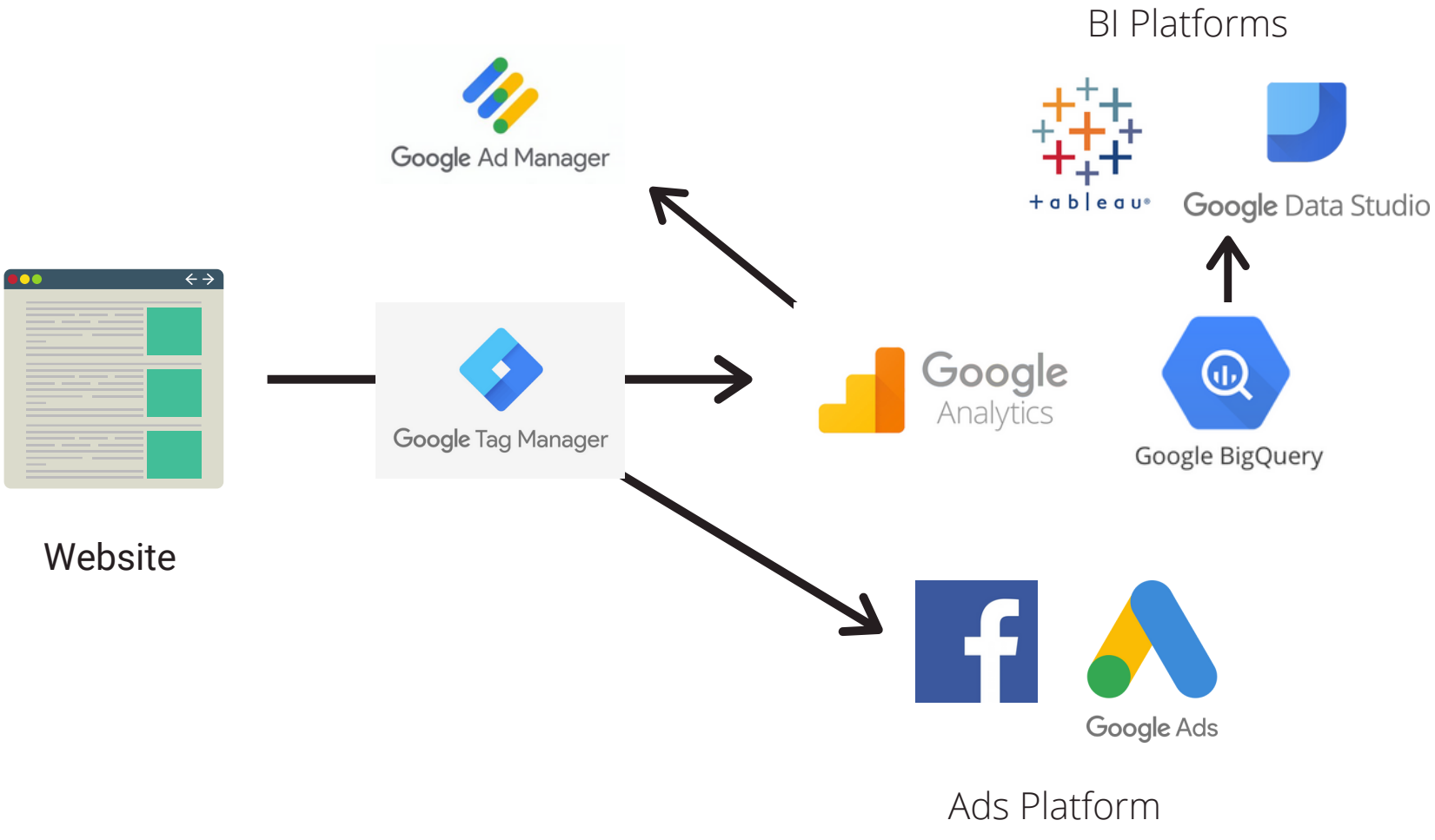
Is everything going as planned?

1. Check if your groups are split by what you intend; if we are testing by cookies and one cookies see both variations, that would be a disaster
2. Check if your control variables are controlled
3. Twyman's law - any statistic that appears interesting is almost certainly a mistake; try to see if the splitting mechanism works as intended
    a. Developers use random() to show variations instead of cookie-based splitting
    b. One variation does not load
    c. Tracking does not work

Sanity Check

Web tracking architecture

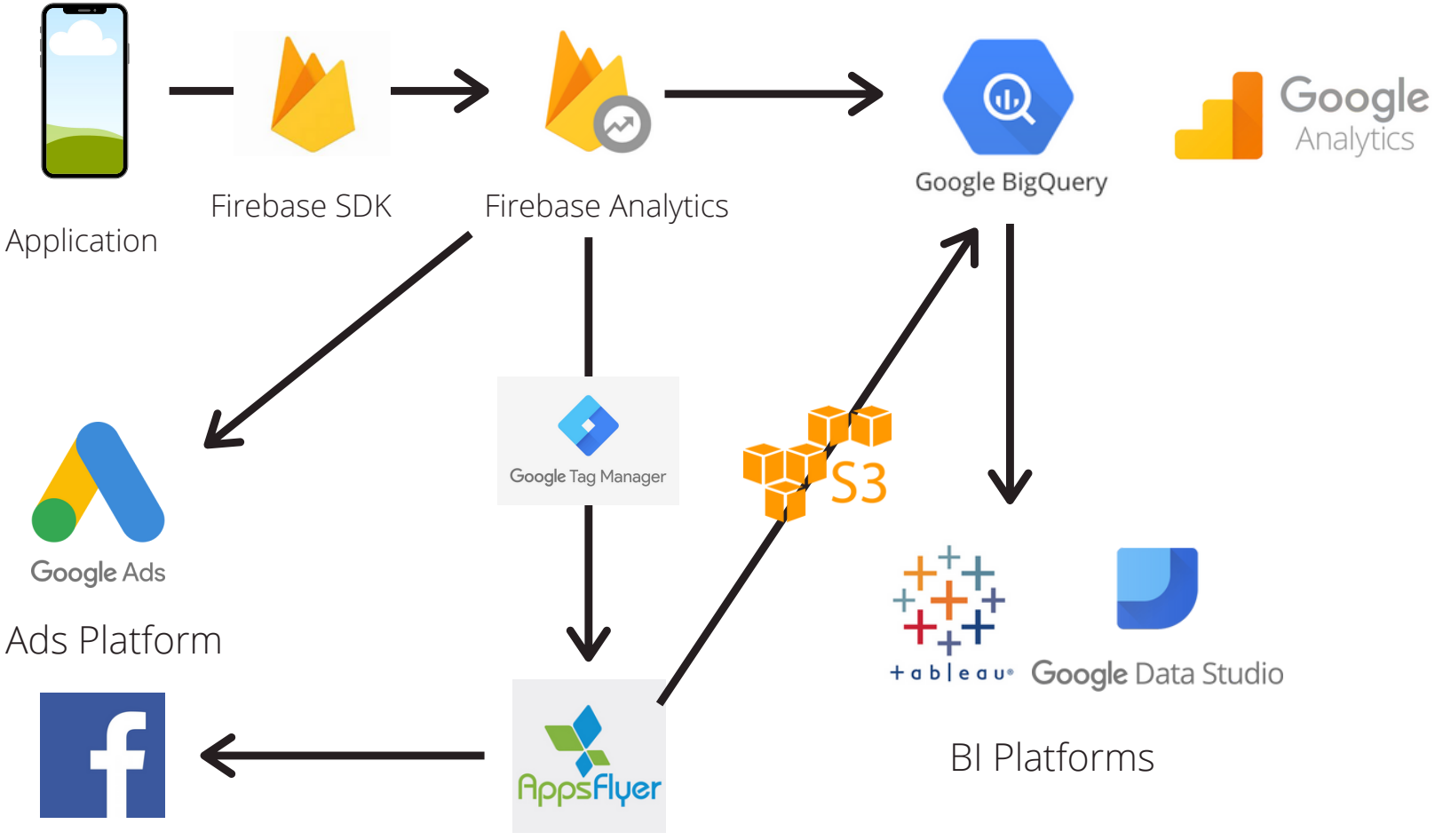
Google Tag Manager is a powerful tool that can aggregate web tracking scripts from different providers and send to analytic dashboards and ad platforms.



Sanity Check

App tracking architecture

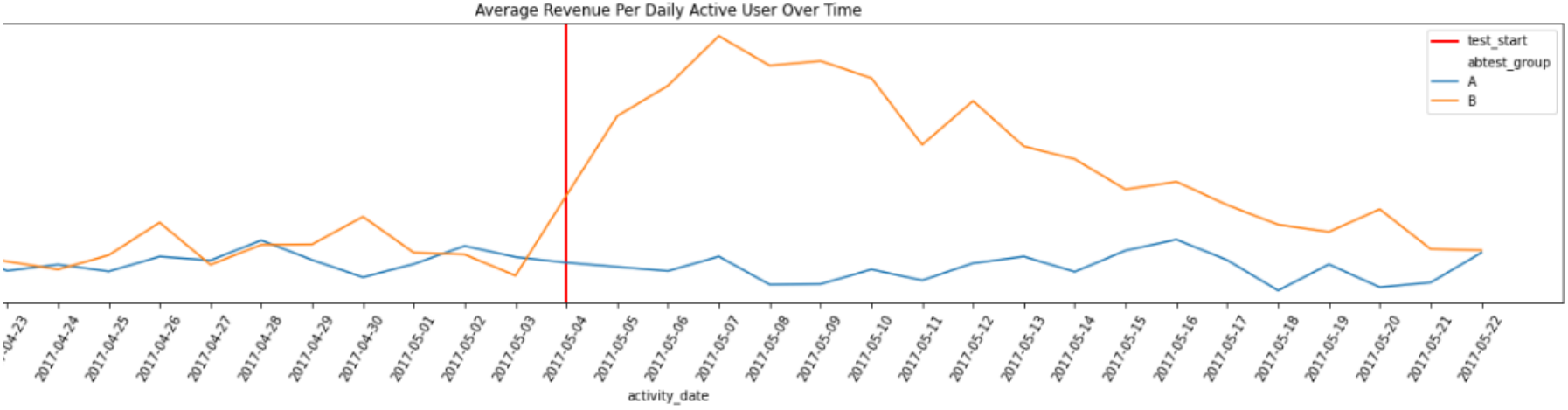
Most app tracking happens through Firebase SDK through which we can forward the data to GTM, analytic dashboards and ads platforms.



Post-test Monitoring

Monitor if the changes persist or just a one-hit wonder

- Customers might be interested because of the novelty but converges to control group over time
- Leave small portion of control group to verify



Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct
3. Implement one-tailed Z test at 5% significance level
4. Assume MDE of 1% based on A/A tests and run until X samples on each variation and at least one week
5. Check if cookies are viewing only assigned variations; leave 10% on control afterwards

Analyze results and suggest actions

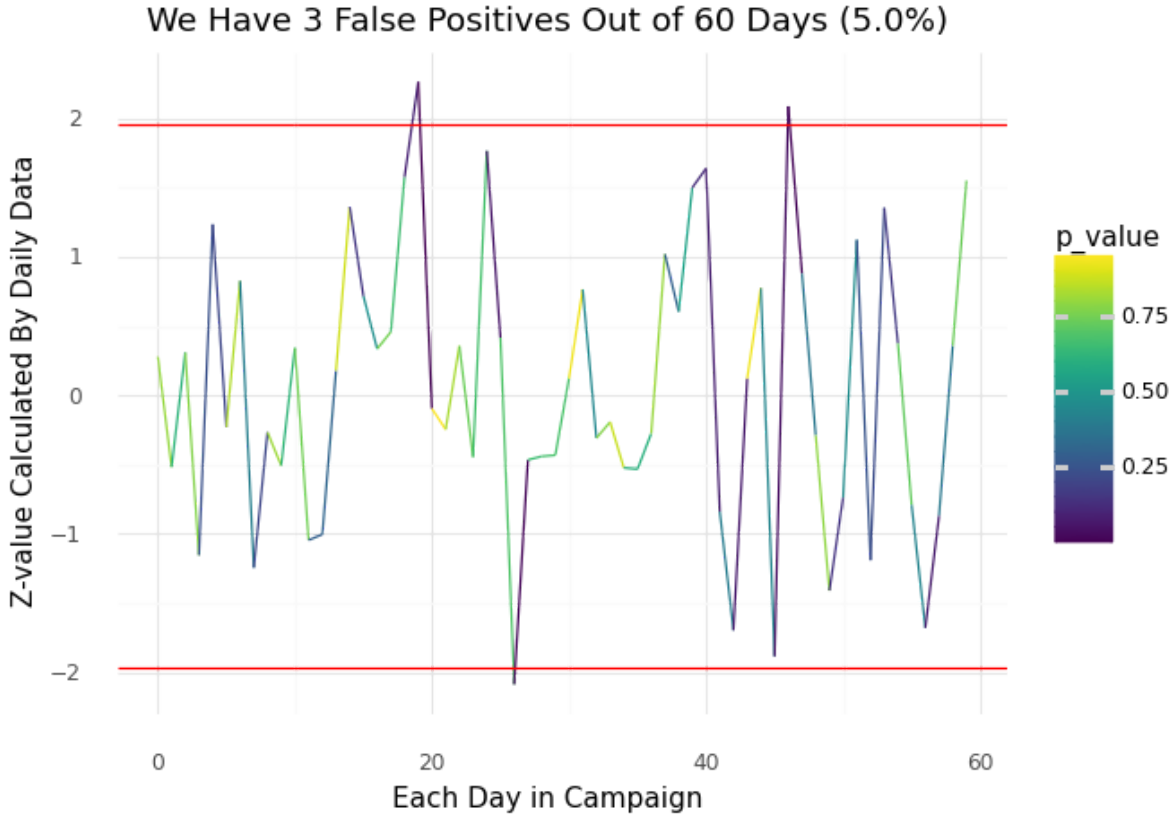
We are running a business after all

1. Multiple test corrections
2. Take actions; understand the risks

Multiple Test Corrections

If you run the tests enough times, you will get statistical significance

If you run 60 tests best on daily data, 3 of them will reject H0, even if it is both groups have the same true conversion rates



Multiple Test Corrections
Bonferroni and Sidak

- **Bonferroni corrections** - divide alpha by number of tests; very conservative
- **Sidak corrections** - assume the tests are independent, the chances that all tests are not statistically significant given null hypothesis is true (aka multiple-test false positive rate) is

  alpha_overall = 1 - (1-alpha_sidak)^(number of tests)
  alpha_sidak = 1 - (1-alpha_overall)^(1/number of tests)

Example: alpha_overall = 0.05; 10 tests

- Bonferroni corrections - 0.05/10 = 0.005
- Sidak corrections - 1 - (1-0.05)^10 = 0.005116

Note that there are also more complicated and less conservative corrections such as the Benjamin-Hochberg method

Take Actions; Understand the Risks

At the end of the day, we are running a business

A Dirty Dozen: Twelve P-Value Misconceptions

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

- Rejecting the null hypothesis does not mean that the alternative hypothesis is better *technically*, but we still make business decisions on it
- We tailor minimum detectable effect so that the winner is worth the investment; change button color vs new machine learning models

Sunflower Seeds Packet A/B Test

Design journal

www.hamzada.com/sunflower-seeds?variation=A

www.hamzada.com/sunflower-seeds?variation=B

1. Metric - # payment / # cookie-days; each cookie is a Bernoulli trial (or the metric follows normalized binomial distribution)
2. 50/50 split on cookies; run separate tests by cookies coming from direct
3. Implement one-tailed Z test at 5% significance level
4. Assume MDE of 1% based on A/A tests and run until X samples on each variation and at least one week
5. Check if cookies are viewing only assigned variations; leave 10% on control afterwards
6. Run one test at the end; use packets that win

Downsides of Frequentist Tests

And how Bayesian tests *might* help

1. It is extremely counterintuitive to explain; we prefer *P(Hypothesis|Data)* over *P(Data|Hypothesis)*
2. Ain't nobody got time for that; peeking problem
3. It does not tell you the effect size; you only know a statistically significant result has larger size than you minimum detectable effect

$$P(H|D) = \frac{P(H \cap D)}{P(D)} \quad (1)$$

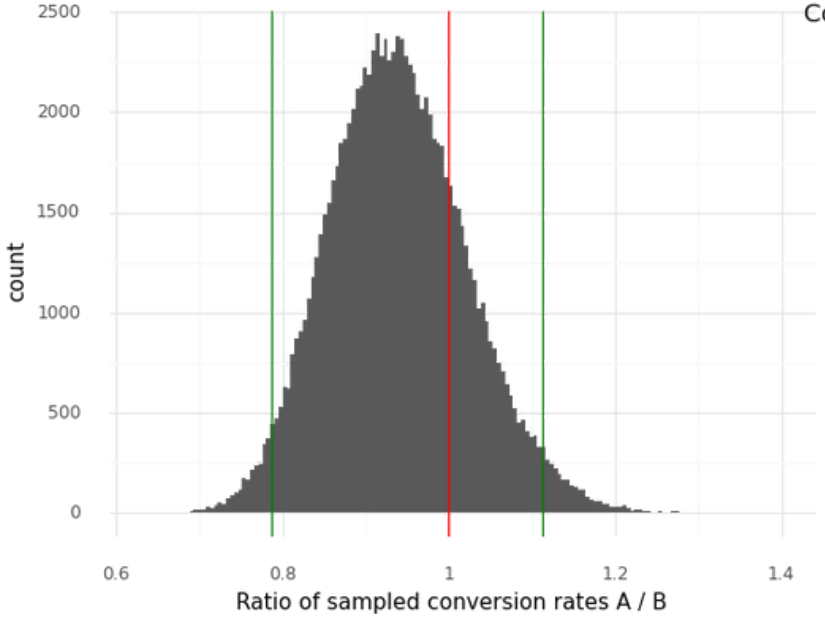
$$= \frac{P(D|H)P(H)}{P(D)}; \text{ chain rule of probability} \quad (2)$$

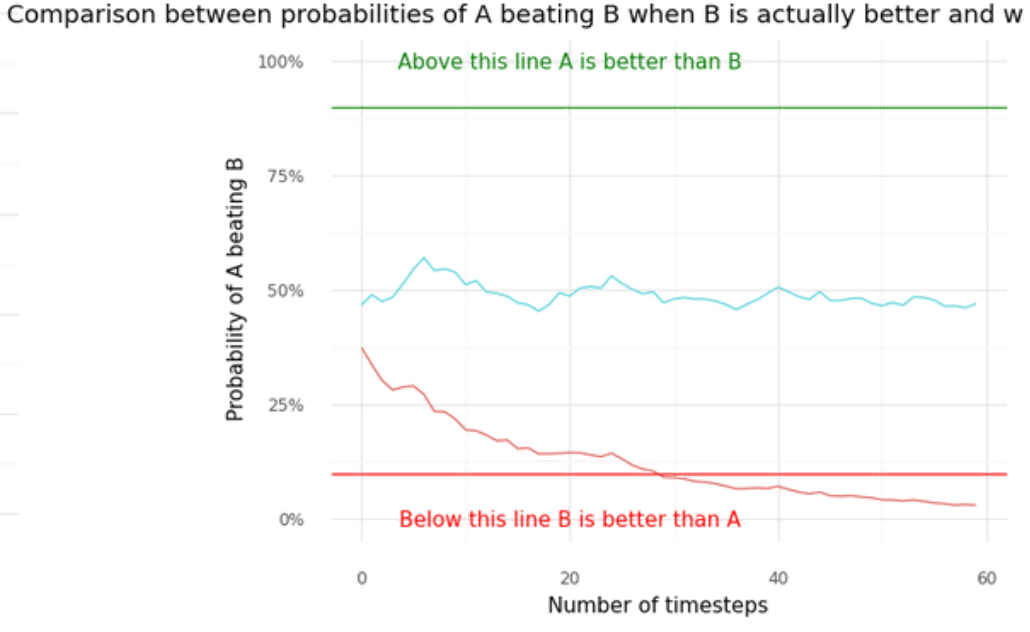
$$= \frac{P(D|H)P(H)}{\sum_{j=1}^{k} P(D|H_j)P(H_j)}; \text{ summing over all possible hypotheses} \quad (3)$$

What Bayesian A/B Test Results Look Like

Run simulation from posterior distribution

Probability that A is greater than B: 0.22743
Average A/B ratio: 0.9394908467261973
Credible interval of A/B ratio: [0.78622231 1.11331594]





Put probability on the hypothesis
Quantify effect size

Sort of resistant to peeking;
DOES NOT guarantee anything

Limitations

When to NOT do an A/B test

1. **Things that cannot be summarized into one or a few metrics**; are our users happy?
2. **Totally new things**; not apple-to-apple comparison; do a survey, focus group, design thinking
3. **Delayed results**; for example, does a change makes customer repurchase more? what if repurchase period is 3 months?
4. **One-off events**; no two big sales events such as 11.11 or Black Friday are the same
5. **If you cannot split groups independently**; offline store layout of the same store

A/B Testing in the Wild

1. Understand your business; you do A/B tests so that you can make real-world decisions, not because you are a calculator
2. Tracking, tracking, tracking
3. Going through the process once is more important than trying to make every step perfect
4. Explicitly state your assumptions and be prepared for the risks
5. Control groups will win a lot; it's how you respond to it

Useful Resources on A/B Testing

- Calculators
  - Evan's Awesome A/B Test Tools
  - A/B Test Calculators by abtestguide.com
  - abtestoo - A/B Tests and MABs made easy
- Courses and readings
  - ExP Experimentation Platform
  - Udacity's A/B Testing by Google
  - p-value and the base rate fallacy
- Frameworks
  - wasabi by Intuit
  - planout by Facebook
  - Google Optimize by Google