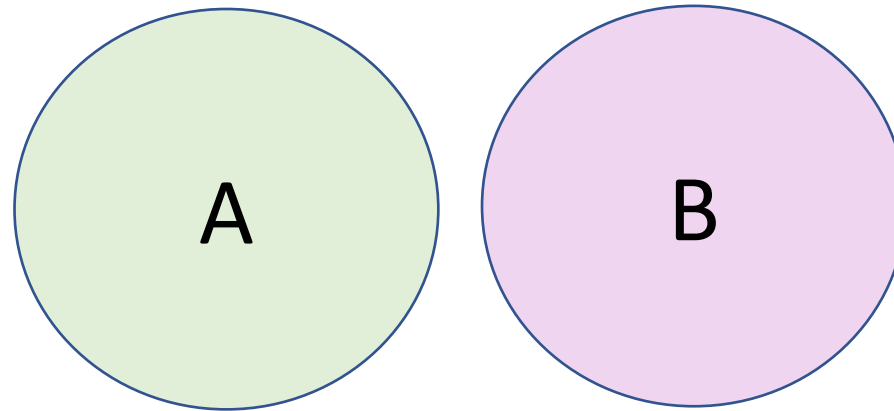


An Introduction to A/B Testing - Part I




- What is A/B Testing?
- What do I need to run an A/B test?
- Sample Sizes and Power
- How is statistical significance determined?
- Exercise: Will it A/B Test?
- Our case study




You're the user experience manager at Bing and want to determine the right color scheme to increase user engagement.

Your analysts claim that changing to a new color scheme will improve user engagement.

Current Version

 Microsoft Bing

What is bologna made of




ALLIMAGESVIDEOSMAPSNEWSSHOPPING

Also try: [how it's made bologna](#) · [what is really in bologna](#) · [homemade bologna recipe in casing](#)



6,080,000 ResultsAny time ▾

What is Bologna made of?

Bologna is a cooked, smoked sausage made of cured beef, cured pork or a mixture of the two. The bologna might include choice cuts, depending on who's making it, but usually contains afterthoughts of the meat industry - organs, trimmings, end pieces and so on.

Image: huffingtonpost.com

[Glad You Asked: What is bologna made of, and how did it get its na...](#)
[journaltimes.com/news/local/glad-you-asked-what-is-bologna-made-of-and-how...](#)

Was this helpful?  

PEOPLE ALSO ASK

What part of the cow is Bologna made from?

What is the difference between Bologna and baloney?




What animal is Bologna?

How do you make homemade Bologna?


Feedback

Images of What is Bologna made of




[bing.com/images](#)



Current Version

 Microsoft Bing

What is bologna made of




ALLIMAGESVIDEOSMAPSNEWSSHOPPING


Also try: [how it's made bologna](#) · [what is really in bologna](#) · [homemade bologna recipe in casing](#)



6,080,000 ResultsAny time ▾

What is Bologna made of?

Bologna is a cooked, smoked sausage made of cured beef, cured pork or a mixture of the two. The bologna might include choice cuts, depending on who's making it, but usually contains afterthoughts of the meat industry - organs, trimmings, end pieces and so on.

Image: huffingtonpost.com

Glad You Asked: What is bologna made of, and how did it get its na...
 [journaltimes.com/news/local/glad-you-asked-what-is-bologna-made-of-and-how...](#)

Was this helpful?  

PEOPLE ALSO ASK

What part of the cow is Bologna made from? ▾




What is the difference between Bologna and baloney? ▾

What animal is Bologna? ▾


How do you make homemade Bologna? ▾

Feedback




Images of What is Bologna made of
[bing.com/images](#)



Proposed New Version

 Microsoft Bing

What is bologna made of




ALLIMAGESVIDEOSMAPSNEWSSHOPPING


Also try: [how it's made bologna](#) · [what is really in bologna](#) · [homemade bologna recipe in casing](#)



6,080,000 ResultsAny time ▾

What is Bologna made of?

Bologna is a cooked, smoked sausage made of cured beef, cured pork or a mixture of the two. The bologna might include choice cuts, depending on who's making it, but usually contains afterthoughts of the meat industry - organs, trimmings, end pieces and so on.

Image: huffingtonpost.com

Glad You Asked: What is bologna made of, and how did it get its na...
 [journaltimes.com/news/local/glad-you-asked-what-is-bologna-made-of-and-how...](#)

Was this helpful?  

PEOPLE ALSO ASK

What part of the cow is Bologna made from? ▾




What is the difference between Bologna and baloney? ▾

What animal is Bologna? ▾

How do you make homemade Bologna? ▾

Feedback

Images of What is Bologna made of
[bing.com/images](#)



Current Version

Microsoft Bing

What is bologna made of

ALL IMAGES VIDEOS MAPS NEWS SHOPPING

Also try: [how it's made bologna](#) · [what is really in bologna](#) · [homemade bologna recipe in casing](#)

6,080,000 Results Any time ▾

What is Bologna made of?

Bologna is a cooked, smoked sausage made of cured beef, cured pork or a mixture of the two. The bologna might include choice cuts, depending on who's making it, but usually contains afterthoughts of the meat industry - organs, trimmings, end pieces and so on.




Image: huffingtonpost.com

Glad You Asked: What is bologna made of, and how did it get its na...
[journaltimes.com/news/local/glad-you-asked-what-is-bologna-made-of-and-how...](#)

Was this helpful? 👍 👎


PEOPLE ALSO ASK

- What part of the cow is Bologna made from? ▾
- What is the difference between Bologna and baloney? ▾
- What animal is Bologna? ▾
- How do you make homemade Bologna? ▾

Feedback

Images of What is Bologna made of

[bing.com/images](#)



Proposed New Version

Microsoft Bing

What is bologna made of

ALL IMAGES VIDEOS MAPS NEWS SHOPPING

Also try: [how it's made bologna](#) · [what is really in bologna](#) · [homemade bologna recipe in casing](#)

6,080,000 Results Any time ▾

What is Bologna made of?

Bologna is a cooked, smoked sausage made of cured beef, cured pork or a mixture of the two. The bologna might include choice cuts, depending on who's making it, but usually contains afterthoughts of the meat industry - organs, trimmings, end pieces and so on.




Image: huffingtonpost.com

Glad You Asked: What is bologna made of, and how did it get its na...
[journaltimes.com/news/local/glad-you-asked-what-is-bologna-made-of-and-how...](#)

Was this helpful? 👍 👎


PEOPLE ALSO ASK

- What part of the cow is Bologna made from? ▾
- What is the difference between Bologna and baloney? ▾
- What animal is Bologna? ▾
- How do you make homemade Bologna? ▾

Feedback

Images of What is Bologna made of

[bing.com/images](#)



You think that there is no way such a small change will have any impact, but want to give your analysts the benefit of the doubt.

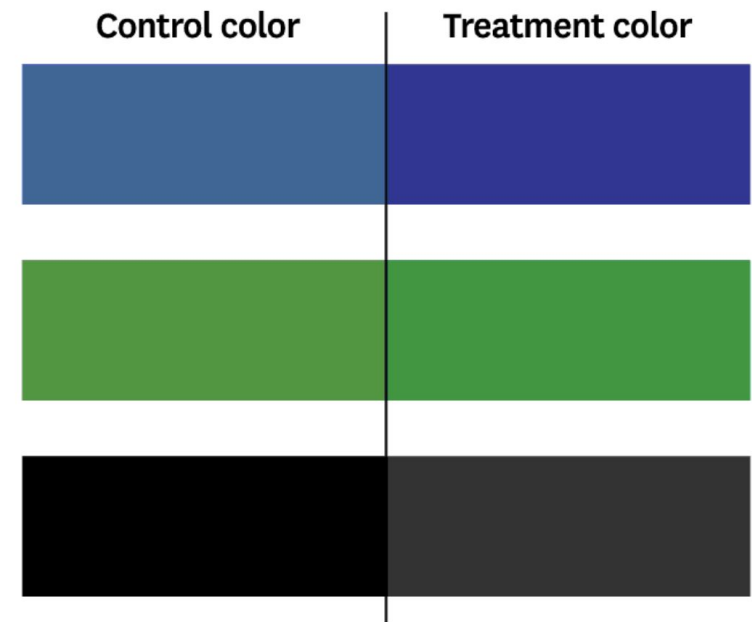
How do you test this claim?

Ask the UI/UX experts?

No - Let the data speak for itself!

In 2013, when Bing experimented with these color changes, they found that users presented with the treatment colors were successful in their searches a larger percentage of the time and that those who found what they wanted did so in significantly less time.

After validating the results this experiment with a much larger sample of 32 million users, the results indicated a projected revenue increase of \$10 million annually due to these simple color changes.



Turning your ideas into Experiments

As you can see, testing even small changes can be enormously lucrative. The best part: you can experiment in almost every discipline there is (you'll find out more about that in one of our case studies!). If you are interested in reading about more case studies, Optimizely recently published a [Big Book of Experimentation](#), showcasing 50 different case studies across different industries.

Let's talk about A/B testing ...

What is A/B Testing?

An A/B test is a **simple controlled experiment**, in which a single variable is altered between two groups (A and B) and a target variable to measure outcome is clearly defined. It is widely used in marketing and web design.

- The idea underlying A/B testing has been around for a long time
 - In the 1920s, statistician and biologist Ronald Fisher ran agricultural experiments to answer how different fertilizer affected the crops on this land
 - Clinical trials often follow A/B test pattern
- Modern usage has been prominent since around the year 2000 (marketing and web design)
 - First modern A/B tests run by companies like Google, which ran an early experiment to determine the optimum number of results to return from a search
 - Today companies like Google and Amazon may run thousands of tests at any given time

Examples of A/B Tests in Marketing

Testing can be done on virtually anything. You can test simple ideas like,

- Does changing my email subject line affect open rates?
- Does changing my website's Calls to Action (buttons, links, etc.) increase signups?
- Does adding an additional field into a form decrease form completions?
- Which Paid Search ad is more successful in driving users to the site?

... or more sophisticated ones:

- Will personalizing a whole website experience increase conversions?
- Will changing my website to dark mode during night time increase user engagement?
- How does radio advertising affect traffic to my website?

Your turn!

Locate at least two experiments currently running on Amazon.com.
Present one experiment and start to think why Amazon might be running this particular test.

How?

- A/B tests are randomly assigned based on your cookies. By opening two different incognito browsers and clearing your cookies, you will be able to spot the experiments currently running on Amazon's website.

Step-by-Step Process to perform A/B testing

A/B testing has five key components. Following these steps creates data-driven problem solving quickly and iteratively to achieve success for your business.



Step-by-Step Process to perform A/B testing

A/B testing has five key components. Following these steps creates data-driven problem solving quickly and iteratively to achieve success for your business.



How to write a good hypothesis

A hypothesis is a clearly defined statement to predict what you think the outcome of the experiment will be. It is **not** an open-ended question which you hope to answer. Through our test, we can either prove that hypothesis true or false. A good way of phrasing hypotheses is through the If/then framework:

If _____,

then _____.

How to write a good hypothesis

A hypothesis is a clearly defined statement to predict what you think the outcome of the experiment will be. It is **not** an open-ended question which you hope to answer. Through our test, we can either prove that hypothesis true or false. A good way of phrasing hypotheses is through the If/then framework:

If we adjust our color scheme for users in search engine results listings,

then users will find what they are looking for in less time.

Step-by-Step Process to perform A/B testing

A/B testing has five key components. Following these steps creates data-driven problem solving quickly and iteratively to achieve success for your business.



How to Decide Which Test is Worth it?

Prioritizing Your Growth Experiments with PIE/ICE

PIE and ICE are the most widely used frameworks and they are very similar to each other. The PIE framework was created by Chris Goward (founder of [WiderFunnel](#)); ICE framework was popularised by [Growthhackers](#).

PIE

Potential
Importance
Ease

ICE

Impact
Confidence
Ease

PIE in Practice

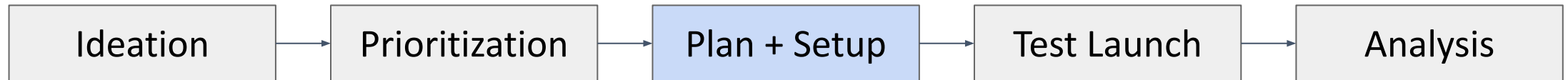
List your experiments, along with their Potential, Importance and Ease on a scale of 1-10. The PIE score is simply the average of these values.

Experiment	Potential	Importance	Ease	PIE Score
Change button color from blue to red	10	7	5	7.33
Another example for a test	8	6	10	8.00
Third idea for an experiment	10	10	1	7.00

A “**Painted Door**” test creates a minimal user experience that mimics a full feature set and measures the number of users who engage with it. Behind the **painted door**, you may include a survey or some other mechanism for collecting additional qualitative data.

Step-by-Step Process to perform A/B testing

A/B testing has five key components. Following these steps creates data-driven problem solving quickly and iteratively to achieve success for your business.



Technical Requirements for A/B Testing

There are **four basic requirements** to be able to perform an A/B test:

1. The difference between groups should be limited to a **single variable**.
2. Assignment to Variant A or Variant B must be **random**. Randomness helps smooth out differences between group members that are not included in what is being tested (mobile vs desktop, for example).
3. Often one variant is the existing condition (control group) and the other is the experimental condition.
4. **Sample size should be defined at the outset.**

Designing your experiment

The first step is to understand how much data you need.

Because it is almost always impossible to measure each member of a population, we can sample the population in order to infer something about it.

Examples:

- Exit polls to estimate the winner in an election
- Testing a new vaccine as a “trial” before making it available as a treatment

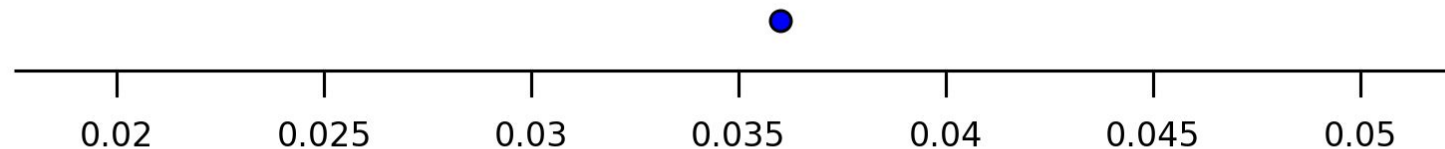
How big a sample you need depends on the ***size of the effect*** you are trying to measure.

Statistical Estimation Refresher

We run an ad which gets 1000 impressions and 36 clicks.

We want to estimate the **true click-through rate** of the entire population that will see the ad.

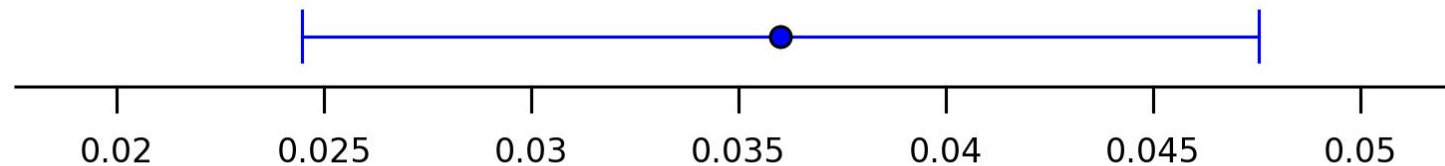
Based on our data, we can generate a **point estimate**: $36 / 1000 = 0.036$.



Statistical Estimation Refresher

But this point estimate also has a **margin of error** since we are only looking at a sample and not the entire population.

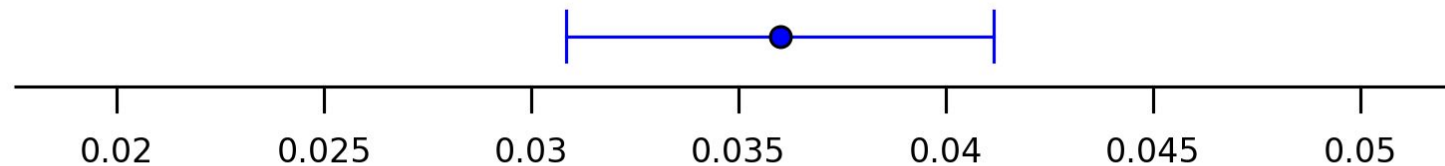
If the true rate is anywhere within the error bars, our observation would not be considered unusual.



Statistical Estimation Refresher

With more data, the margin of error shrinks.

For example, if we had 5000 impressions with 180 clicks, our point estimate would again be $180 / 5000 = 0.036$, but we would have a much smaller margin of error.



Hypothesis Testing

When performing an A/B test, you are not just looking at an estimate of a single click-through rate.

Instead, you are comparing the click-through rates of two versions, A and B, to determine if there is a difference in these click-through rates.

Hypothesis Testing

Start with a **null hypothesis** that there is *no difference* in click-through rates between version A and version B.

Then gather your data and calculate how unusual your data would be if the null were in fact true. The probability of seeing data at least as extreme as you observed is called the ***p*-value**.

If your observation is *unlikely enough*, reject the null hypothesis in favor of the **alternative hypothesis** that **there is a difference** in click-through rates.

Traditionally, the cutoff for rejecting the null is a $p\text{-value} < 0.05$. This cutoff is called the **significance level**.

Hypothesis Testing

The p -value is **not**:

- The probability that the null hypothesis is true
- The probability that version A is better than B or vice versa
- The probability that you would get a different result if you reran the experiment
- The probability that the result is due to chance

The p -value indicates the ***probability of seeing the difference you did see (or a more extreme difference), if there is in fact no difference between A and B.***

Example

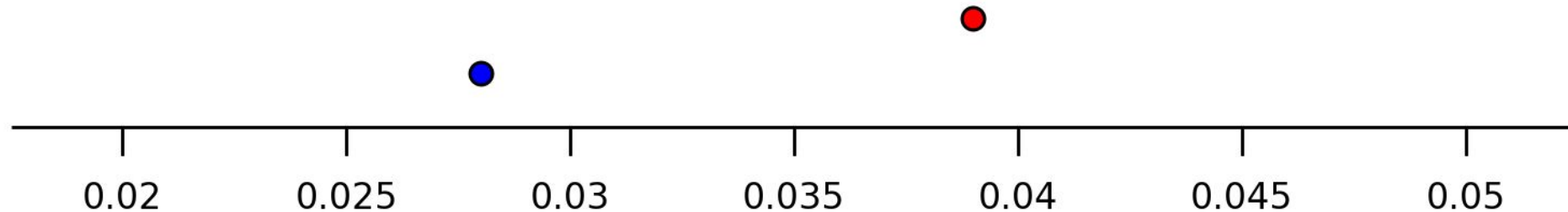
We run two versions of an ad and want to test if there is a statistically significant difference in the click-through rates. Here is the data we gather.

Version	Impressions	Clicks	Point Estimate
A	1000	28	$28/1000 = 0.028$
B	1000	39	$39/1000 = 0.039$

Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

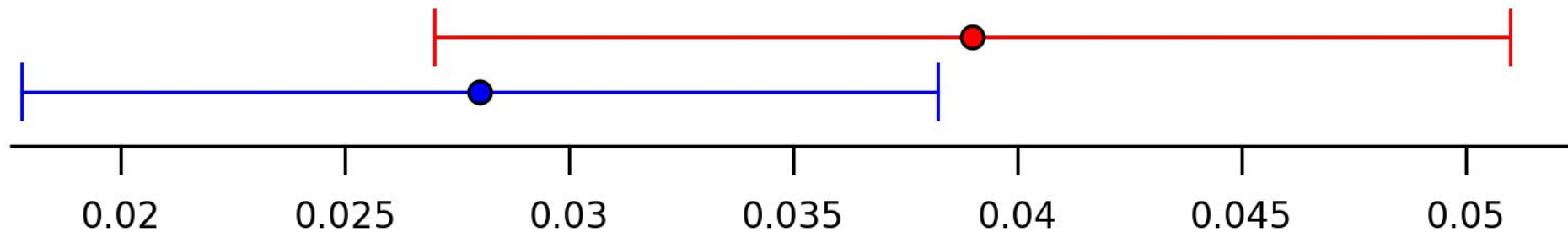
Version	Impressions	Clicks	Point Estimate
A	1000	28	$28/1000 = 0.028$
B	1000	39	$39/1000 = 0.039$



Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	1000	28	$28/1000 = 0.028$
B	1000	39	$39/1000 = 0.039$



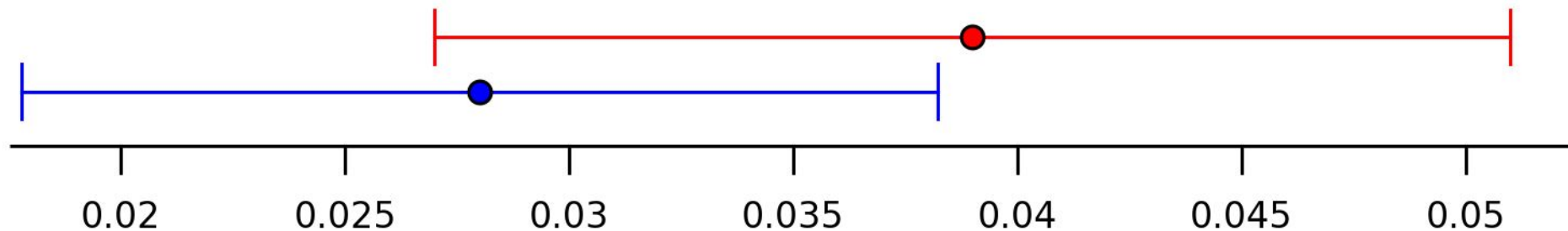
Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	1000	28	$28/1000 = 0.028$
B	1000	39	$39/1000 = 0.039$

p-value: 0.172

Do not reject the null.



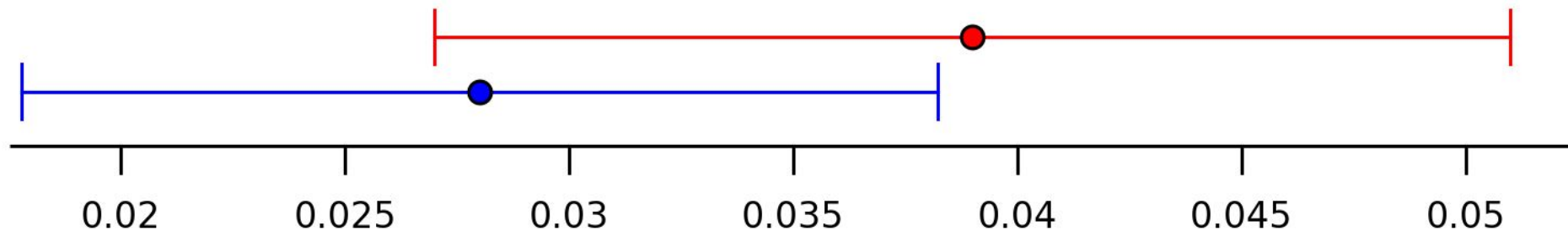
Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	1000	28	$28/1000 = 0.028$
B	1000	39	$39/1000 = 0.039$

p-value: 0.172

Do not reject the null.



Even though there was a difference in the point estimates, such an observed difference would not be *that* unusual if there was really no difference in click-through rate due to the high uncertainty in our estimates.

Example

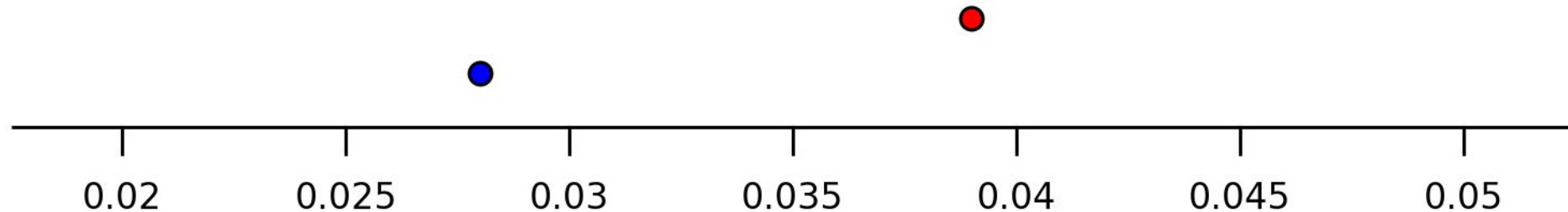
Now, we want to rerun our trial with a larger sample size. Here is the data we gathered.

Version	Impressions	Clicks	Point Estimate
A	5000	140	$140/5000 = 0.028$
B	5000	195	$195/5000 = 0.039$

Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

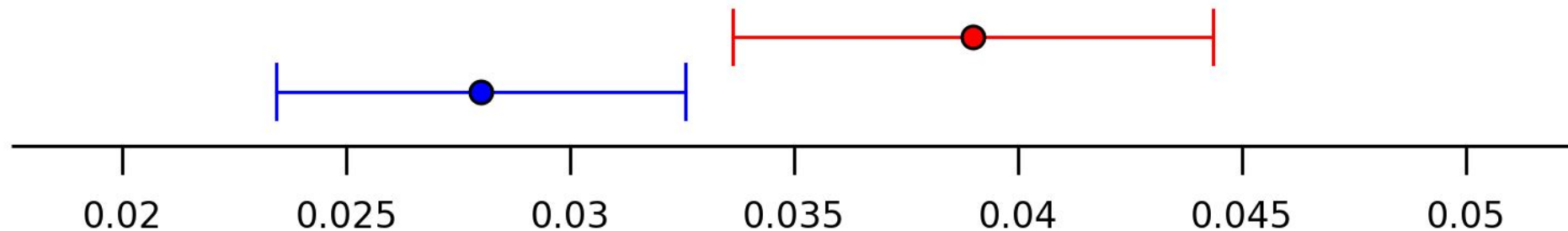
Version	Impressions	Clicks	Point Estimate
A	5000	140	$140/5000 = 0.028$
B	5000	195	$195/5000 = 0.039$



Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	5000	140	$140/5000 = 0.028$
B	5000	195	$195/5000 = 0.039$



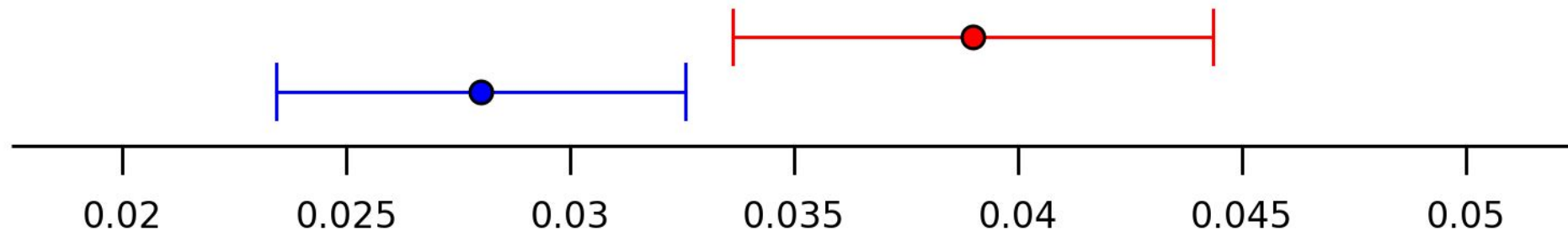
Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	5000	140	$140/5000 = 0.028$
B	5000	195	$195/5000 = 0.039$

p-value: 0.002

Reject the null.

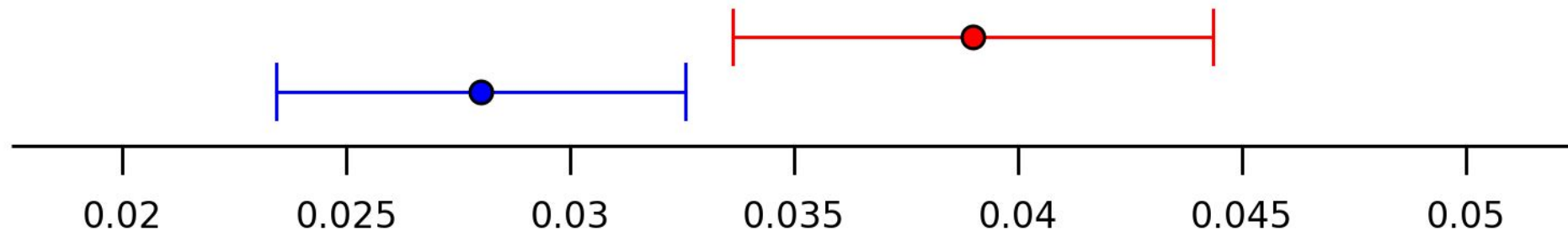


Example

When comparing two options, we have to consider not just the point estimates, but also the margin of error.

Version	Impressions	Clicks	Point Estimate
A	5000	140	$140/5000 = 0.028$
B	5000	195	$195/5000 = 0.039$

p-value: 0.002
Reject the null.



The larger sample size reduces the uncertainty in our estimates. The observed difference would be very unlikely if in reality there was no difference in click-through rates.

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.035	1000	27	$27/1000 = 0.027$
B	0.035	1000	44	$44/1000 = 0.044$

You have collected the results of an experiment (shown above).

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.035	1000	27	$27/1000 = 0.027$
B	0.035	1000	44	$44/1000 = 0.044$

p-value: 0.04

Reject the null.

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.035	1000	27	$27/1000 = 0.027$
B	0.035	1000	44	$44/1000 = 0.044$

p-value: 0.04
Reject the null.

Here, the null hypothesis is actually true, but we rejected it.

We have committed a **Type I Error** by falsely concluding that there is a difference in rates.

The significance level can also be stated as **the probability of a Type I error, given that the null hypothesis is true.**

Choosing the best statistical testing method (traditional / frequentist methods)

How do you actually compute the p -value?
It depends on what you are measuring.

Comparing averages between A and B (eg. average transaction amount)

- Welch's t-test (more reliable when samples have uneven size or variance)
- Student's t-test (assumes samples are normally distributed and have equal variance)

Comparing proportions between A and B (eg. click-through rates)

- Two proportion z-test. Uses a normal approximation, which is very accurate for large samples.
- Fisher's exact test (examines significance of frequencies distributed among categories).

Power

Even if there is a difference in click-through rates, we won't detect it if the margin of error on our estimates is too large.

We want to give ourselves a decent chance to detect a difference, if one does exist. This probability of rejecting the null hypothesis when there is a certain difference in rates is called the **power** of the test.

When estimating sample size, you must choose a desired power level. It is standard to set the desired power to be 0.8.

When estimating a minimum sample size, you need to estimate a **base rate** (often based on historical data) and the **minimum detectable effect**.

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.03	400	12	$12/400 = 0.03$
B	0.02	400	8	$8/400 = 0.02$

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.03	400	12	$12/400 = 0.03$
B	0.02	400	8	$8/400 = 0.02$

p -value: 0.365

Do not reject
the null.

Hypothesis Testing: Errors

Version	True Rate	Impressions	Clicks	Point Estimate
A	0.03	400	12	$12/400 = 0.03$
B	0.02	400	8	$8/400 = 0.02$

p-value: 0.365

Do not reject
the null.

Here, the null hypothesis is actually false, but we did not reject it (even though we got the correct point estimates!).

We have committed a **Type II Error** by not concluding that there is a difference in rates.

The power of a test is $(1 - P(\text{Type II Error}))$ for a given base rate and minimum effect size.

Hypothesis Testing: Errors

TYPE I ERROR



TYPE II ERROR



<https://codingwithmax.com/evaluate-ab-test-controlled-experiments/>

Power Example

You know that historically, a particular ad has a 2.5% click-through rate.

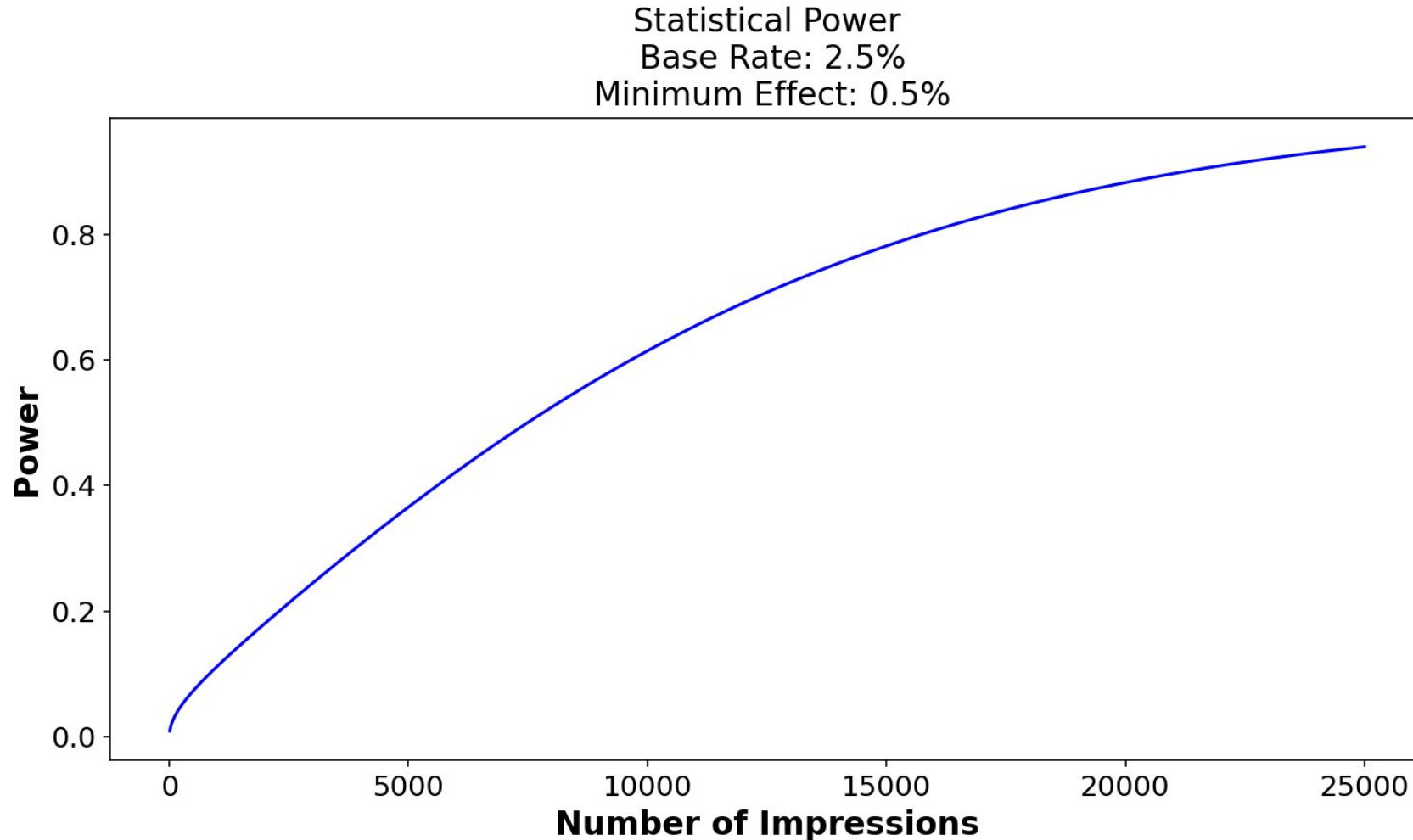
There is a new variant that you want to test and the minimal detectable effect you need to measure in order to make a change is 0.5%

How many impressions do you need in order for there to be an 80% chance of detecting this difference?

That is, **what sample size is needed in order to have a *power* of 0.8?**

Power

As sample size increases, the power increases (nonlinearly).



Power

There are numerous sample size calculators available. For example, <https://www.evanmiller.org/ab-testing/sample-size.html>

Question: How many subjects are needed for an A/B test?

Baseline conversion rate:

%

2.5%

[[link](#)]

Minimum Detectable Effect:

%

2% – 3%

The Minimum Detectable Effect is the smallest effect that will be detected (1- β)% of the time.

☒ Absolute
☐ Relative

Conversion rates in the gray area will not be distinguishable from the baseline.

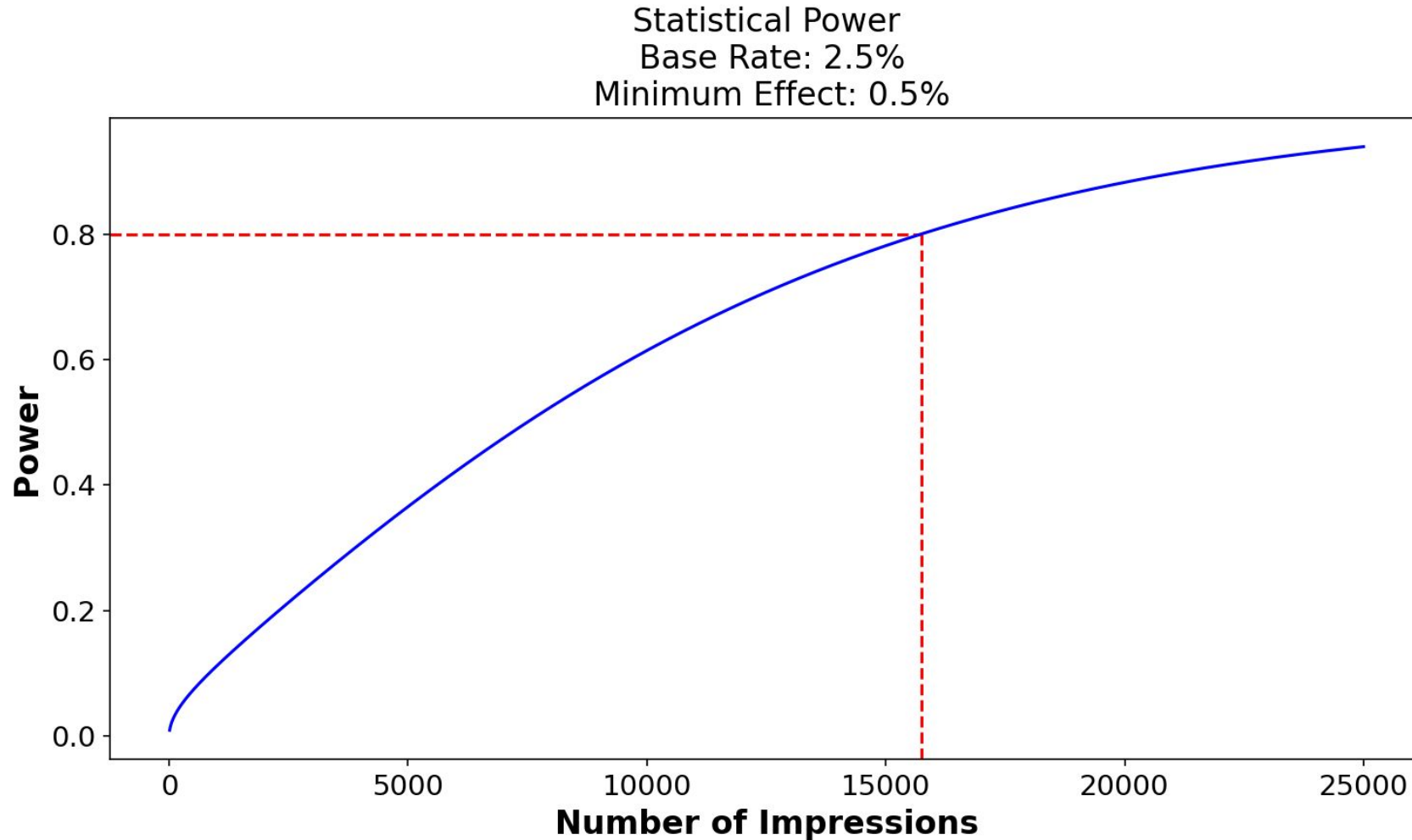
Sample size:
15,744
per variation

Statistical power $1-\beta$: 80% *Percent of the time the minimum effect size will be detected, assuming it exists*

Significance level α : 5% *Percent of the time a difference will be detected, assuming one does NOT exist*

Power

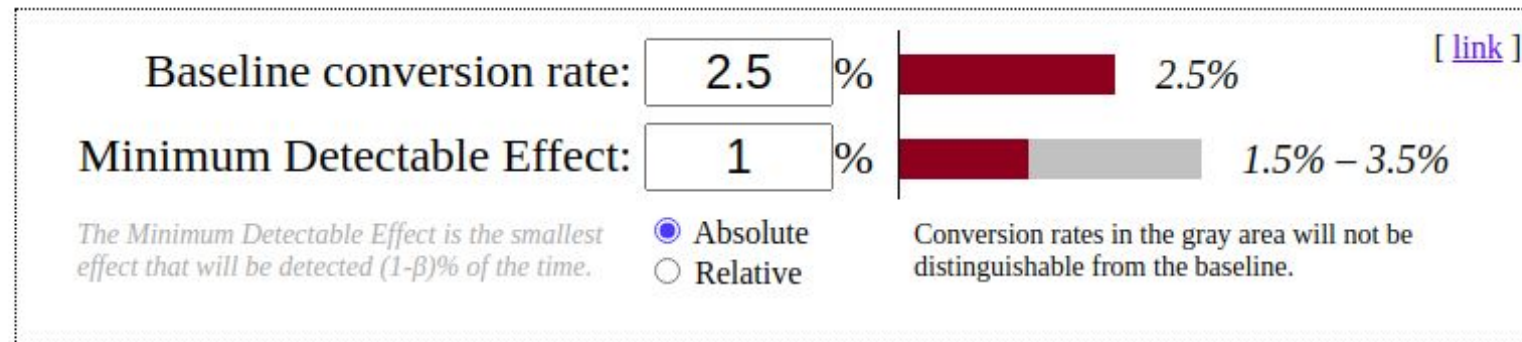
We can see the sample size requirement on our plot.



Power

For larger effect, a smaller sample size is needed for the same amount of power.

Question: How many subjects are needed for an A/B test?



Sample size:

4,041

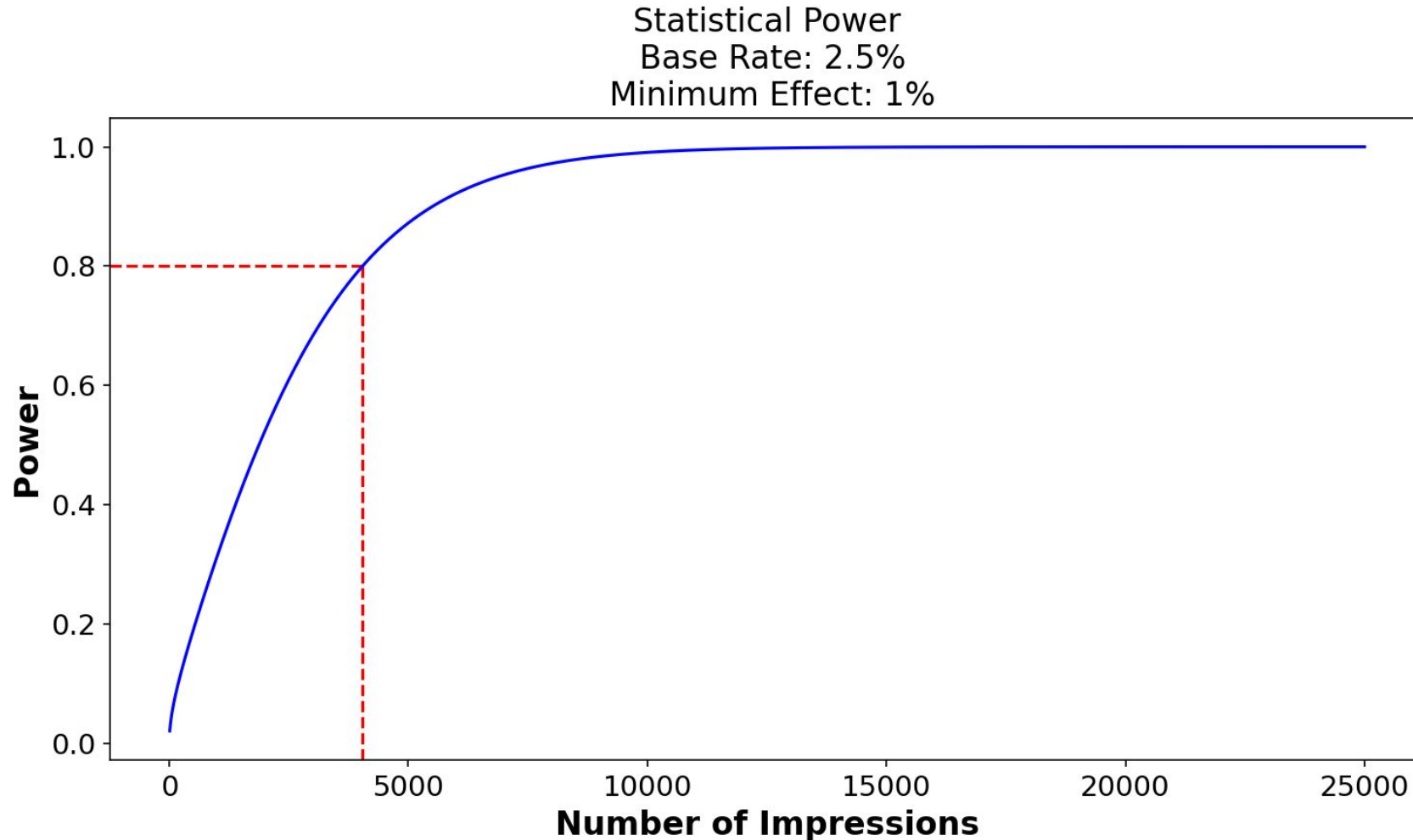
per variation

Statistical power $1-\beta$: 80% *Percent of the time the minimum effect size will be detected, assuming it exists*

Significance level α : 5% *Percent of the time a difference will be detected, assuming one does NOT exist*

Power

For larger effect, a smaller sample size is needed for the same amount of power.



Choosing the best statistical testing method (Bayesian)

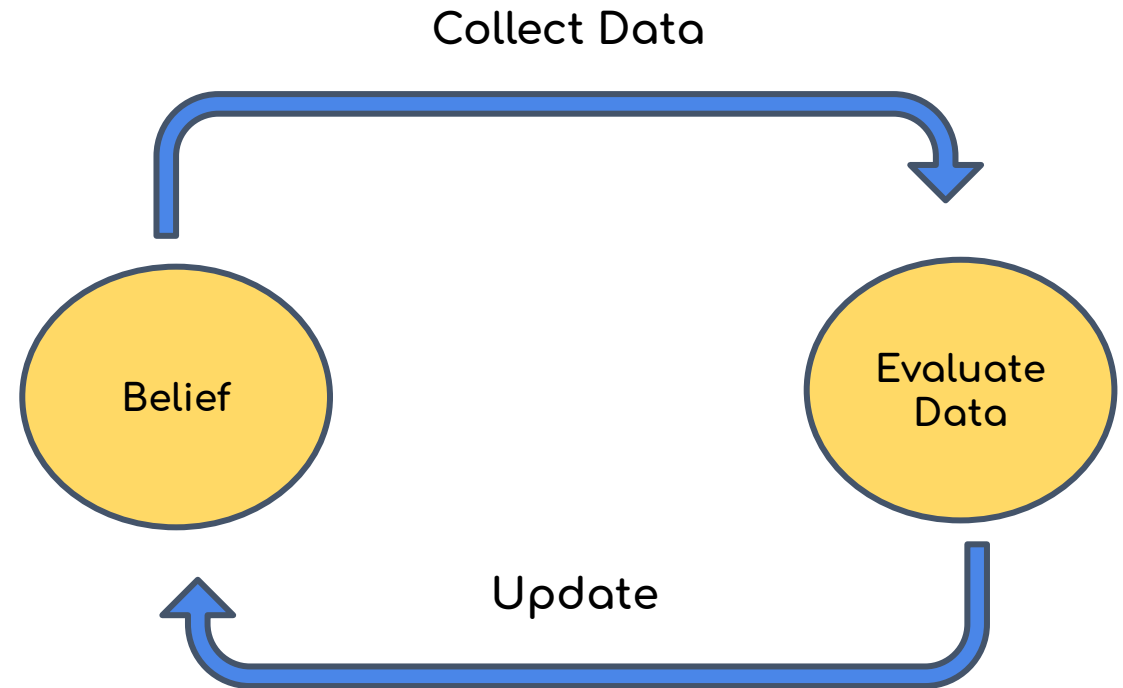
The frequentist approach has the disadvantage that you can't talk directly about the probability of A being better than B.

The Bayesian approach allows one to directly talk about the probability that variant A leads to a significantly different outcome than variant B.

Bayesian Analysis Steps:

1. Begin with an assumption/belief
2. Learn from the data collected
3. Update your belief in a way that combines the initial assumption/belief with what has been learned from the data.

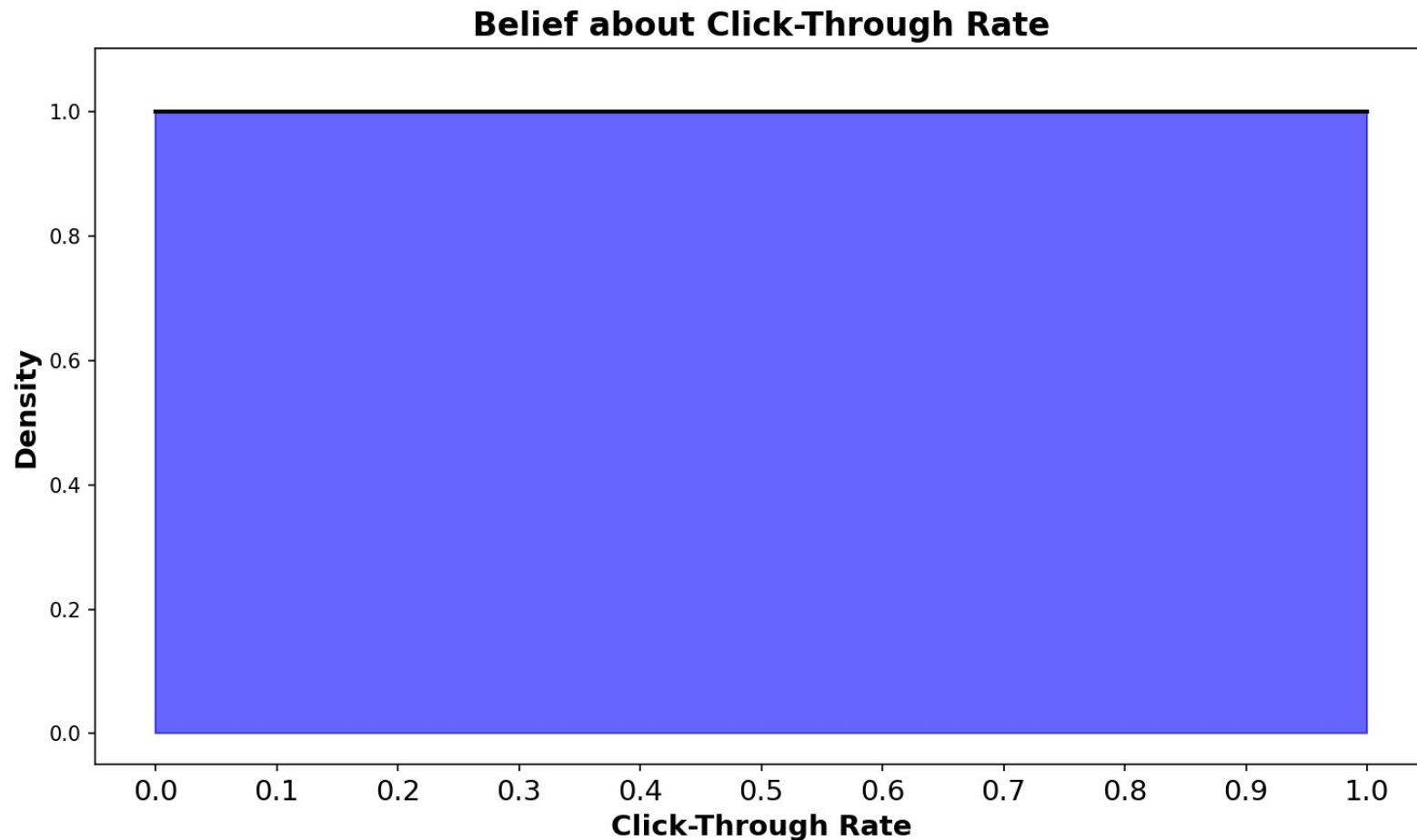
As you gather more data, you can continually update your belief.



Bayesian Example

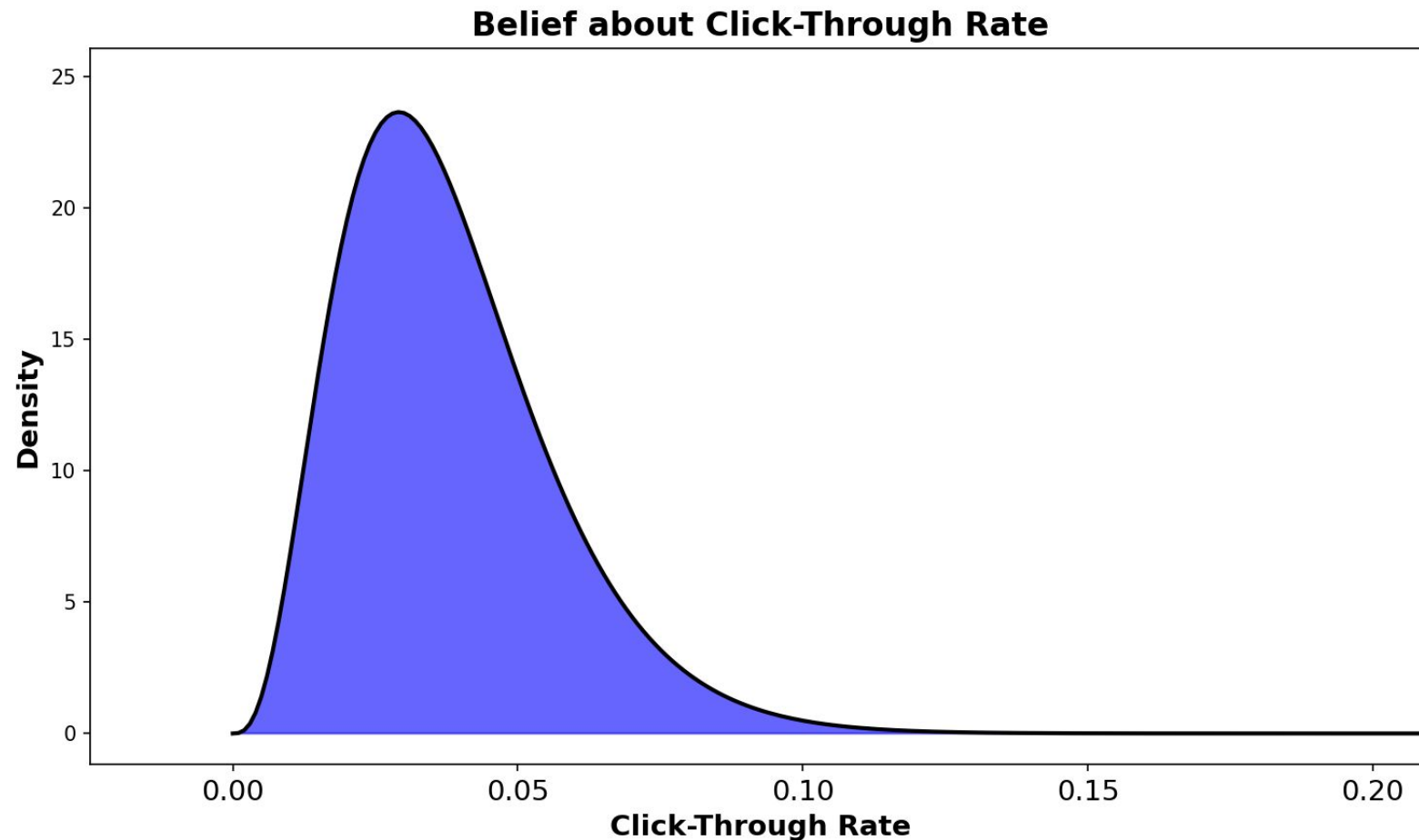
You are launching a new ad campaign. As you gather data, you want to estimate the click-through rate.

You want to start your estimation with no prior belief at all, meaning that all click-through rates are equally likely.



Bayesian Example

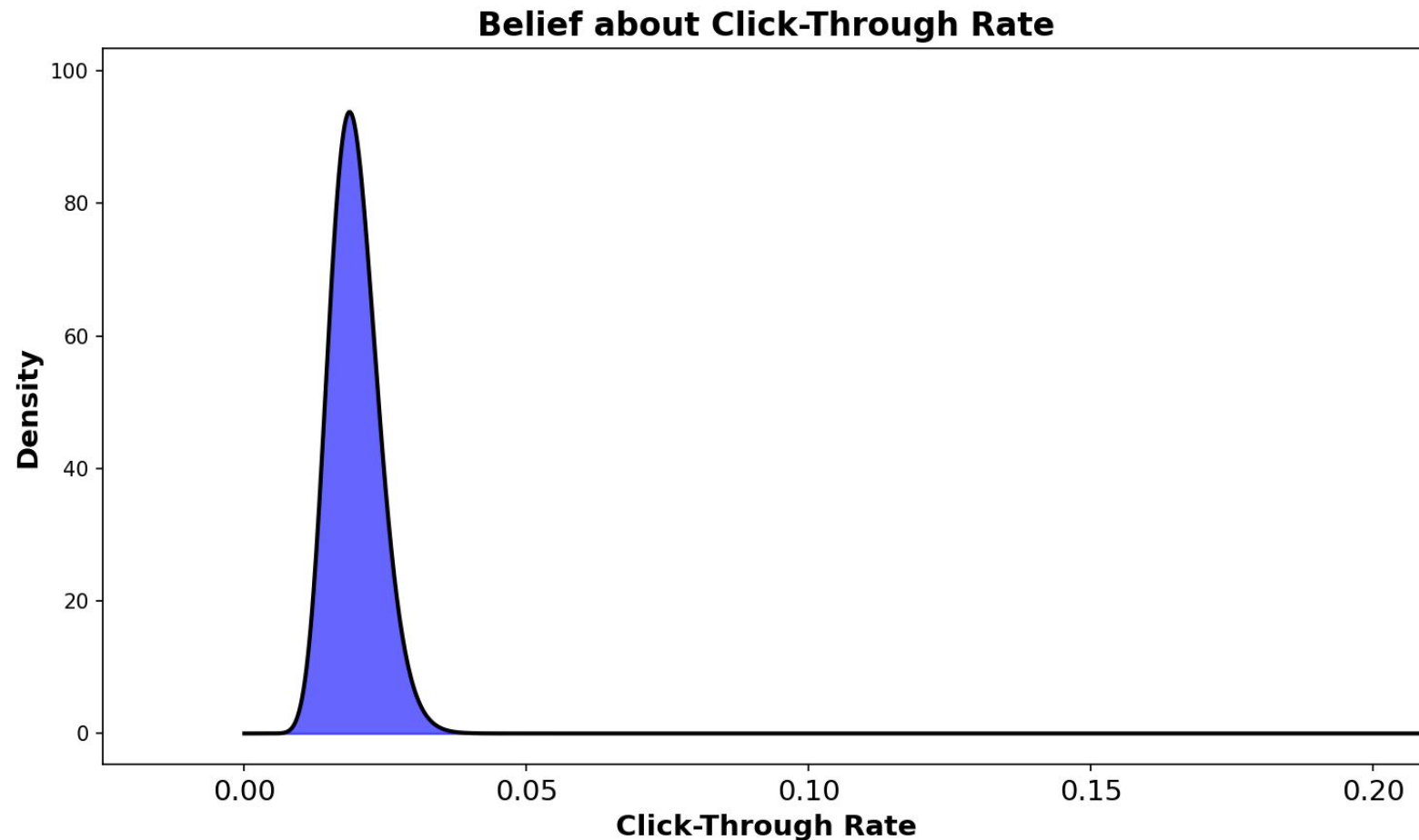
After 100 impressions, there are 3 clicks.
Based on this data, we can update our belief.



Bayesian Example

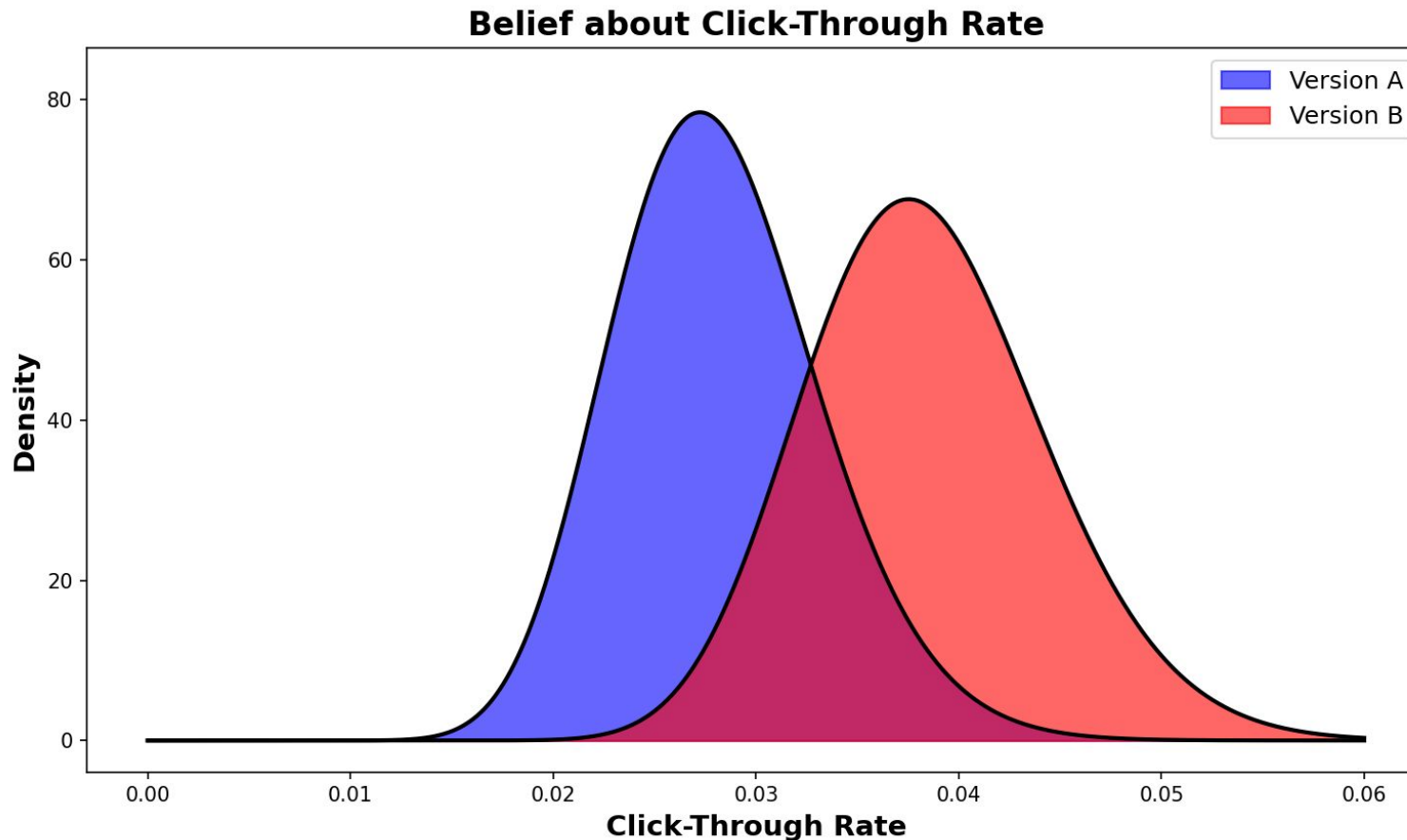
After 1000 impressions, there are 19 clicks.

With this much data, the range of probable click-through rates becomes very narrow.



Bayesian Example

What if we want to compare click-through rates?



Borrowing the data from the prior example:

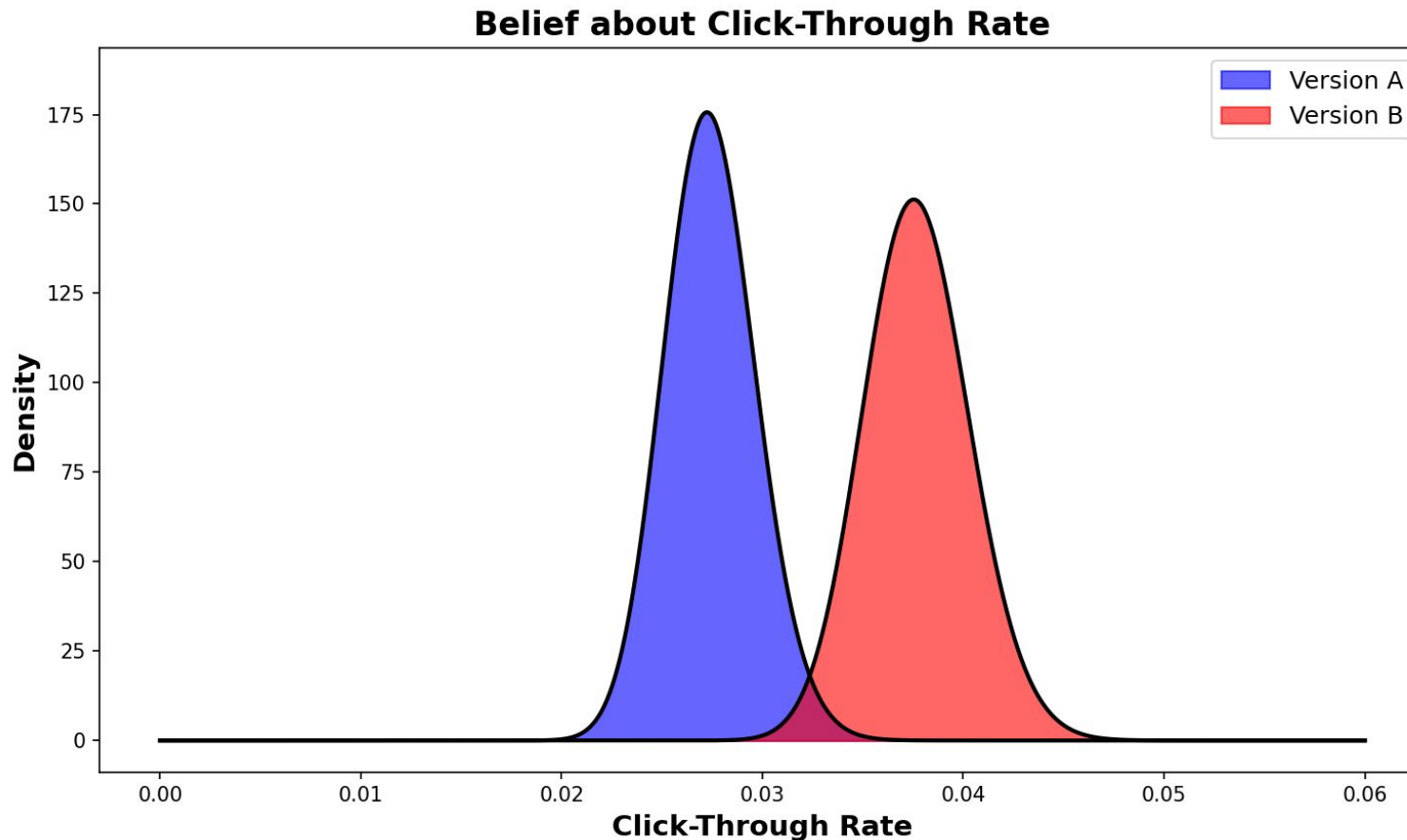
Version	Impressions	Clicks
A	1000	28
B	1000	39

Probability that **B** has a higher click-through rate than **A**:

0.905

Bayesian Example

What if we want to compare click-through rates?



Borrowing the data from the prior example:

Version	Impressions	Clicks
A	5000	140
B	5000	195

Probability that **B** has a higher click-through rate than **A**:

0.998

Real-life Examples of Bayesian vs. Frequentist Models

There are applications for each statistical method and different statisticians prefer different methods. Thus, it is important to know that different platforms might use different statistical approaches:

Frequentist



Adobe
Target

Bayesian



Google
Optimize

Hybrid



Optimizely

Step-by-Step Process to perform A/B testing

A/B testing has five key components. Following these steps creates data-driven problem solving quickly and iteratively to achieve success for your business.



Recalling that A/B testing is iterative...

The key is to run a test, analyze it, derive learnings and then tweak the test to continually improve. There is a misconception that a winner has to be found on the first try, however, it is really about the *learning* you gain from it. It's an excellent way of uncovering user behavior in the wild, even if a test does not have a conclusive winner.



Will it A/B test?

Travis thinks that people will be more likely to read a blog post if he features it on the main page of his website instead of relying on visitors to click through the menu to the page of blog posts.

Is there a single variant?

Yes – blog featured on landing page vs blog archive

Can users be randomly assigned?

Yes – can determine on website page load whether to show version A or version B

Does he have a target outcome to measure?

Yes – If he requires the user to click to read most or all of the blog post

Can he define the sample size?

Yes - he can set the number of visits to his website to sample

Will it A/B test?

A local radio station wants to determine if listeners are more likely to call in and share stories if the subject of the stories is **personal triumphs** versus **funny things that happened to me**. The station manager sets up Funny Things Call-in Night on Saturday and Personal Triumph Night on Sunday and counts the number of callers over a 3-month trial.

Is there a single variant?

No – story theme and night of show (potentially also the show's host) are differences

Can users be randomly assigned?

No – they are self-selecting

Does he have a target outcome to measure?

Yes – the number of participants in each show

Can he define the sample size?

Yes – he has set a time period during which he expects to have a sufficient sample

Will it A/B test?

An experimental drug to enhance problem solving has been developed by PharmaNow. 1000 volunteers are recruited and randomly assigned to receive either the new drug or a placebo. Afterwards, all volunteers are given an identical set of puzzles to solve.

Is there a single variant?

Yes – the new drug

Can users be randomly assigned?

Yes – they are assigned to receive either the drug or a placebo at random

Does he have a target outcome to measure?

Yes – solve rate for the puzzles

Can he define the sample size?

Yes – sample set at 1000 volunteers

Will it A/B test?

Jodie is running for mayor. She wants to see if people are more likely to donate to her campaign if a **Donate Now** button is placed at the top of her “**How to help**” page instead of letting it remain at the bottom of the page where it is currently placed.

Is there a single variant?

Yes – placement of the button

Can users be randomly assigned?

Yes – can decide upon loading the page whether to load Variant A or Variant B

Does she have a target outcome to measure?

Yes – can count the number of donors with each version of the page

Can she define the sample size?

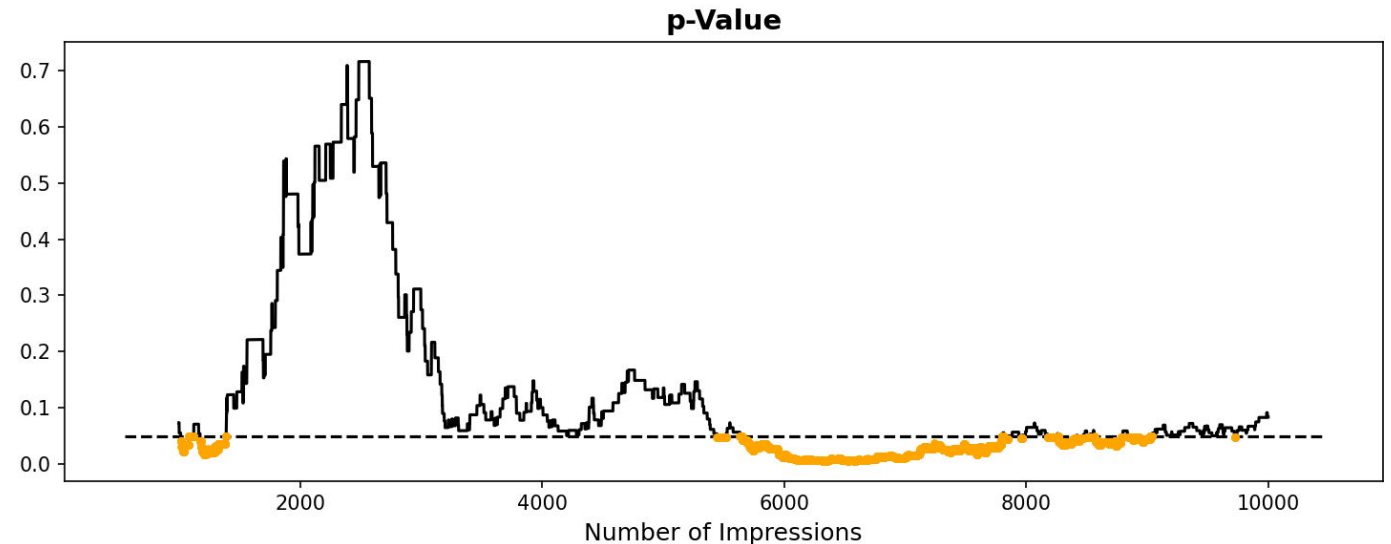
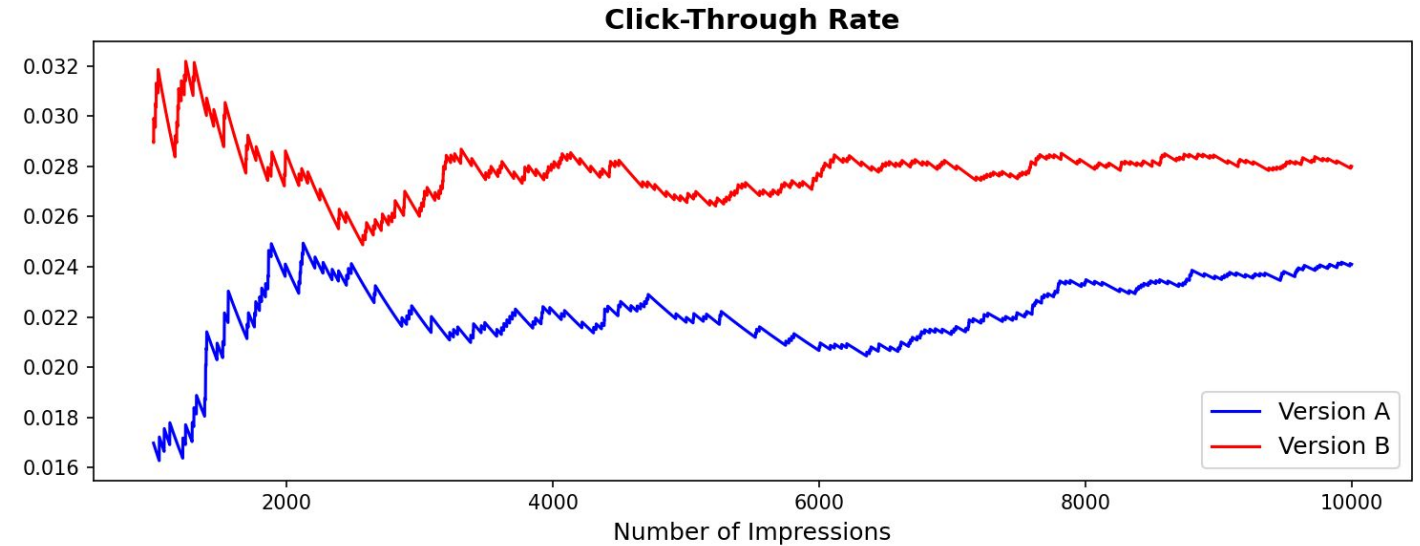
Yes – she can set the number of visitors to the How to Help page

Important Considerations for A/B Testing

Sample size must be fixed ahead of time and must be kept.

Do not stop the test until there is a statistically significant result.

The probability that the p -value is less than 0.05 at some point during the test, given that the null hypothesis is true, is *much* higher than 0.05



Important Considerations for A/B Testing

Decide which metrics to monitor ahead of time to avoid the multiple-testing problem/spurious correlations.

Beware of Simpson's paradox if there are multiple input streams.

Retest frequently; likes and preferences change.

Only make one change at a time so that you can identify which feature is responsible for the change. To test multiple features simultaneously, you need to do **multivariate testing**, which requires a much larger sample size.

Limitations of A/B Testing and Alternatives

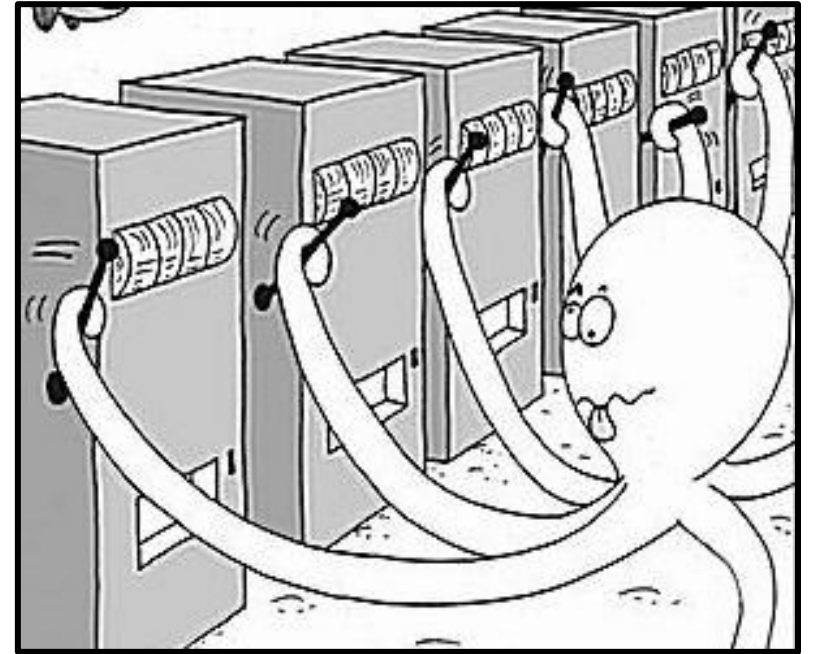
Requires fixed sample size for each variant from the outset.

You spend the whole test using a less-than optimal option, potentially costing clicks or revenue.

A/B testing limits you to two variants, unless you significantly increase sample size and test time.

Multi-Armed Bandits balance out *exploration* (searching for the optimal variant) with *exploitation* (directing more users to the optimal variant).

The mathematics are significantly more complicated for bandit algorithms and are an active area of research.



Case Study - Simultaneous Google Ad Campaigns

Variant A – emphasis on launching a career

Ad • learn.nashvillesoftwareschool.com...

Launch A Career in Analytics | Nashville'...

Learn to apply statistical reasoning with...

Example ad

Ad group

Training

Type

Standard

Keywords

SQL training, Tableau training, analytics training
+ 7 more

Ad extensions

Sitelink extension, Callout extension, Dynamic
structured snippet

Variant B – emphasis on learning

Ad • learn.nashvillesoftwareschool.com...

Learn Data Analytics | Nashville's Nonpr...

Learn to apply statistical reasoning with...

Example ad

Ad group

Root/Tools

Type

Standard

Keywords

Excel, Power BI, SQL, Tableau, business
intelligence, data analysis + 1 more

Ad extensions

Sitelink extension, Callout extension, Dynamic
structured snippet

Case Study

The ad campaigns are identical except for the headlines. They use the same ad groups, which define the keywords that might lead to the ad being returned at the top of the search results.

The screenshot shows the Google Ads 'Ad groups' page. At the top, a line chart displays performance from Sep 4, 2020, to Sep 10, 2020, with a value of 0.00%. Below the chart is a table of ad groups. The table has columns for Ad group, Status, Ad group type, Clicks, Impr., CTR, Avg. CPC, Cost, Conv. value, Conv. rate, Conversions, and Cost / conv. The ad groups listed are Root/Tools, Skills, Training, Learn, Class/Bootcamp, and Competition, all with a status of 'Eligible' and a type of 'Standard'. A 'Total' row shows 0 clicks, 2 impressions, and a 0.00% CTR.

Ad group	Status	Ad group type	Clicks	Impr.	CTR	Avg. CPC	Cost	Conv. value	Conv. rate	Conversions	Cost / conv.
Root/Tools	Eligible	Standard	0	0	—	—	\$0.00	0.00	0.00%	0.00	\$0.00
Skills	Eligible	Standard	0	0	—	—	\$0.00	0.00	0.00%	0.00	\$0.00
Training	Eligible	Standard	0	1	0.00%	—	\$0.00	0.00	0.00%	0.00	\$0.00
Learn	Eligible	Standard	0	0	—	—	\$0.00	0.00	0.00%	0.00	\$0.00
Class/Bootcamp	Eligible	Standard	0	1	0.00%	—	\$0.00	0.00	0.00%	0.00	\$0.00
Competition	Eligible	Standard	0	0	—	—	\$0.00	0.00	0.00%	0.00	\$0.00
Total...			0	2	0.00%	—	\$0.00	0.00	0.00%	0.00	\$0.00
Total...			0	2	0.00%	—	\$0.00	0.00	0.00%	0.00	\$0.00

The skills ad group contains keywords like “data analysis,” “business intelligence,” and “Excel.”

The screenshot shows two side-by-side tables from the Google Ads interface. The left table is 'Ad groups' and the right table is 'Keywords'. Both tables show performance metrics for the 'Skills' ad group.

Ad groups	Avg. CPC	Clicks	CTR
Root/Tools	\$0.00	0	0.00%
Skills	\$0.00	0	0.00%
Training	\$0.00	0	0.00%

Keywords	Cost	Impressions	CTR
"data analysis"	\$0.00	0	0.00%
"business intelligence"	\$0.00	0	0.00%
"Excel"	\$0.00	0	0.00%

Case Study

Null Hypothesis: Click rates will be identical for each of the two ads

Alternative Hypothesis: Searchers will be more likely to click on one ad and visit the Nashville Software School website than they will the other ad

Case Study

Is there a single variant?

Yes – the ad headline

Can users be randomly assigned?

Maybe – this is a bit of a black box, in that we don't know what's happening behind the scenes. Essentially the ads are competing against one another with identical keywords and identical budgets so they should get equal time. Ideally, we would have more direct control of the randomization process.

Do we have a target outcome to measure?

Yes – clicks on the ad

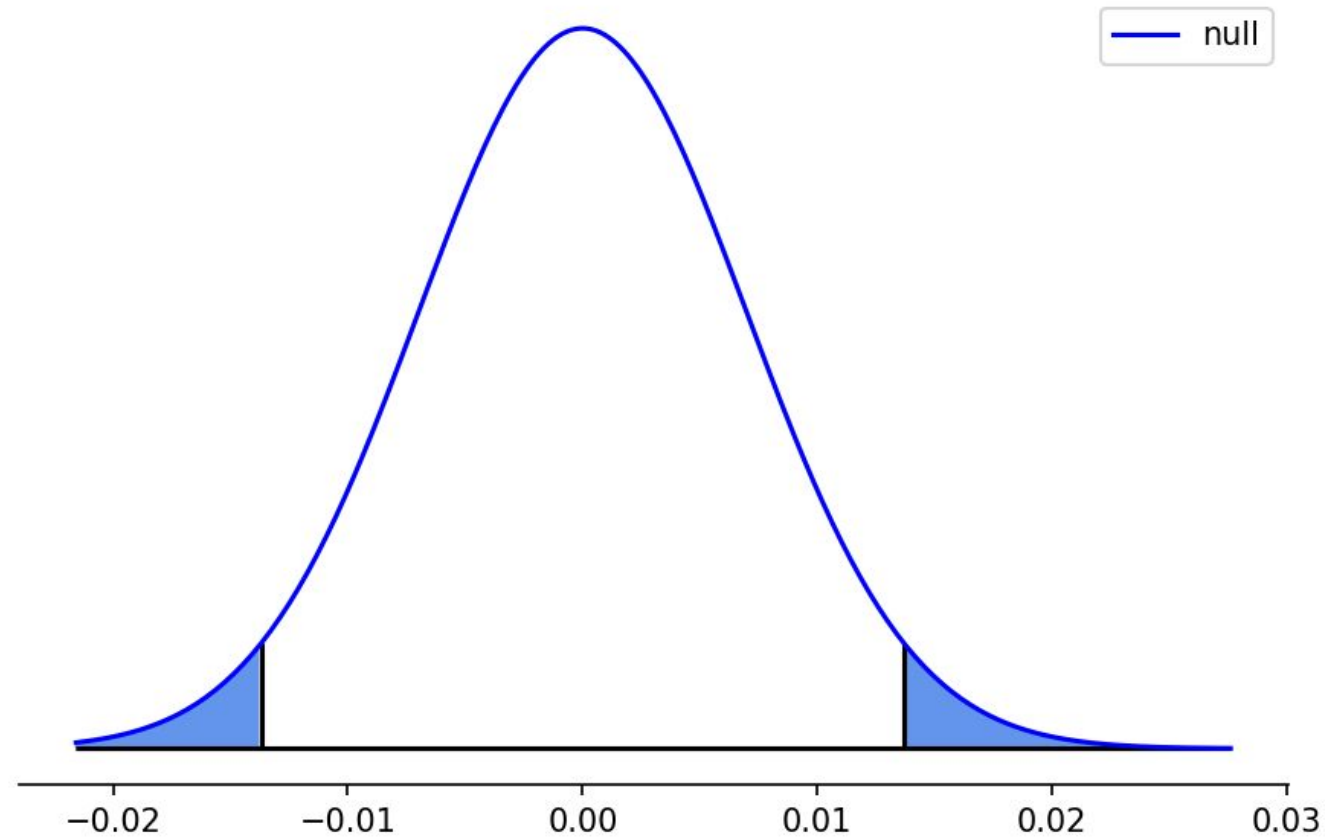
Can we define the sample size?

Yes – Historically, on average the click-through rate has been around 1.50 percent. A noticeable effect from this baseline is defined as 1.0 percent. So we have targeted at least 2535 impressions (presentation of each ad) and -- based on past ad campaigns – set the competing campaigns to run for 2 weeks

Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

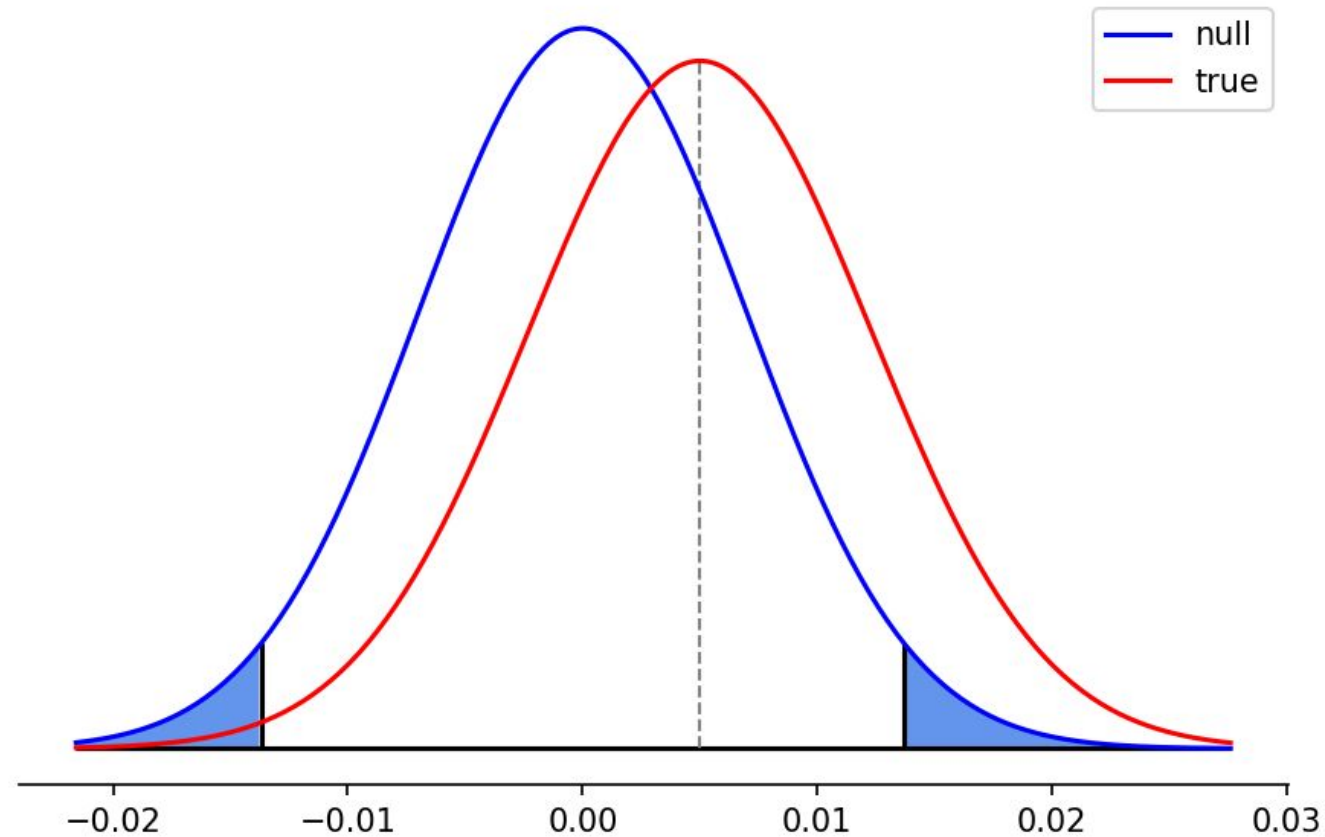
Sample Size = 1000



Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

Sample Size = 1000

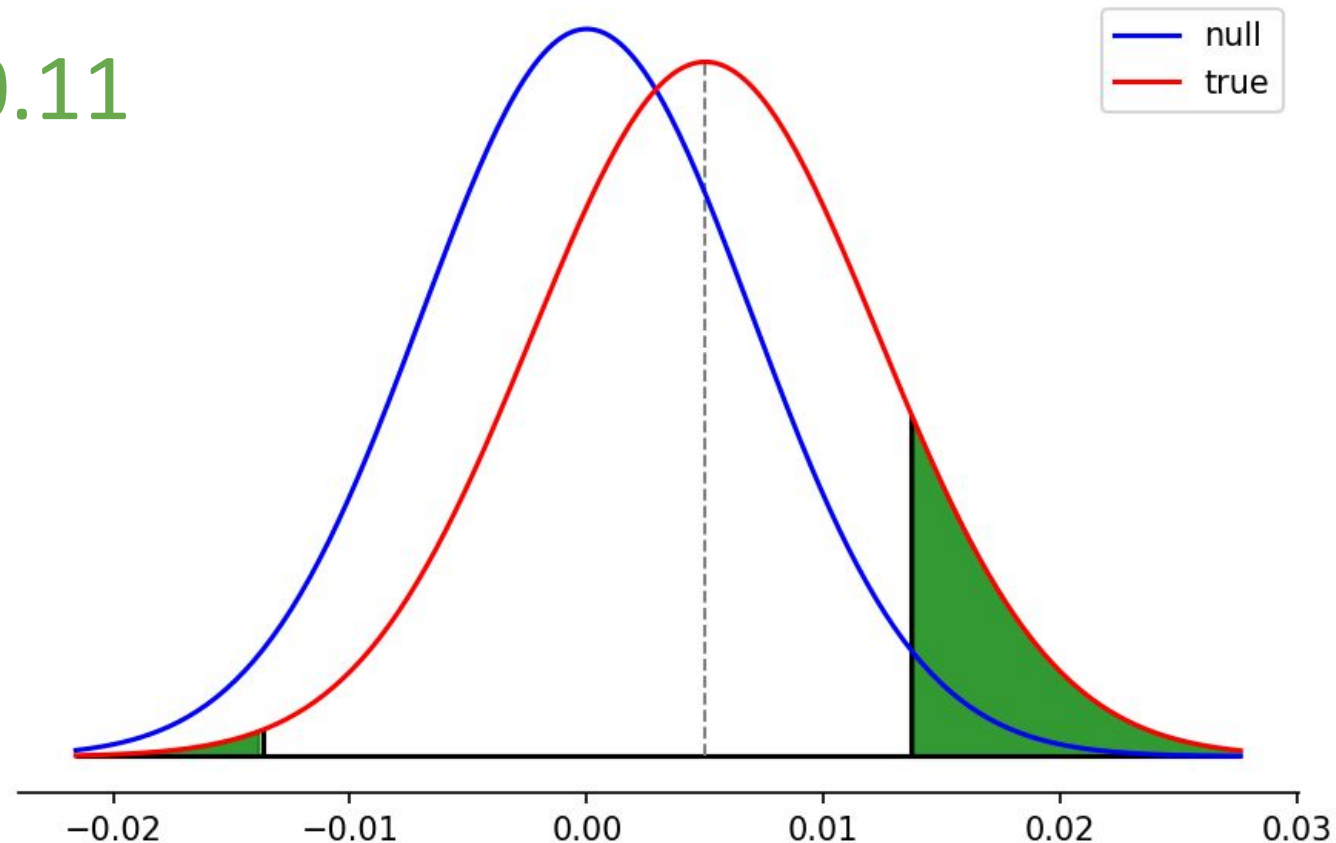


Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

Sample Size: 1000

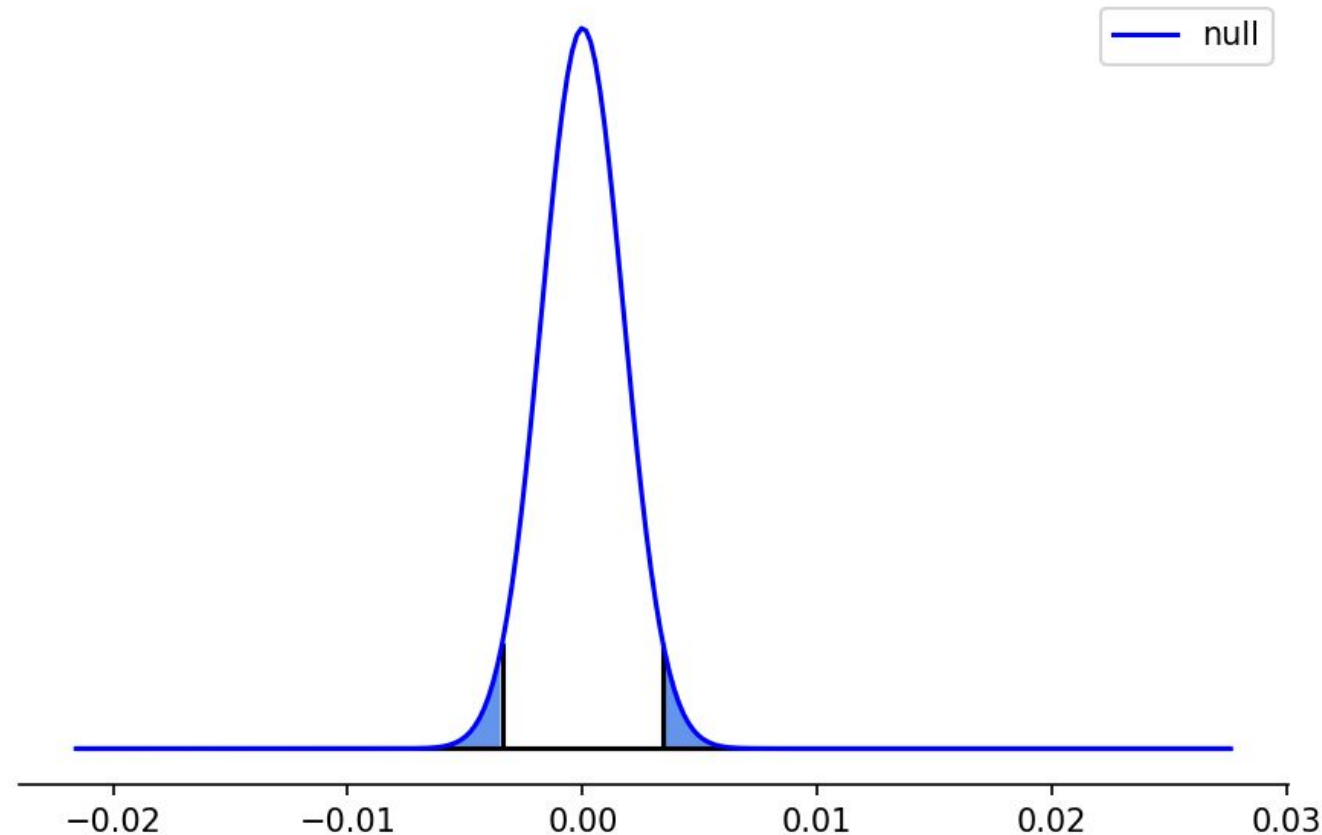
power: 0.11



Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

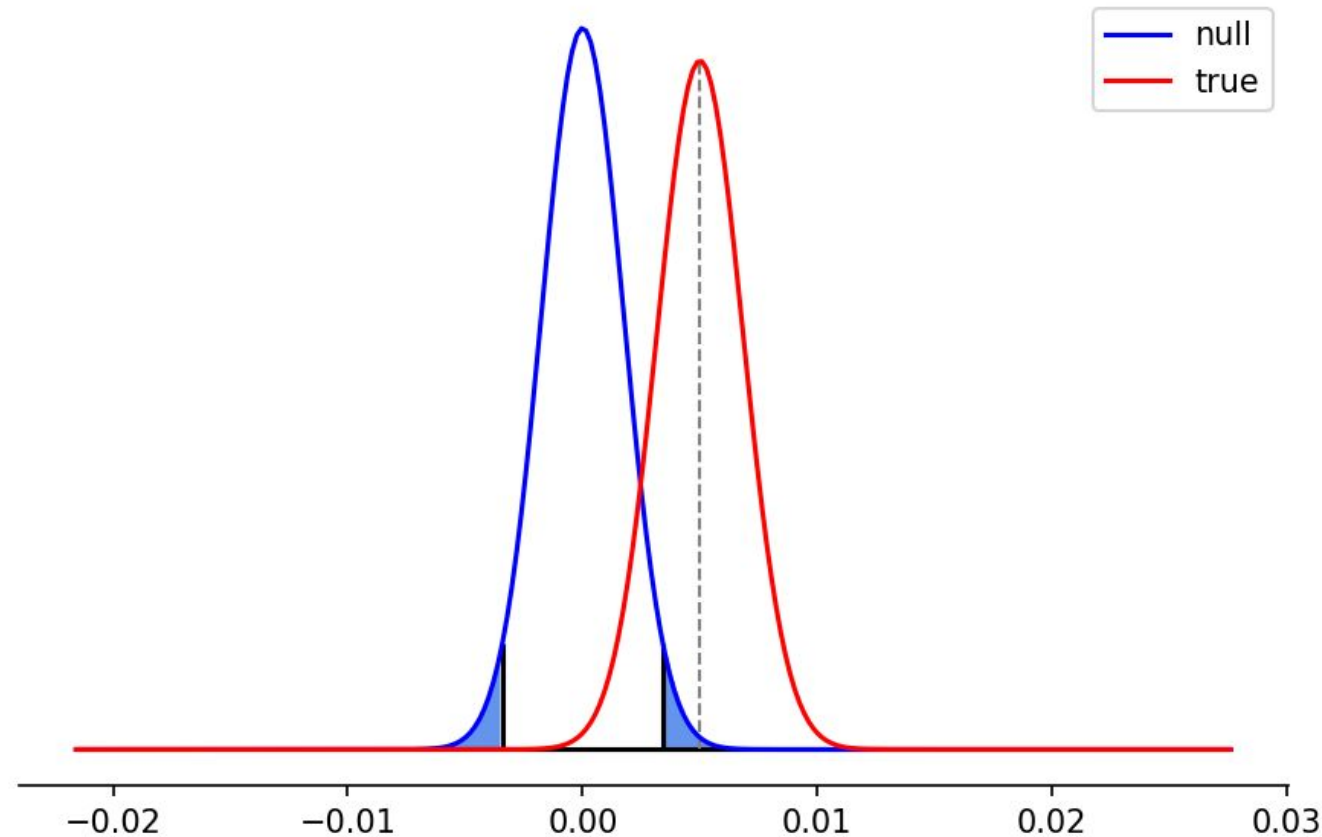
Sample size = 16000



Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

Sample Size = 16000



Power

Fact: the sampling distributions for the difference in proportions are approximately normal with a standard deviation which shrinks as sample size grows.

Sample Size: 16000

power: 0.81

