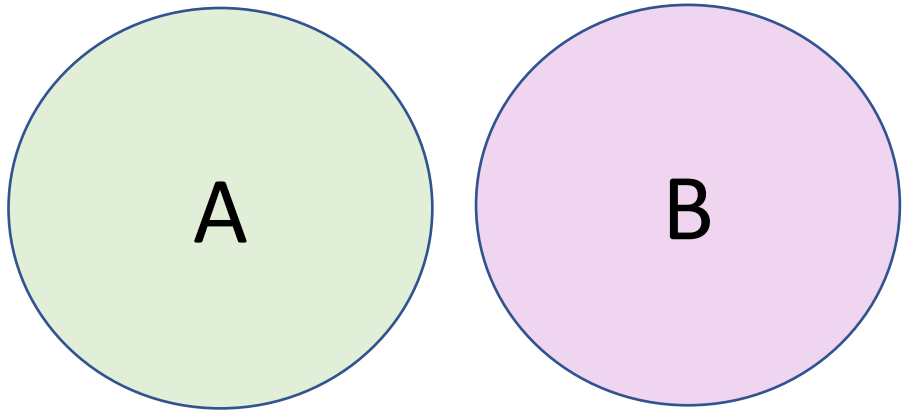
An Introduction to A/B Testing - Part I



NASHVILLE SOFTWARE SCHOOL

- What is A/B Testing?
- History
- Present day use
- How is statistical significance determined?
- Exercise: Will it A/B Test?
- Our case study

What is A/B Testing?

An A/B test is a simple controlled experiment, in which a single variant 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.

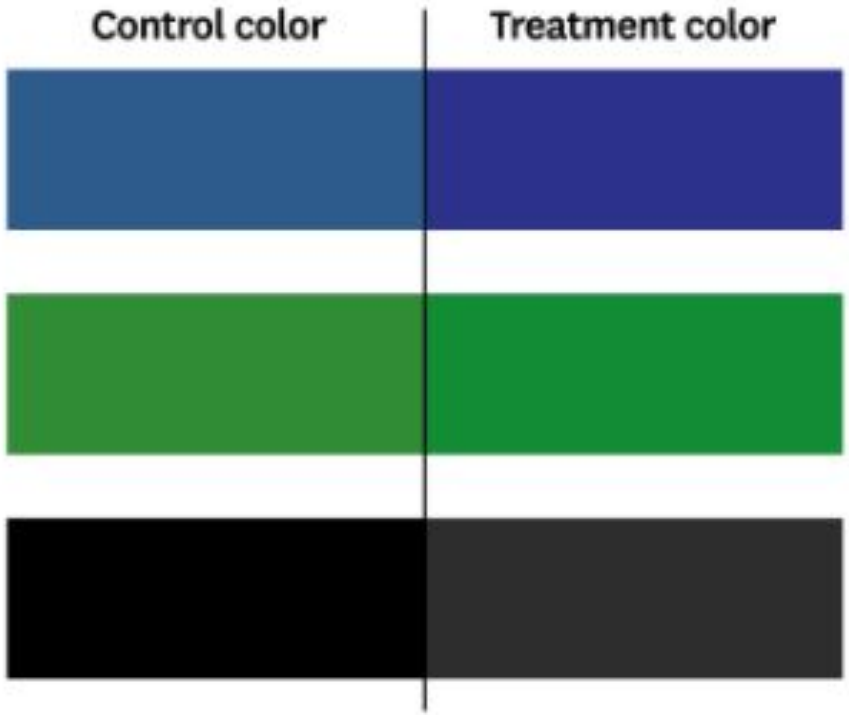
- The difference between groups should be limited to a **single variable**.
- Assignment to Variant A or Variant B must be **random**. Randomness helps smooth out differences between group members that are not included in what is being tested (mobile vs desktop, for example).
- Often one variant is the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

History of A/B Testing

- 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 questions like how will this fertilizer affect crops on this land?
  - Clinical trials often follow A/B test pattern
- Modern usage of A/B testing 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

https://hbr.org/2017/06/a-refresher-on-ab-testing

In 2013, Bing experimented with color in captions and titles on their search results pages.
Users shown titles in darker blues and darker greens along with captions in a lighter black color were more successful in their searches.



Twyman's Law: Any figure that looks interesting or different is usually wrong

Repetition is key to validating results like the one claimed by Bing. In fact, this experiment was repeated with a much larger sample of 32 million users and the results indicated a projected revenue increase of $10 million due to these simple color changes.

https://hbr.org/2017/09/the-surprising-power-of-online-experiments

Commonly Tested Website and Marketing Artifacts

- Titles and Headers
- Call to Action
- Forms
- Navigation
- Images
- Page Structure
- Website Landing Page
- Algorithms (e.g. purchase funnel models)

Designing your experiment - decide 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.

Example: exit polls to estimate the winner in an election

Example: testing a new cancer drug on a sample of patients before making it available as a treatment

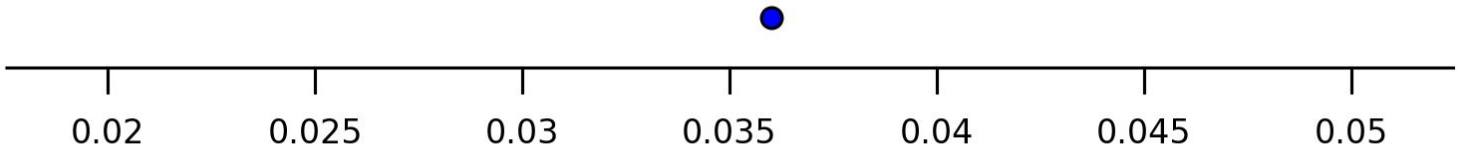
How big a sample you need depends on the size of the effect you are trying to measure.

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.

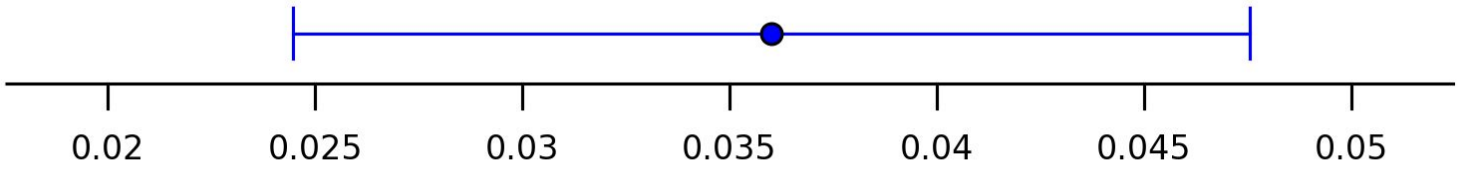
Based on our data, we can generate a **point estimate:** 36 / 1000 = 0.036.



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.

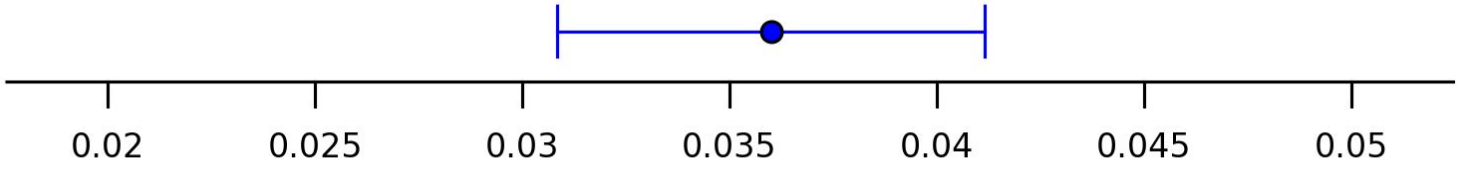
If the true rate is anywhere within the error bars, our observation would not be considered unusual.



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-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 that you did *or* a more extreme difference, if there is in fact no difference between A and B.

Hypothesis Testing

One possible type of error when conducting a hypothesis test is to reject the null hypothesis when the null hypothesis is actually true.

For A/B testing this means concluding that there is a difference in click-through rates when there really is no difference.

This type of error is called a **Type I Error.**

The significance level is the probability of a type I error if the null hypothesis is actually true.

Example

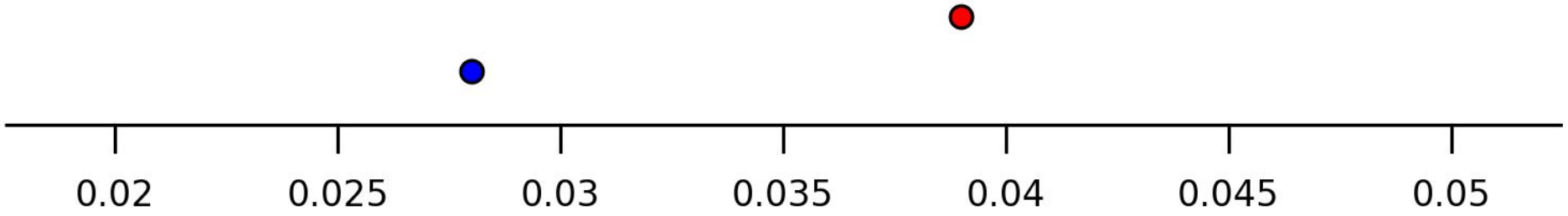
We run two versions of an ad and want to test if there is a statistically significant difference in the click-through rates. Here is the data we gather.

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

Example

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

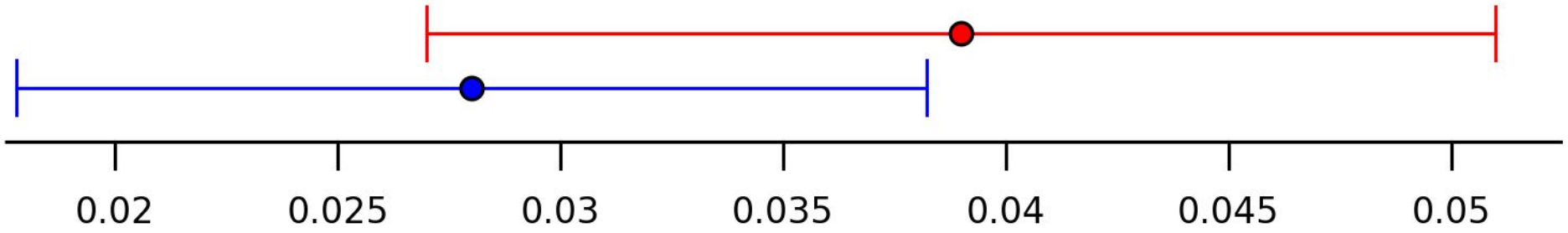
When comparing two options, we have to consider not just the point estimates, but also the margin of error.



Example

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

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



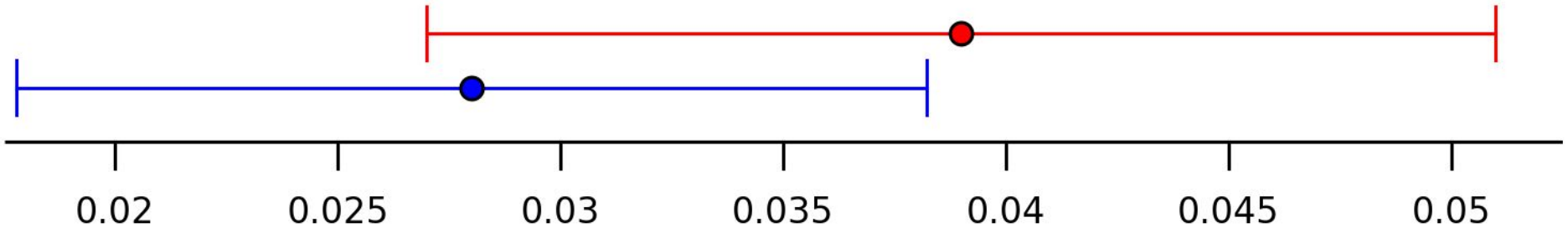
Example

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.



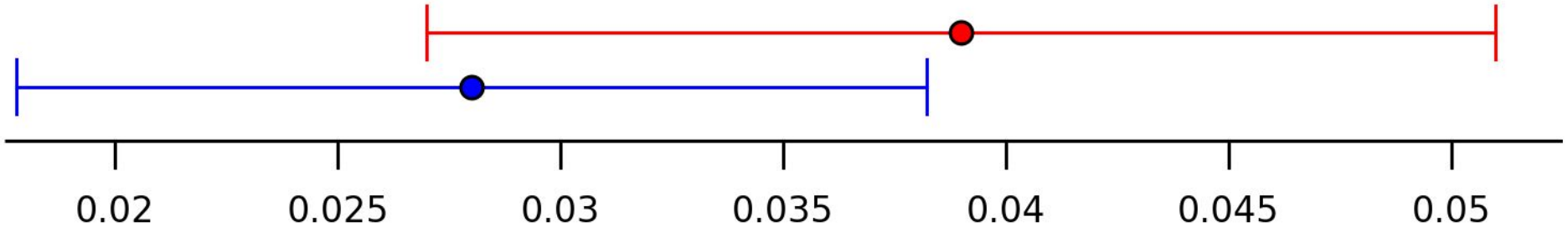
Example

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

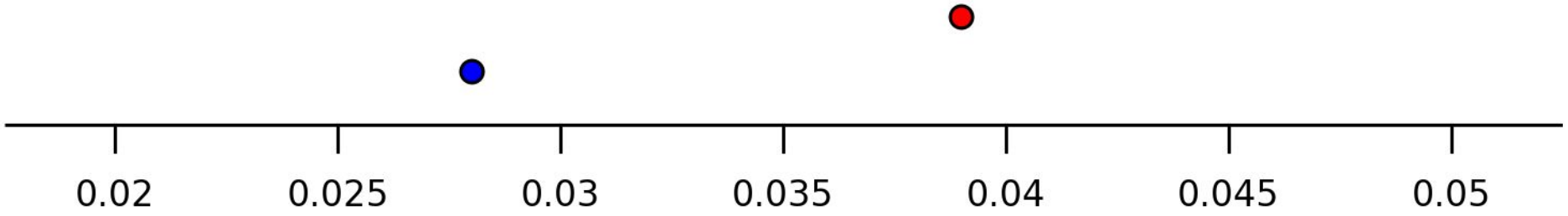
Now, we want to rerun out trial with a larger sample size. Here is the data we gather.

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

Example

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

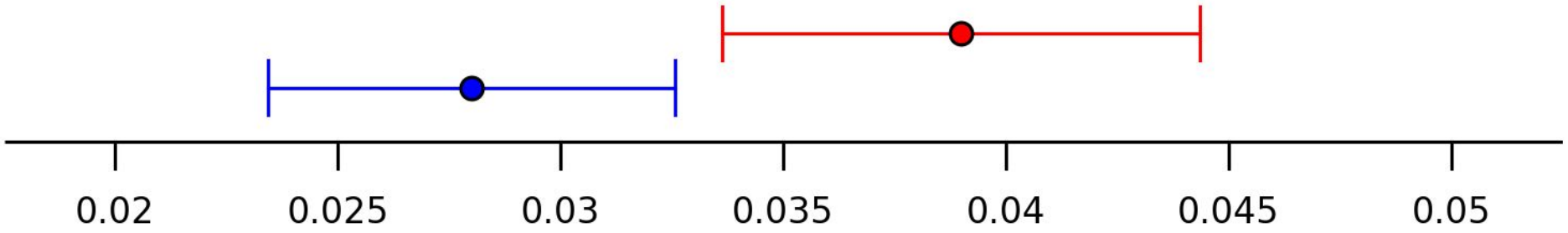
When comparing two options, we have to consider not just the point estimates, but also the margin of error.



Example

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

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



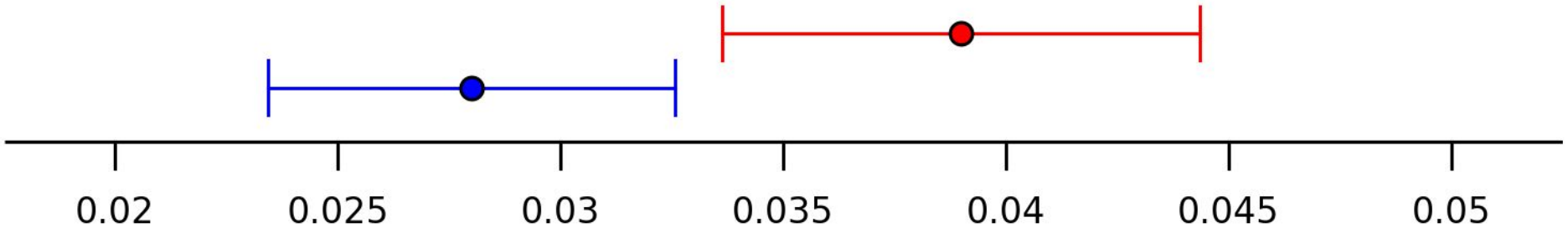
Example

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.



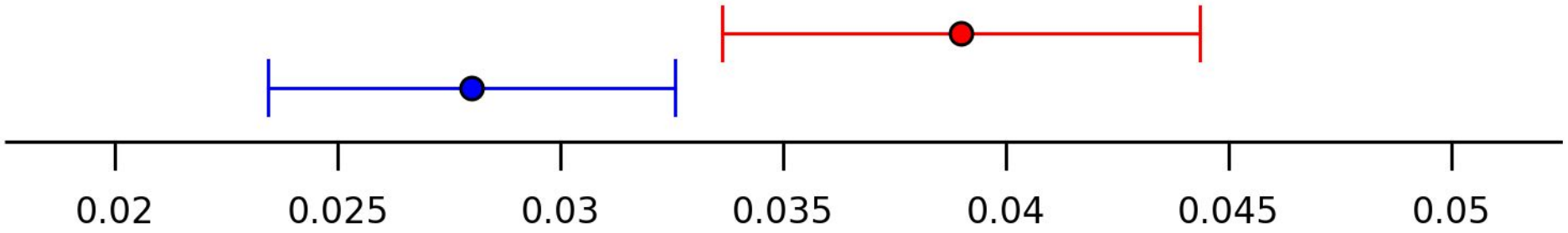
Example

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 reduce the uncertainty in our estimates. The observed difference would be very unlikely if in reality that was no difference in click-through rates.

Choosing the best statistical testing method
(traditional / frequentist)

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 sample sizes.
- 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, given that it is false, 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**.

Power

A second possible type of error when conducting a hypothesis test is to not reject the null hypothesis when the null hypothesis is false.

For A/B testing this means concluding that there is no difference in click-through rates when there really is a difference.

This type of error is called a **Type II Error.**

The power of the test is the probability of a type II error if the null hypothesis is false, given the expected base rate and effect size.

Power Example

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

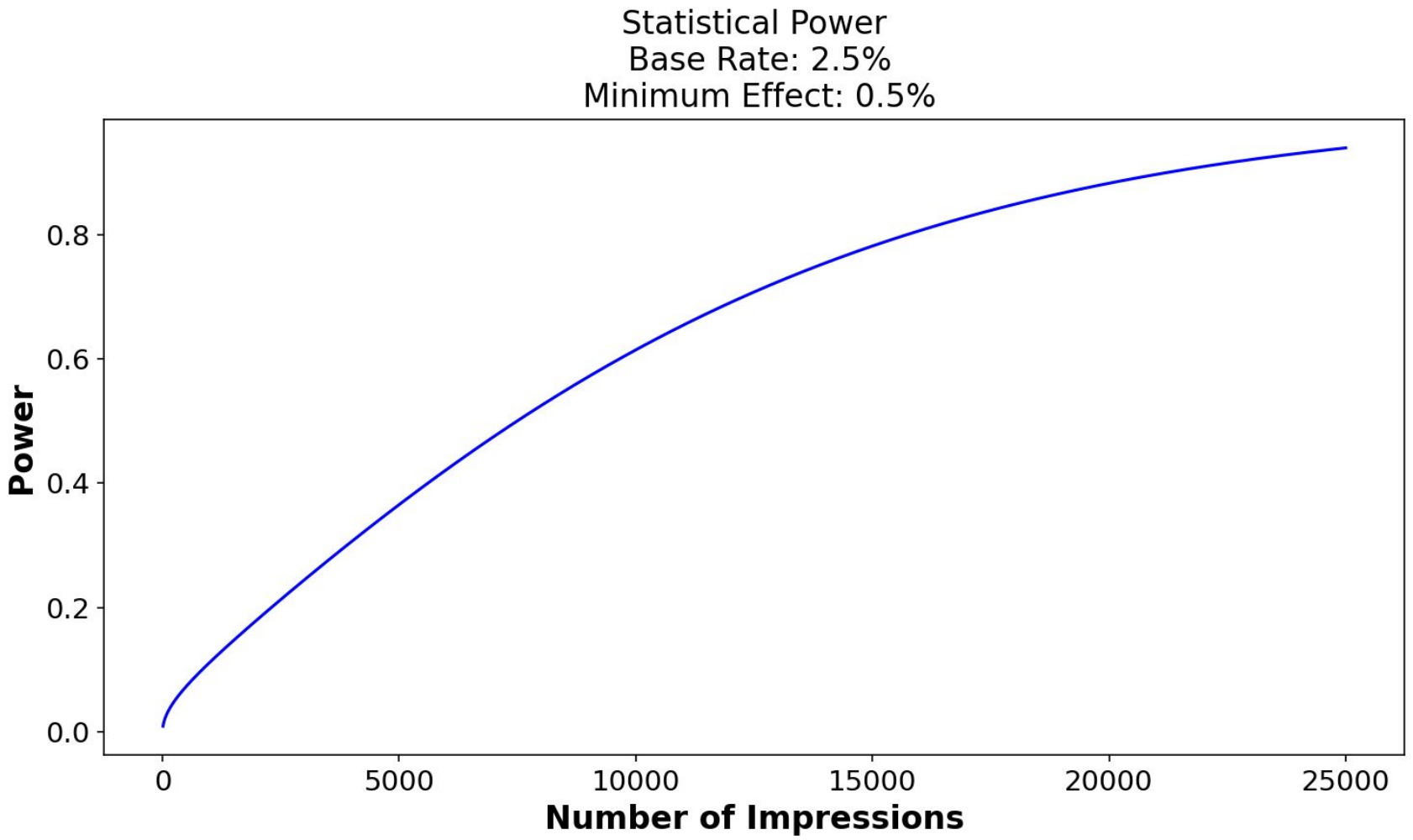
There is a new variant that you want to test and the minimal detectable effect you need to measure in order to make a change is 0.5%

How many impressions do you need in order for there be 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?

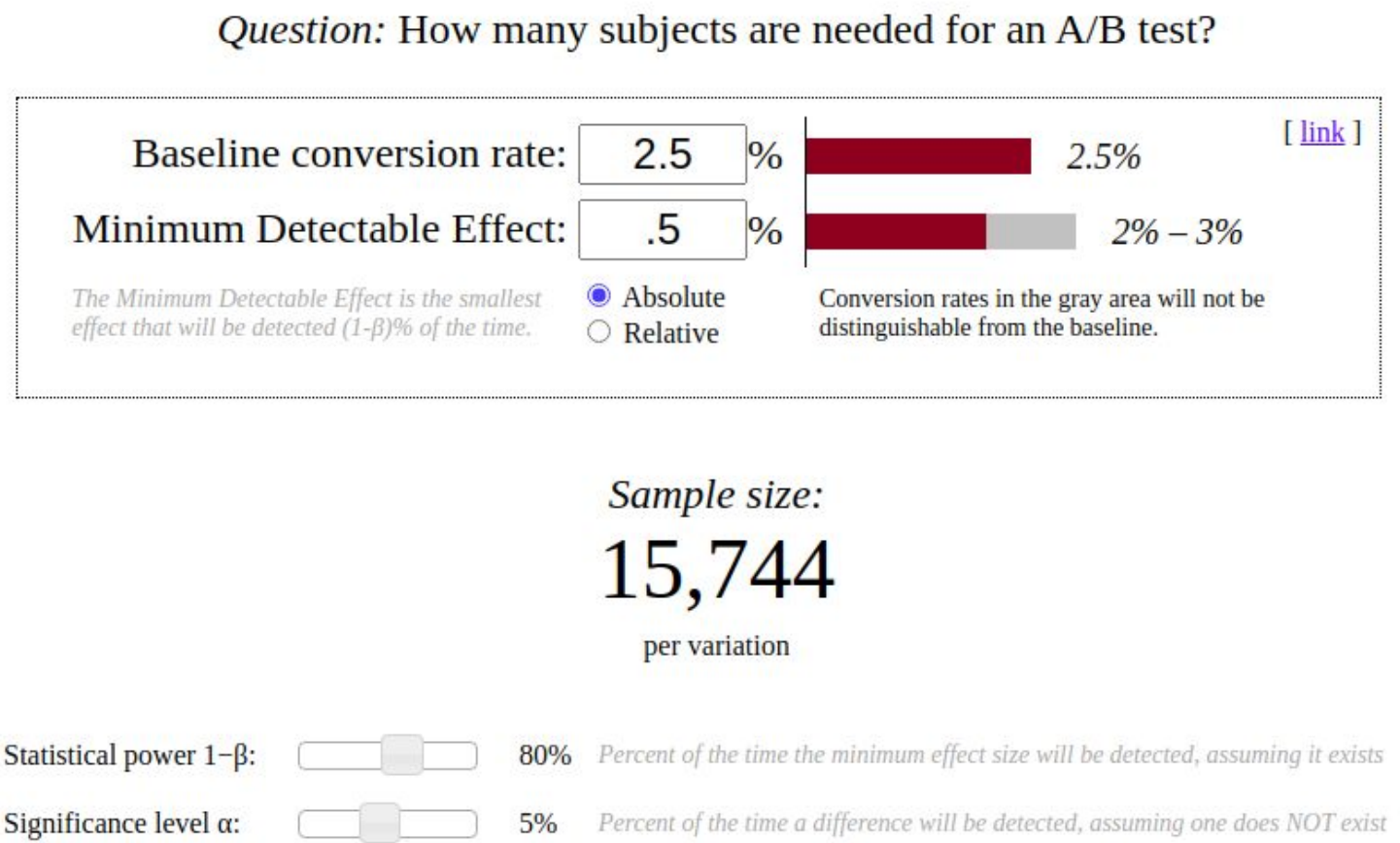
Power

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



Power

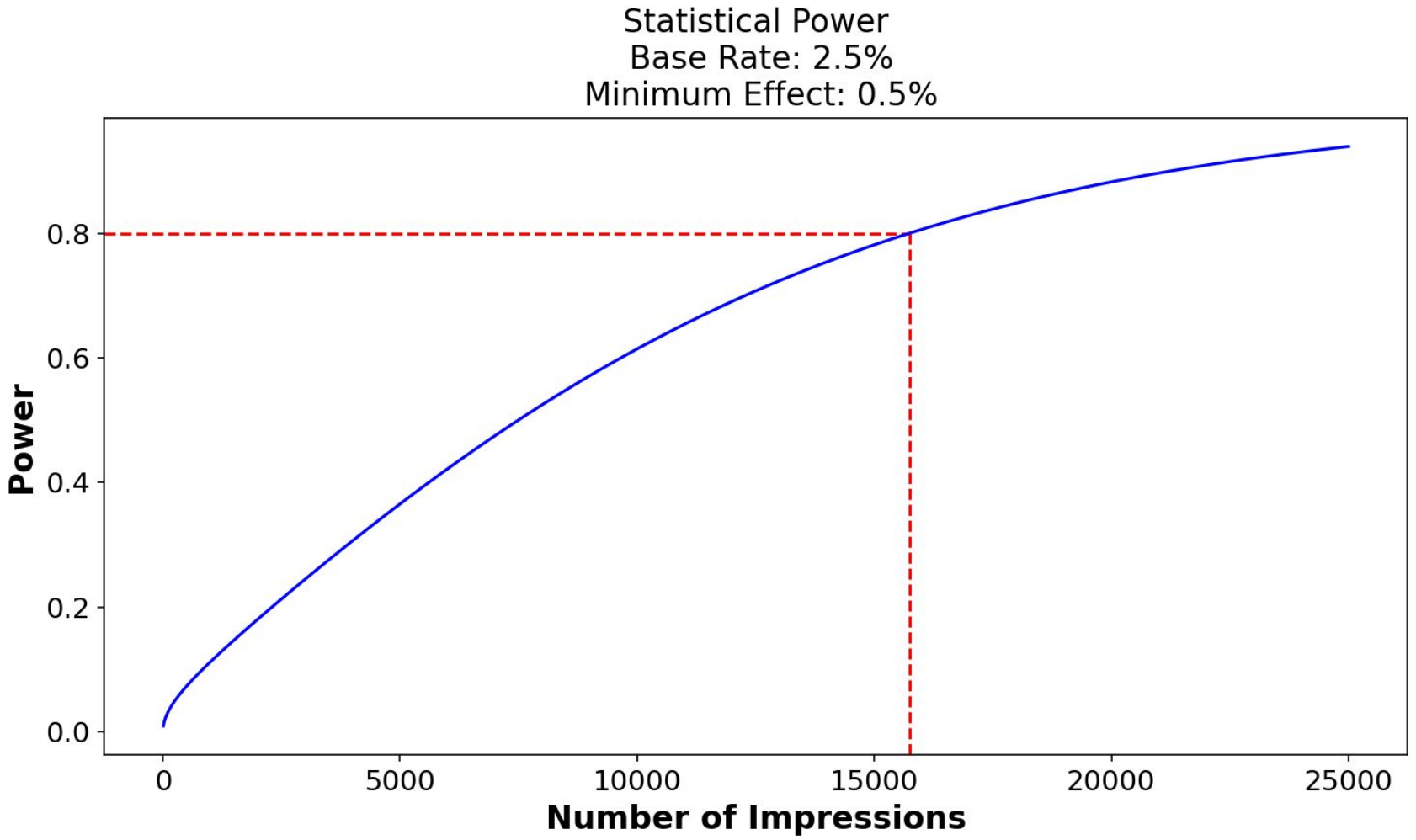
There are numerous sample size calculators available. For example, https://www.evanmiller.org/ab-testing/sample-size.html



NASHVILLE SOFTWARE SCHOOL

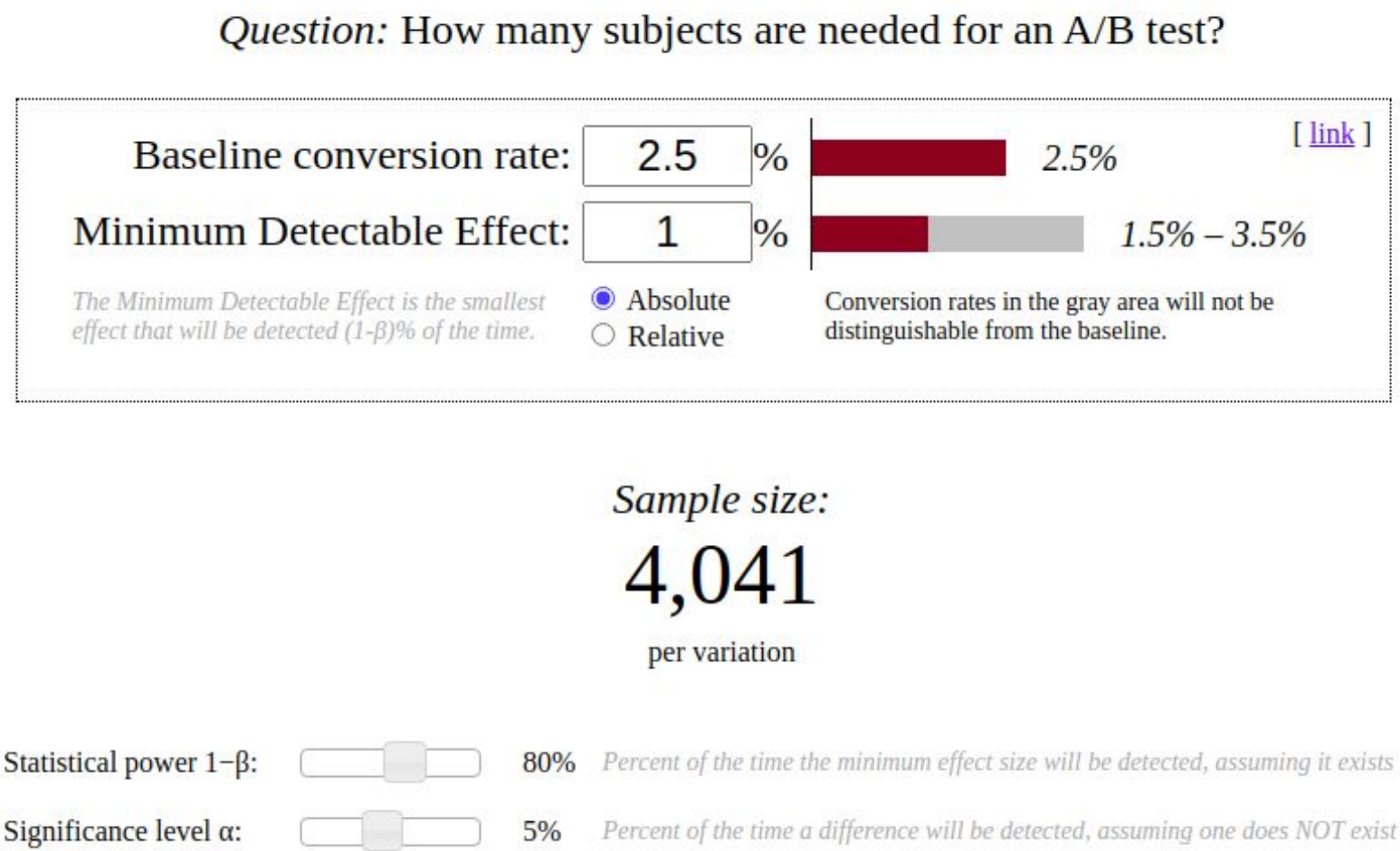
Power

We can see the sample size requirement on our plot.



Power

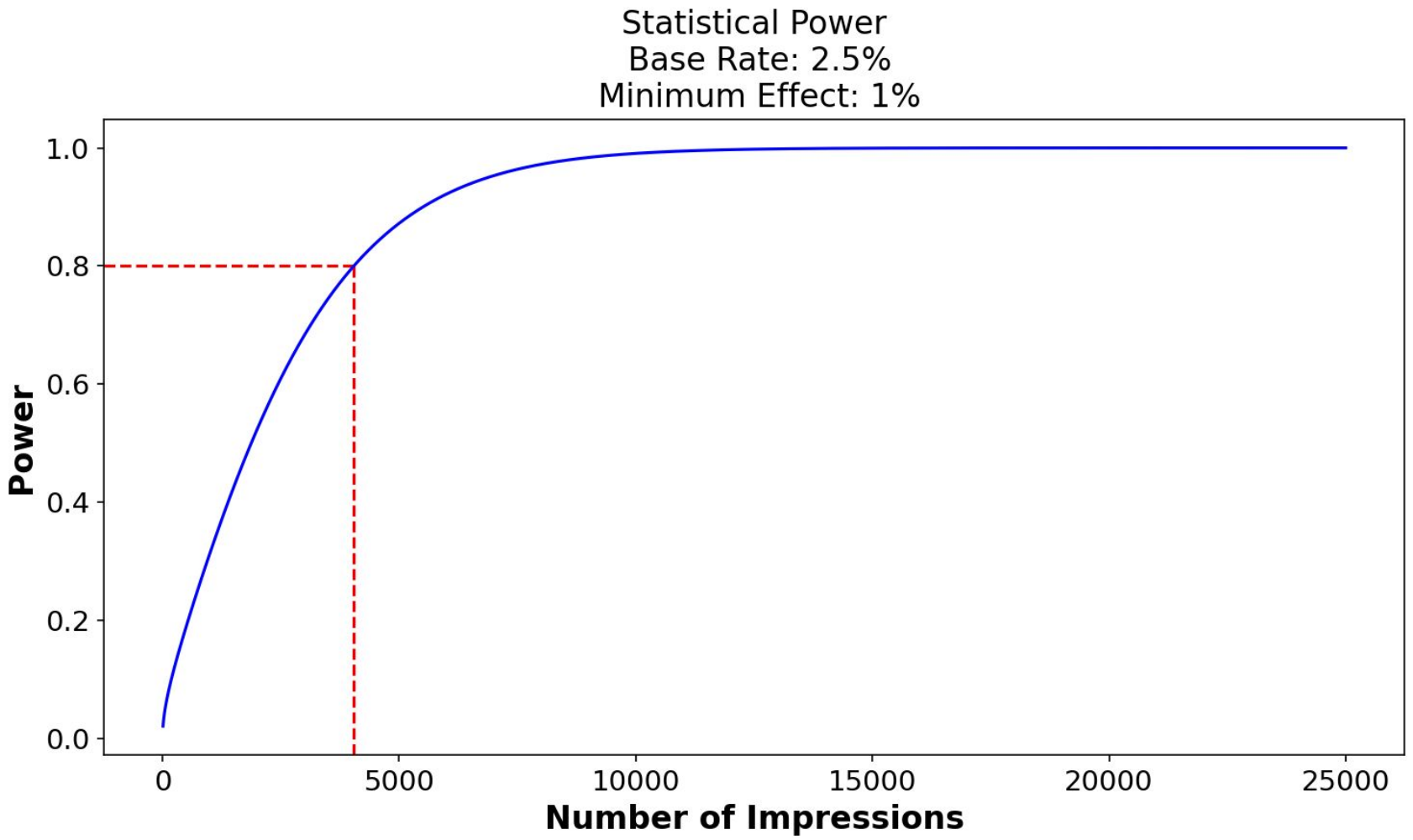
For larger effect, a smaller sample size is needed for the same amount of power.



NASHVILLE SOFTWARE SCHOOL

Power

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



NASHVILLE SOFTWARE SCHOOL

Choosing the best statistical testing method (Bayesian)

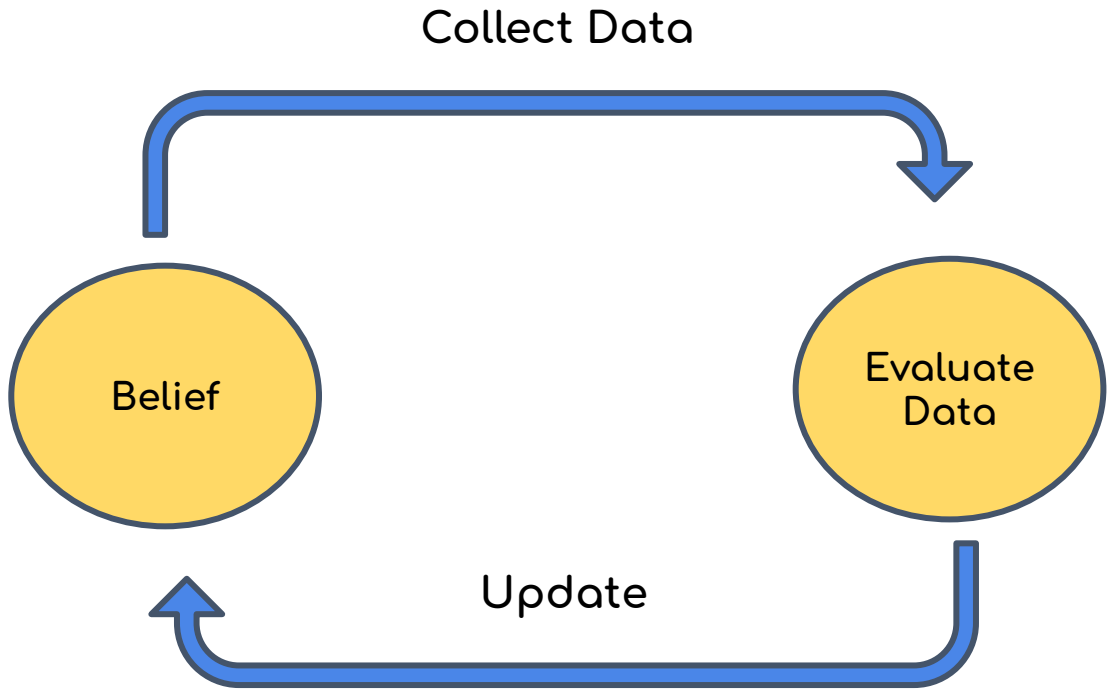
The frequentist approach has the disadvantage that you can't talk directly about the probability of A being better than B.

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

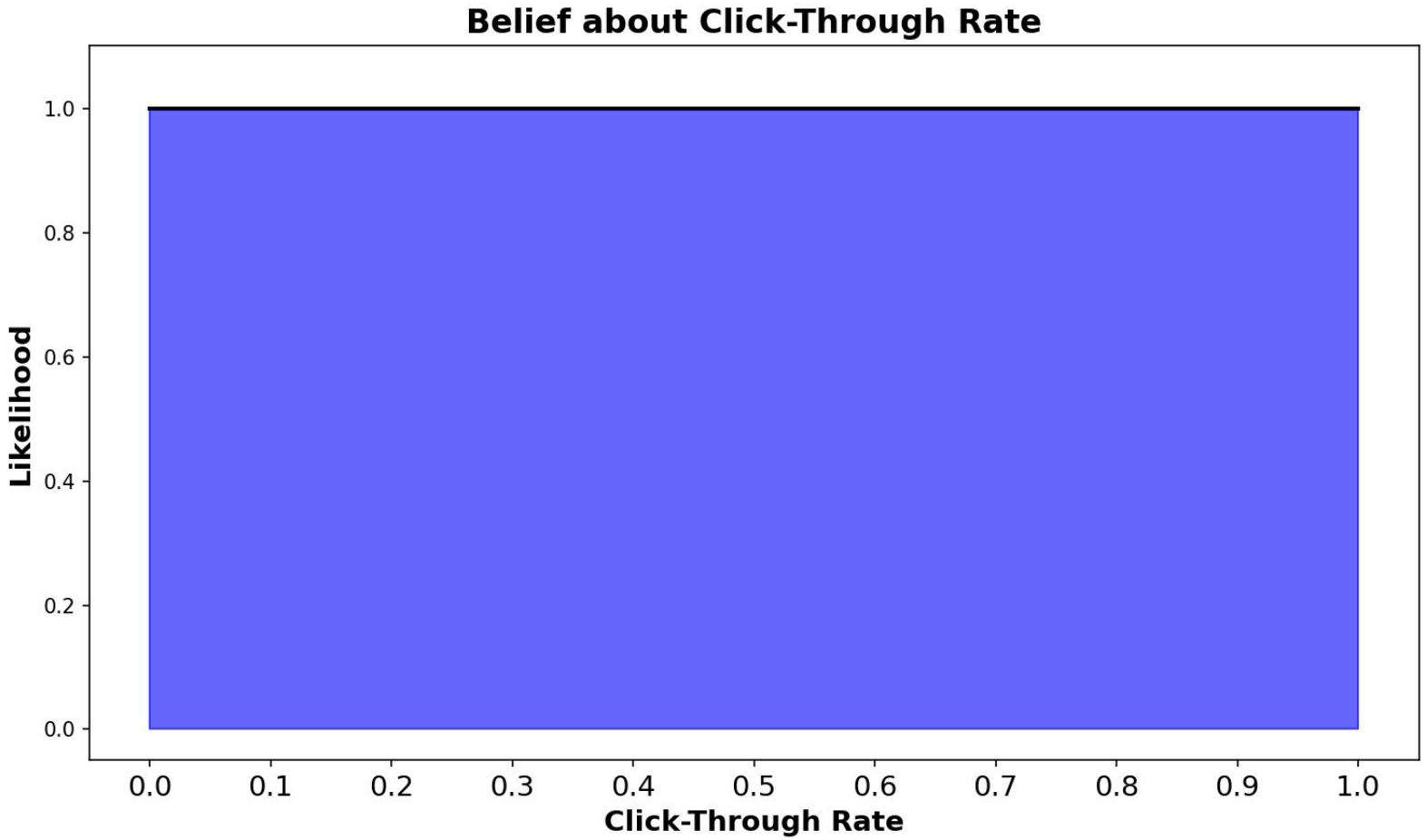
As you gather more data, you can continually update your belief.



Bayesian Example

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

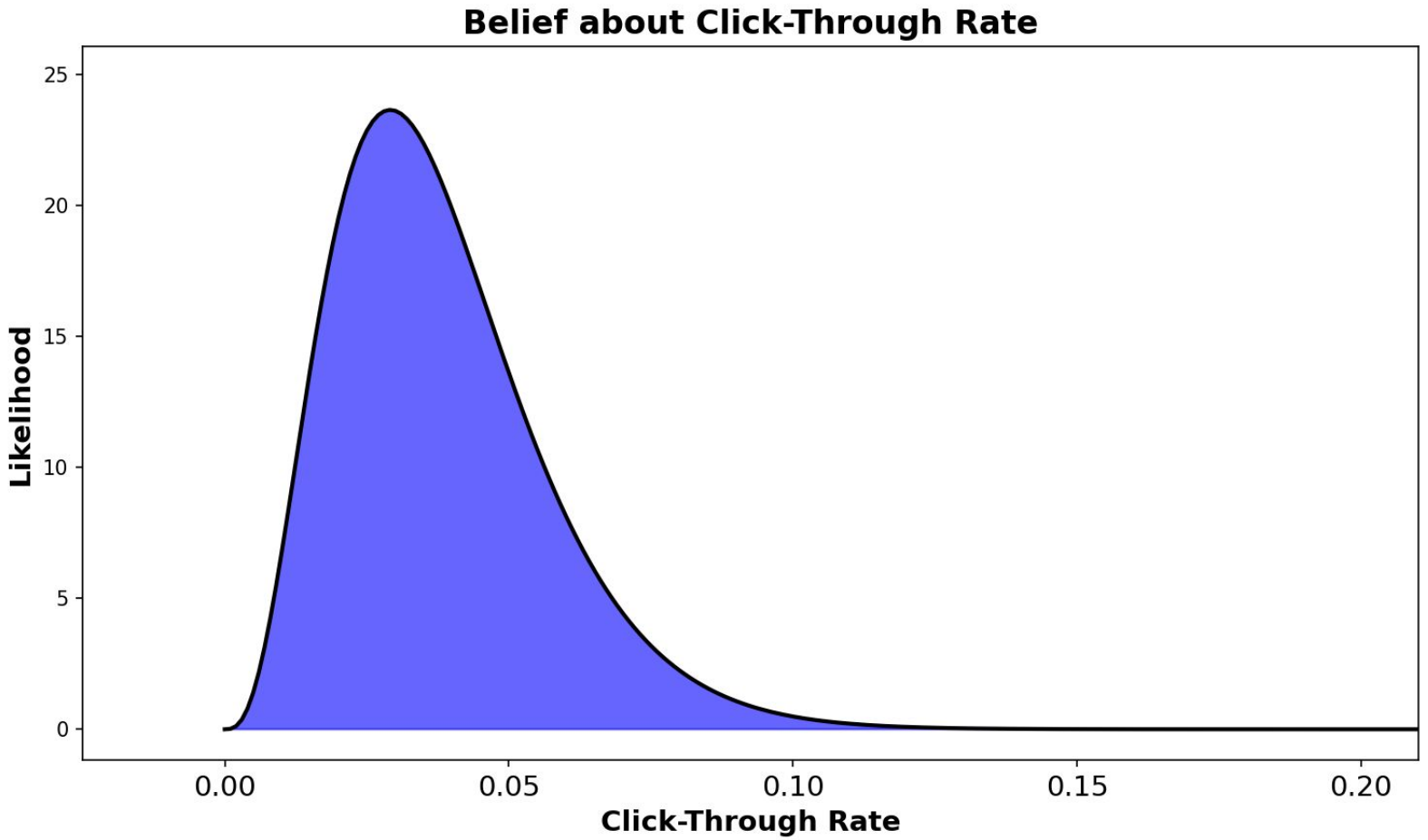
You want to start your estimation with no prior belief at all, meaning that all click-through rates are equally likely.



NASHVILLE SOFTWARE SCHOOL

Bayesian Example

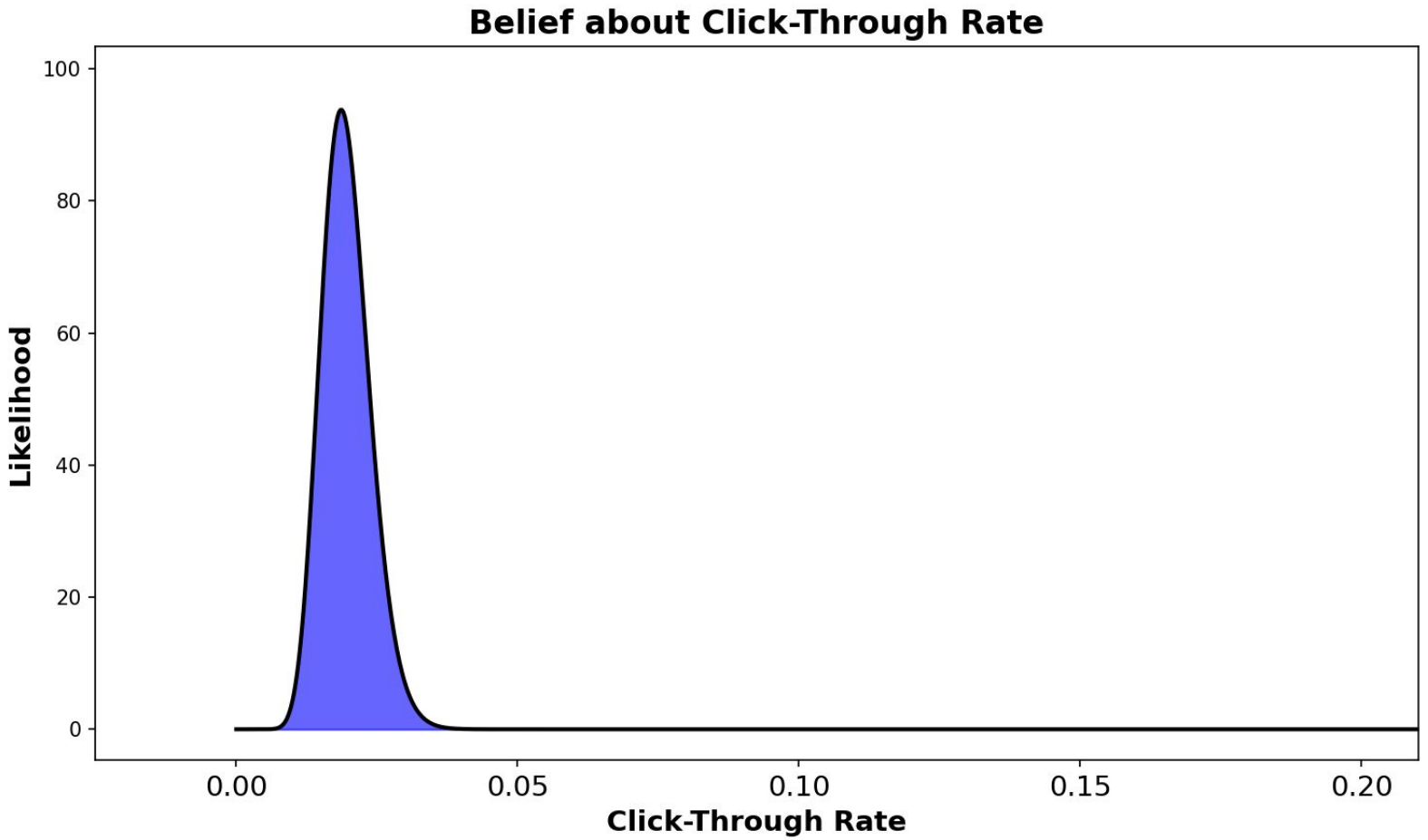
After 100 impressions, there are 3 clicks.
Based on this data, we can update our belief.



NASHVILLE SOFTWARE SCHOOL

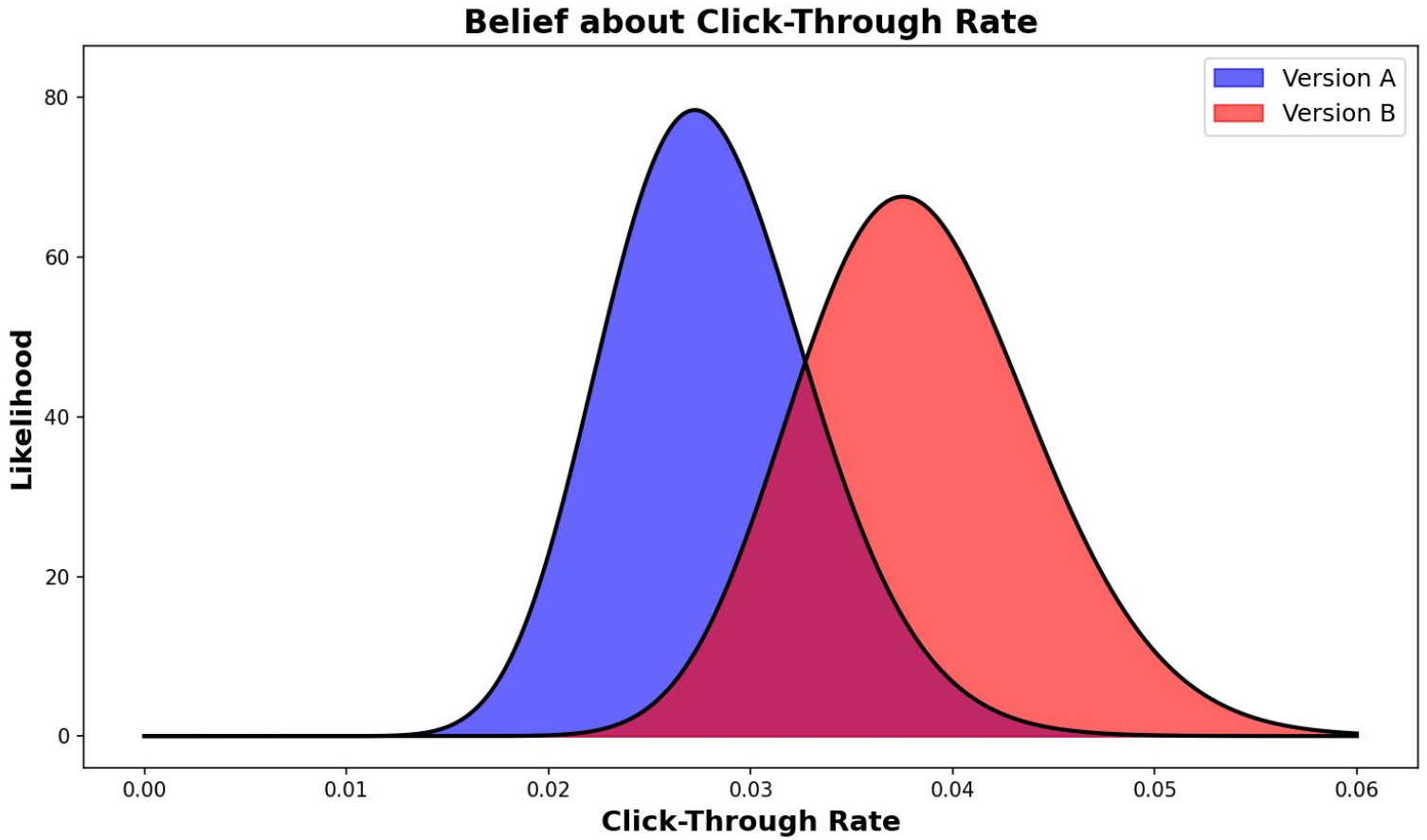
Bayesian Example

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

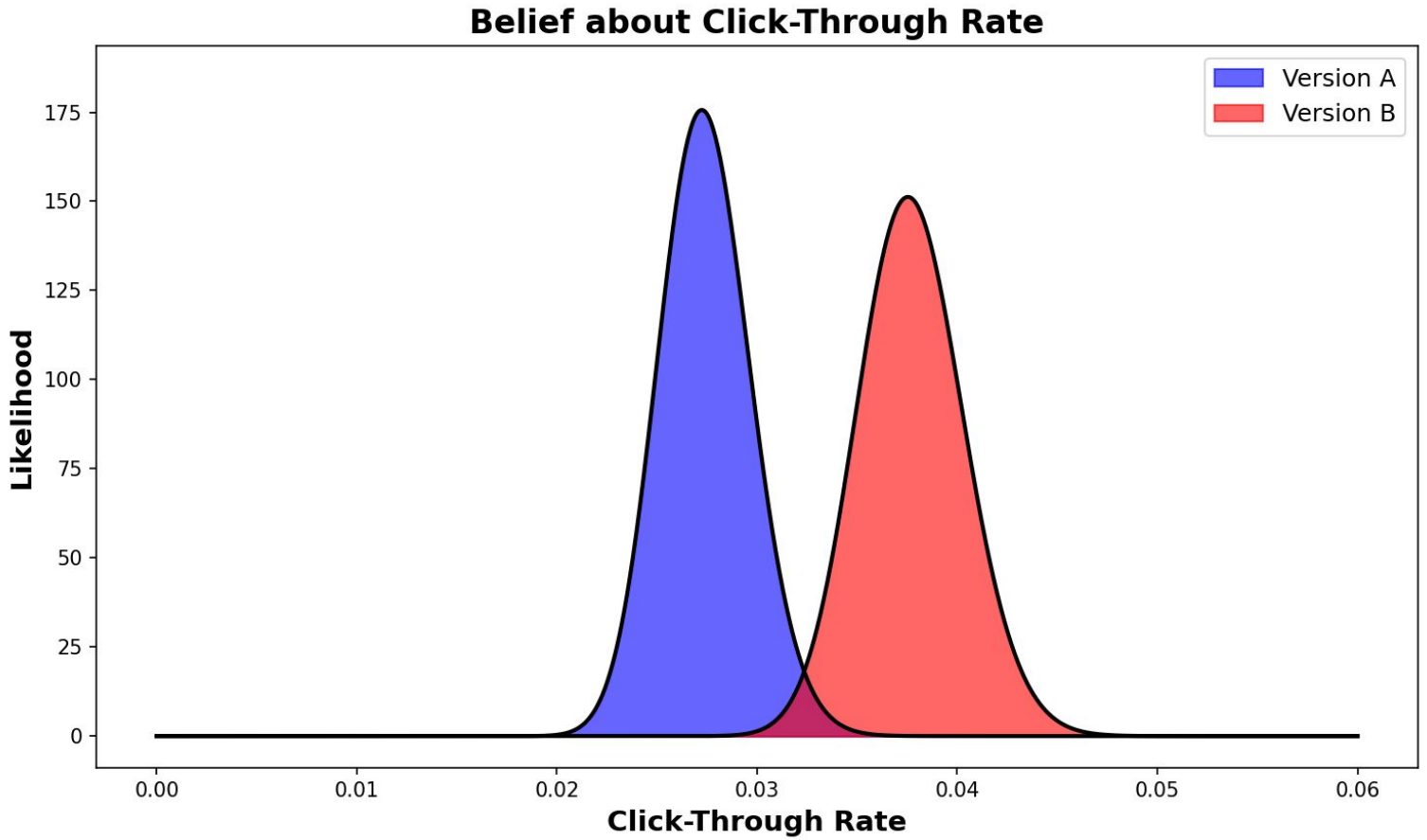
Probability that **B** has a higher click-through rate than **A**:

0.905

NASHVILLE SOFTWARE SCHOOL

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

NASHVILLE SOFTWARE SCHOOL

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?

Yes – story theme

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 he have a target outcome to measure?

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

Can he define the sample size?

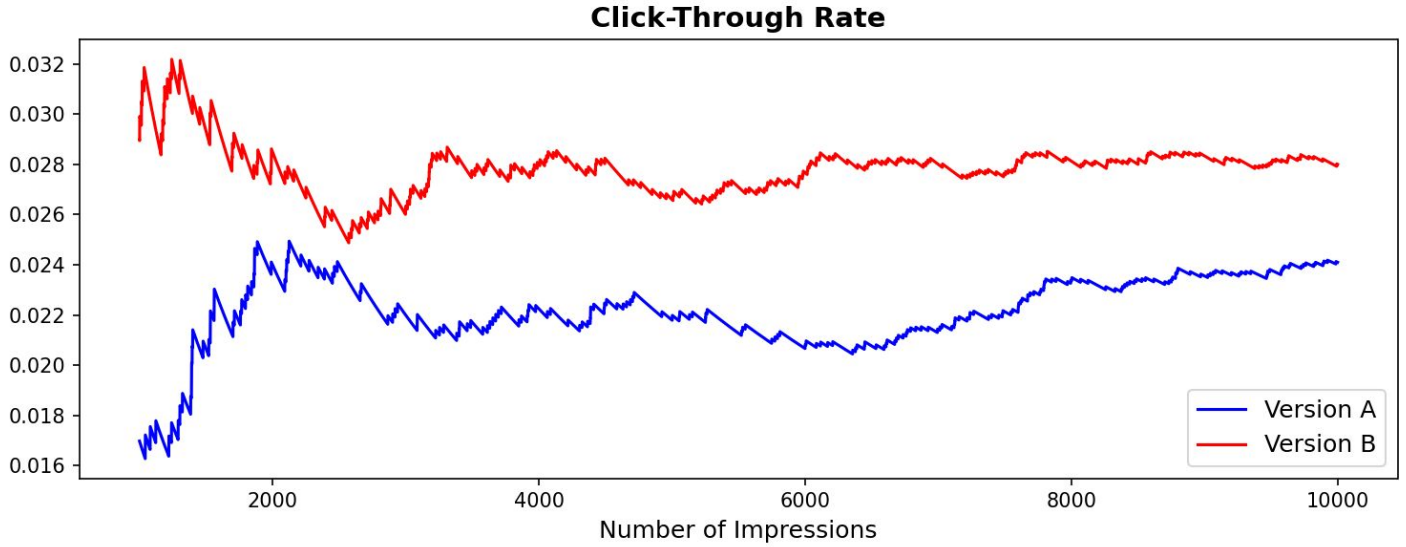
Yes – she can set the number of visitors to the How to Help page

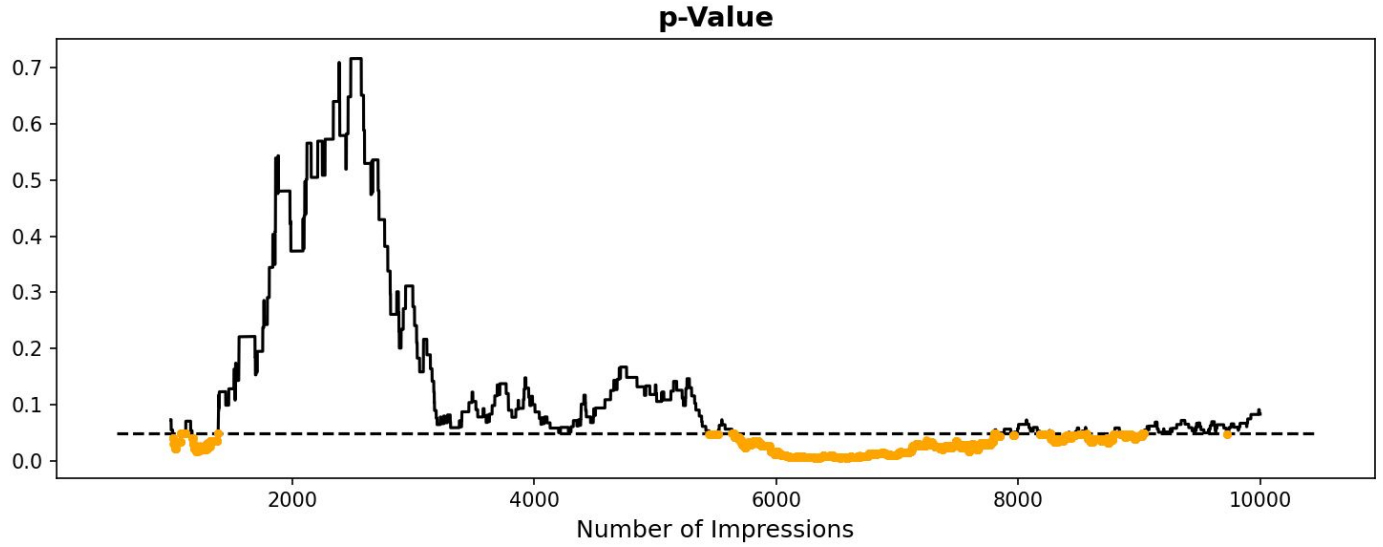
Important Considerations for A/B Testing

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

Do not run 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





NASHVILLE SOFTWARE SCHOOL

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

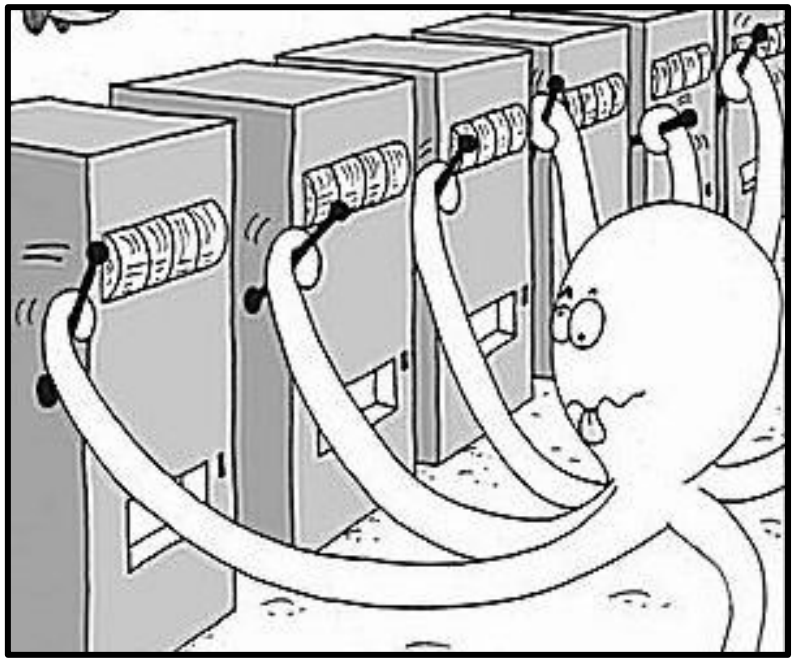
Requires fixed sample size for each variant from the outset.

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...

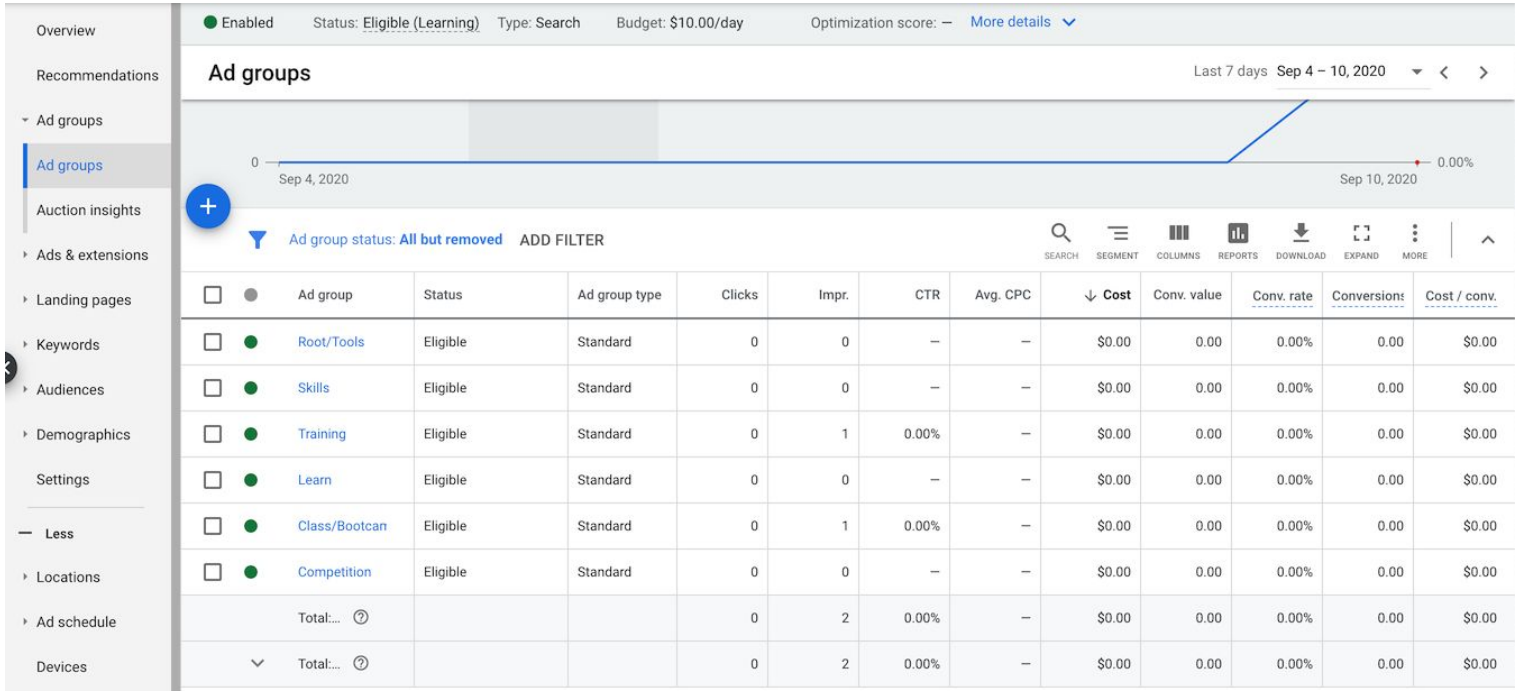
Example ad

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

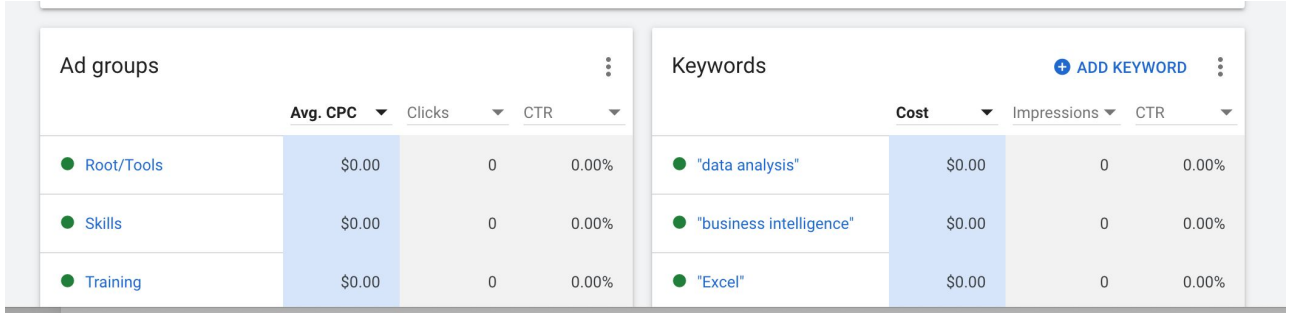
NASHVILLE SOFTWARE SCHOOL

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



NASHVILLE SOFTWARE SCHOOL

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