Introduction to Bayesian A/B testing

Johann de Boer

2022-10-22

Agenda

- Setting the scene
- Priors and probability distributions
- Running a Bayesian A/B test
- Posterior analysis
- When to stop a Bayesian A/B test?
- Summary and some final remarks
- Questions?
- Extras

Setting the scene

Randomised Control Trials (RCTs)

A simplistic example:

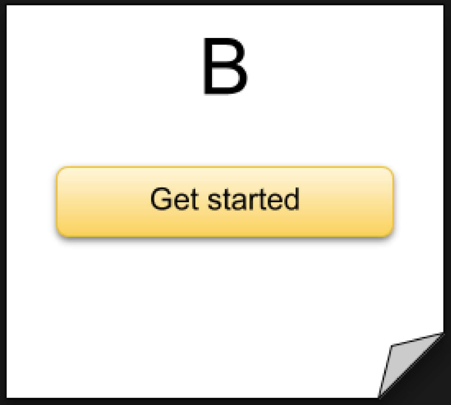
- Users are assigned at **random** to two groups, A and B, with equal probability.
- Let A be our **control** group and B be our **treatment** group.

We want to know what effect our treatment has.

Hypothetical scenario

- A button on a landing page that takes users to a sign up form.
- At present, the button is labelled “Register your interest”.
- Test whether changing it to “Get started” will result in an increased click-through rate (CTR).
- “Get started” was suggested by an experienced and skilled UX designer.

A
Register your interest



Priors and probability distributions

Prior knowledge and beliefs

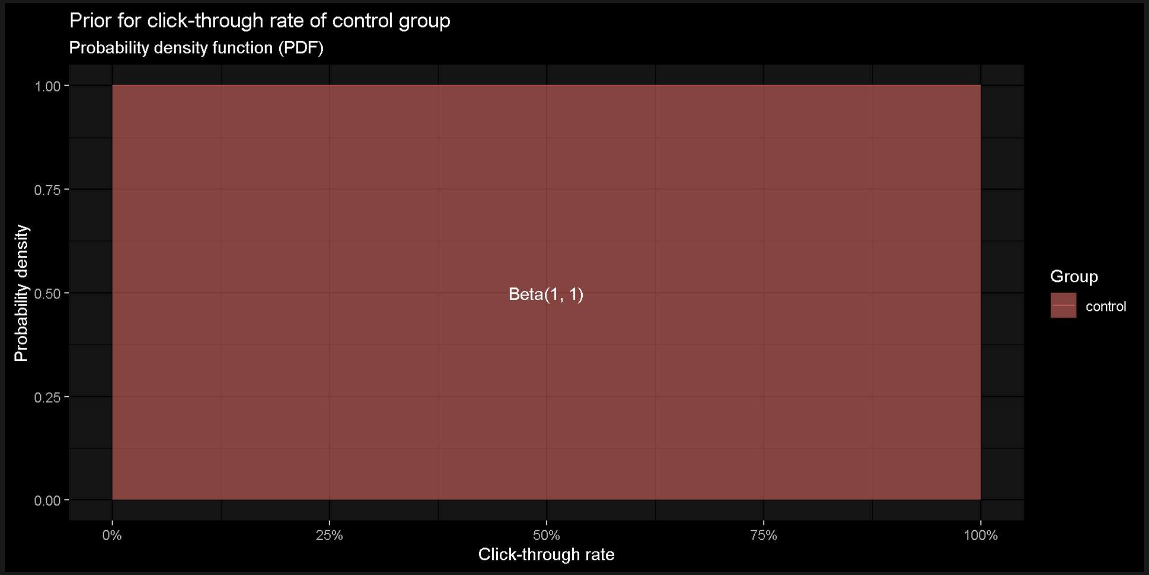
Before running an experiment, we form opinions and gather evidence such as:

- The baseline click-through rate of the button (with its current label) and knowledge of any outside variables that affects click-through rate, e.g. seasonality
- Effects we have seen from similar previous experiments
- Qualitative research, such as usability tests, focus groups, and surveys that are related to the test
- Opinions (including critical) from stakeholders and experts

Priors are probability distributions

Express prior beliefs about the click-through rate of the control group using a **probability distribution**.

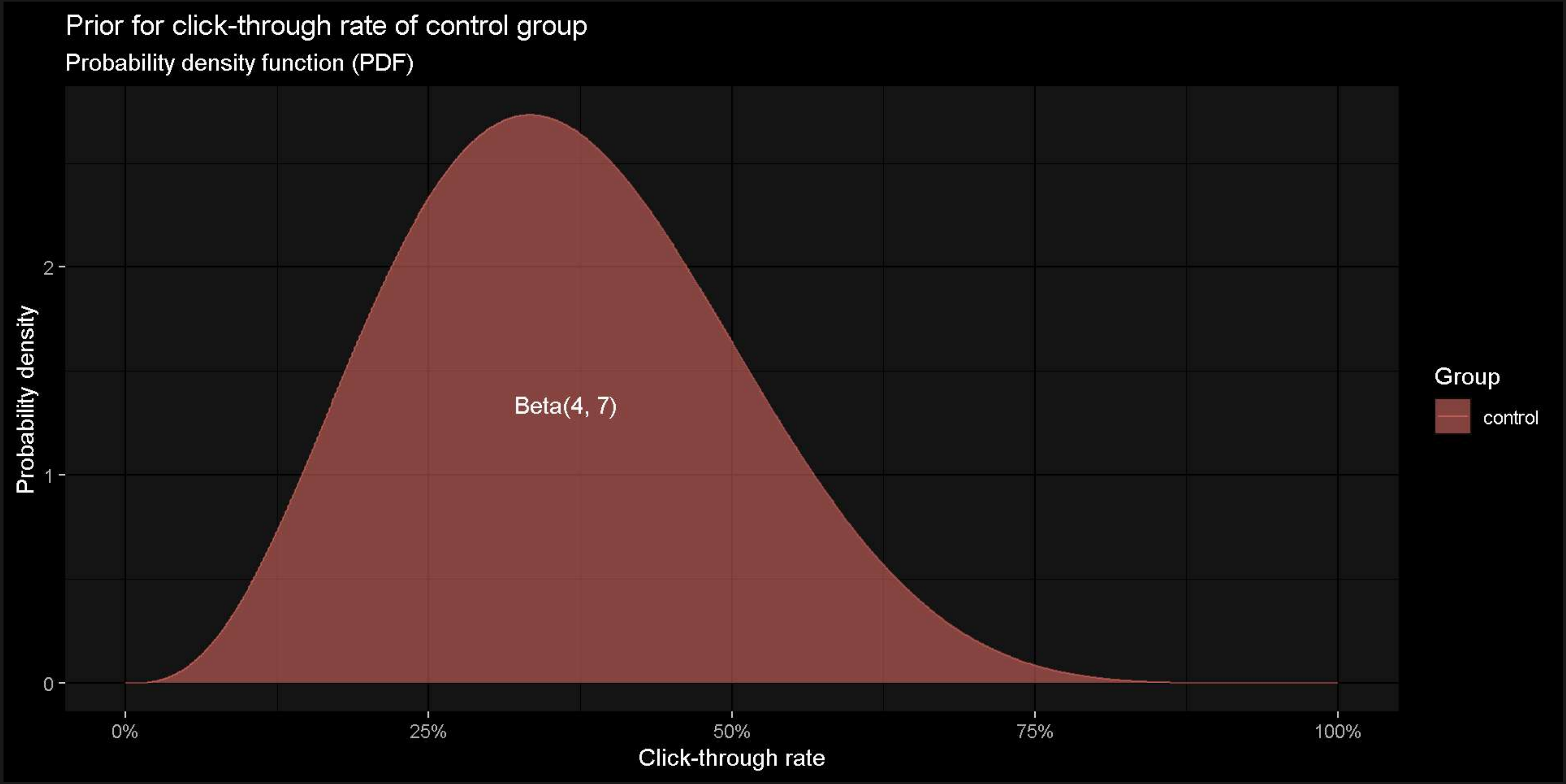
Here's an example of an extremely uninformative prior – a uniform prior that says any range of click-through rate is as probable as any other equally wide range, i.e. naive.



Tip

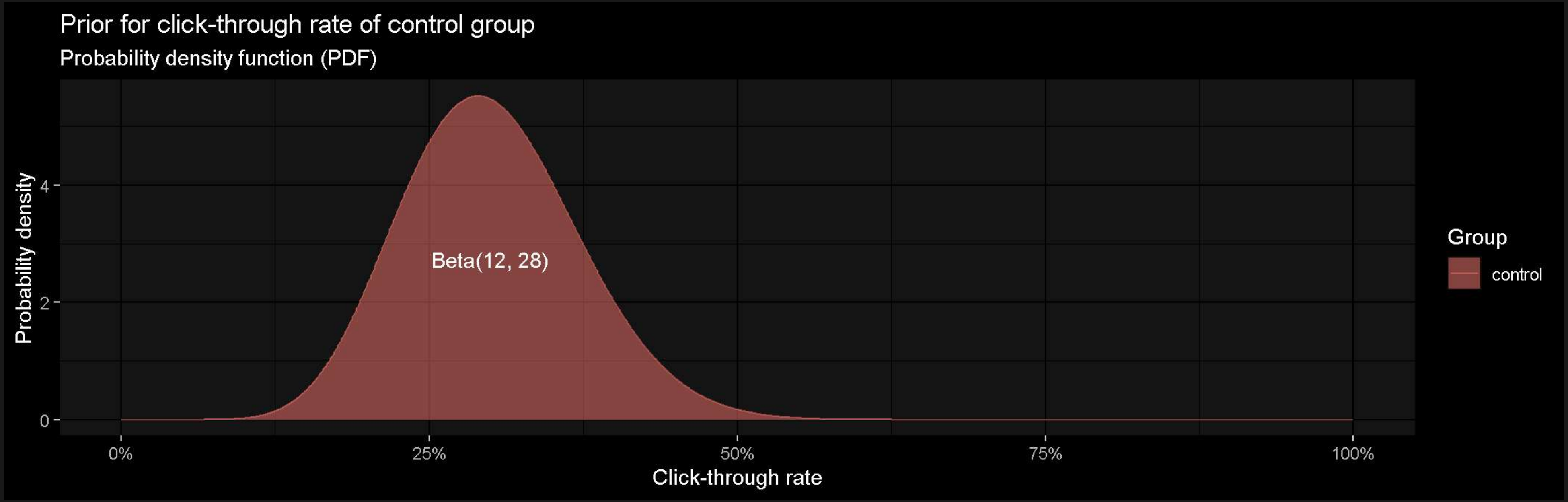
The **Beta distribution** is a **probability density function (PDF)** with two **shape parameters:** $B(shape1, shape2)$. It's used to describe proportions, such as click-through rate.

Something a little more informative



As the curve narrows, notice that the shape parameters of the Beta distribution increase.

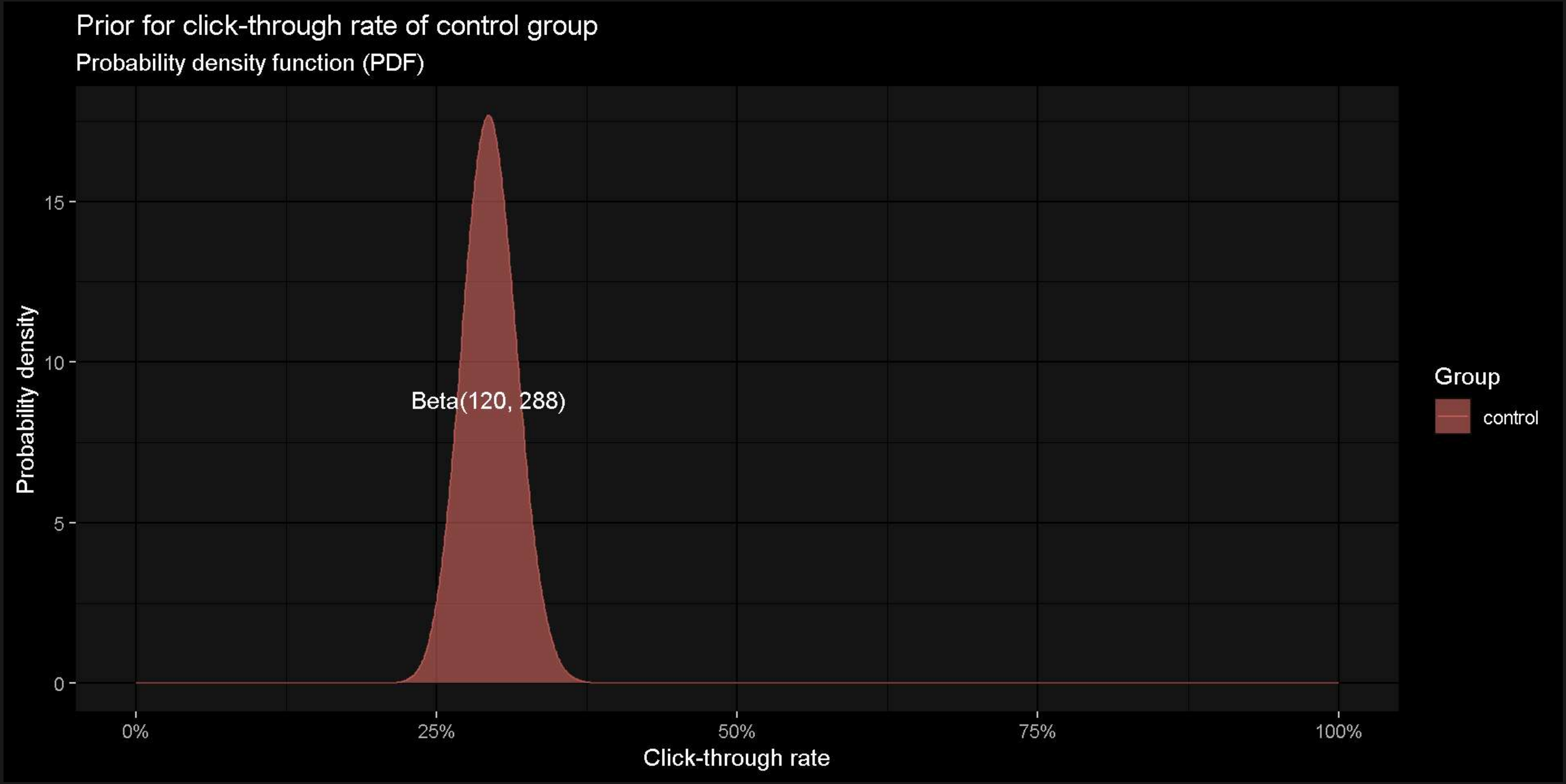
Something even more informative



The shape parameters (shape1 and shape2) of the Beta distribution can be considered counts of **successes** and **failures**, respectively. The mean probability of success (i.e. average click-through rate) can be calculated by this formula:

$$\frac{shape1}{shape1 + shape2}$$

Let’s say we’ve settled on this:

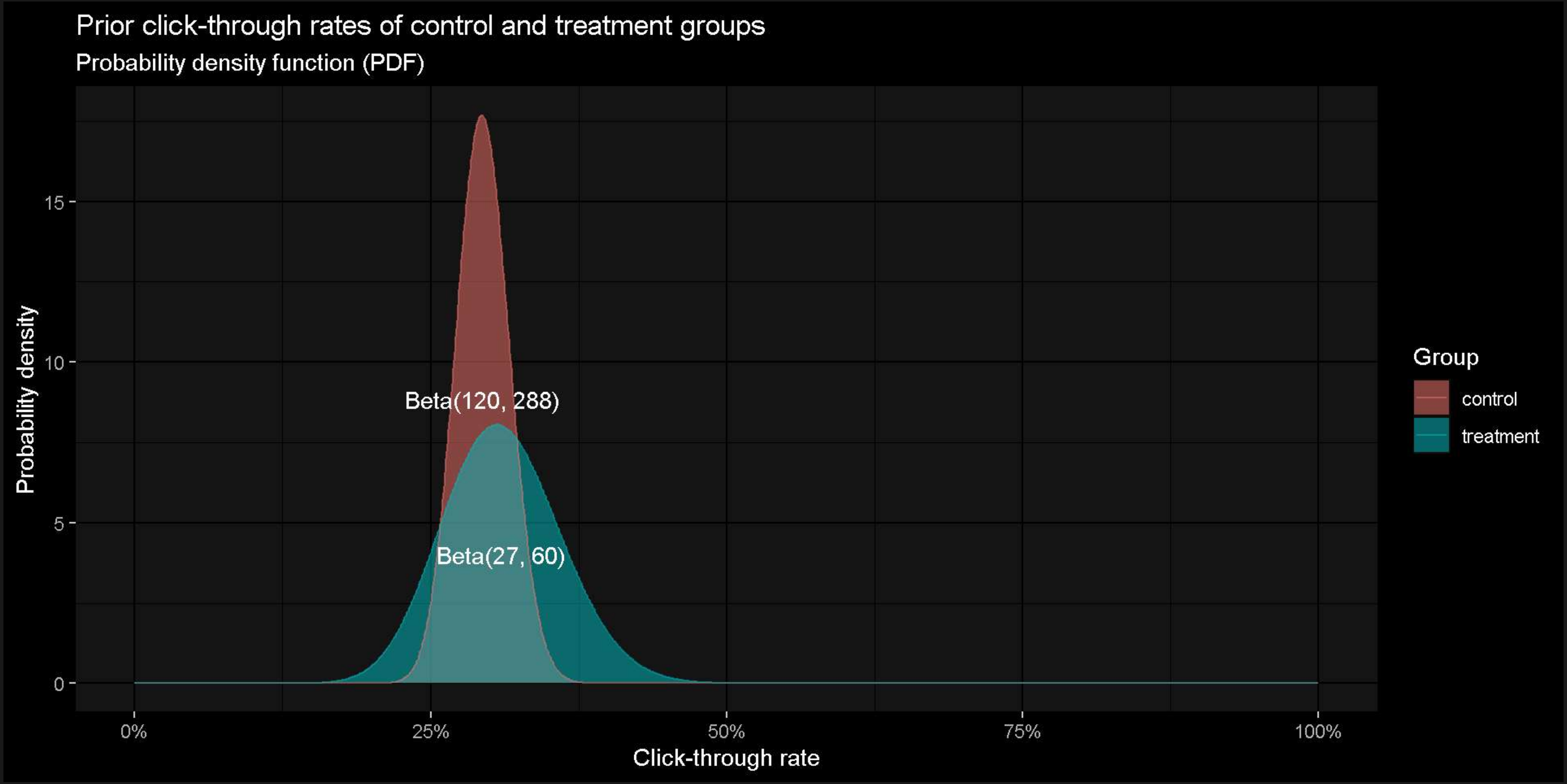


The more confident we are about our beliefs, the narrower the curve.

What about the treatment group?

- We expect that the click-through rates of the treatment and control groups will be correlated.
- We're unsure about how correlated they will be, but we're not expecting a dramatic difference.
- We're more confident than not that the treatment will be an improvement, but we're open to other possibilities.
- We don't want to bias the experiment results in favour of treatment or control, or towards a conclusion of there being a difference or no difference.

We’ve settled on these priors:



Running a Bayesian A/B test

Prior agreement

- Agreement must be reached on the priors before collecting and analysing data from the experiment.
- Once the priors are agreed to and locked in, we can start the experiment.

Here’s a summary of the priors we have chosen:

group	shape1	shape2	mean	sd
control	120	288	29.4%	2.3 p.p.
treatment	27	60	31.0%	4.9 p.p.

Let’s run an experiment

We’ll generate some fake data to mimic a real experiment.

It’ll be rigged though, as we’ll already know the click-through rates for control and treatment, which are:

- Control: 32%
- Treatment: 35%

That’s a relative uplift of 9.4%.

If we’re successful at applying Bayesian inference then we should hope (but can’t guarantee due to randomness) that the results somewhat match with these expected CTRs.

The next day

Let’s pretend that on average 150 users enter our experiment each day, and we’ve received the following data from day 1:

group	total_users	clicked	not_clicked	CTR
control	84	25	59	29.76%
treatment	62	18	44	29.03%

Let’s now incorporate our priors

Posteriors represent your updated beliefs once you’ve incorporated experiment data with your priors. Like priors, posteriors represent your beliefs about the metric of interest, which in our case is click-through rate.

For each experiment group, we derive our posterior shape parameters through simple arithmetic addition:

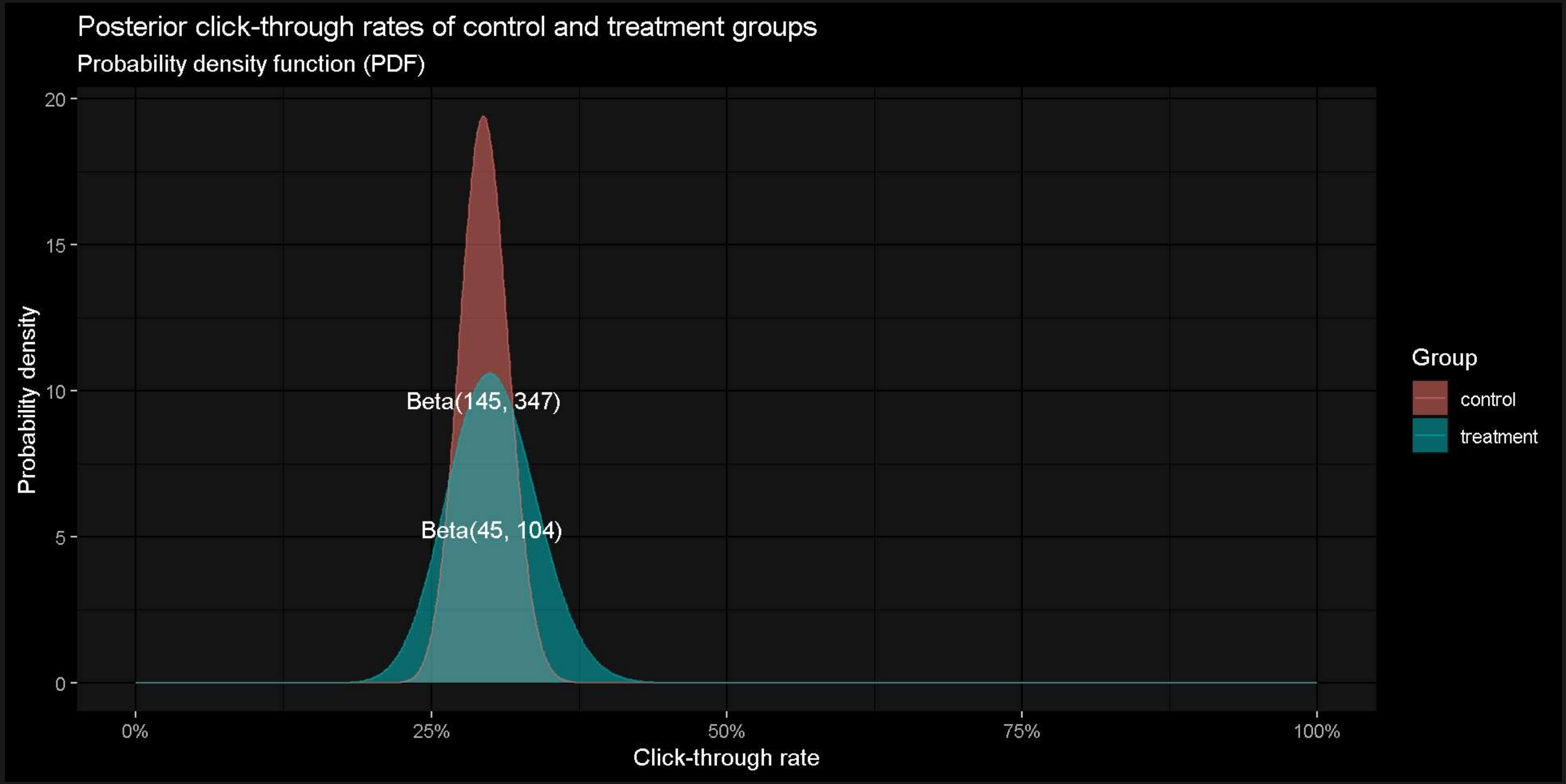
- Increment the first shape parameter by the count users who had **clicked**
- Increment the second shape parameter by the count users who **didn’t click**

count	**control**	**treatment**
prior_shape1	120	27
clicked	25	18
posterior_shape1	145	45

count	**control**	**treatment**
prior_shape2	288	60
not_clicked	59	44
posterior_shape2	347	104

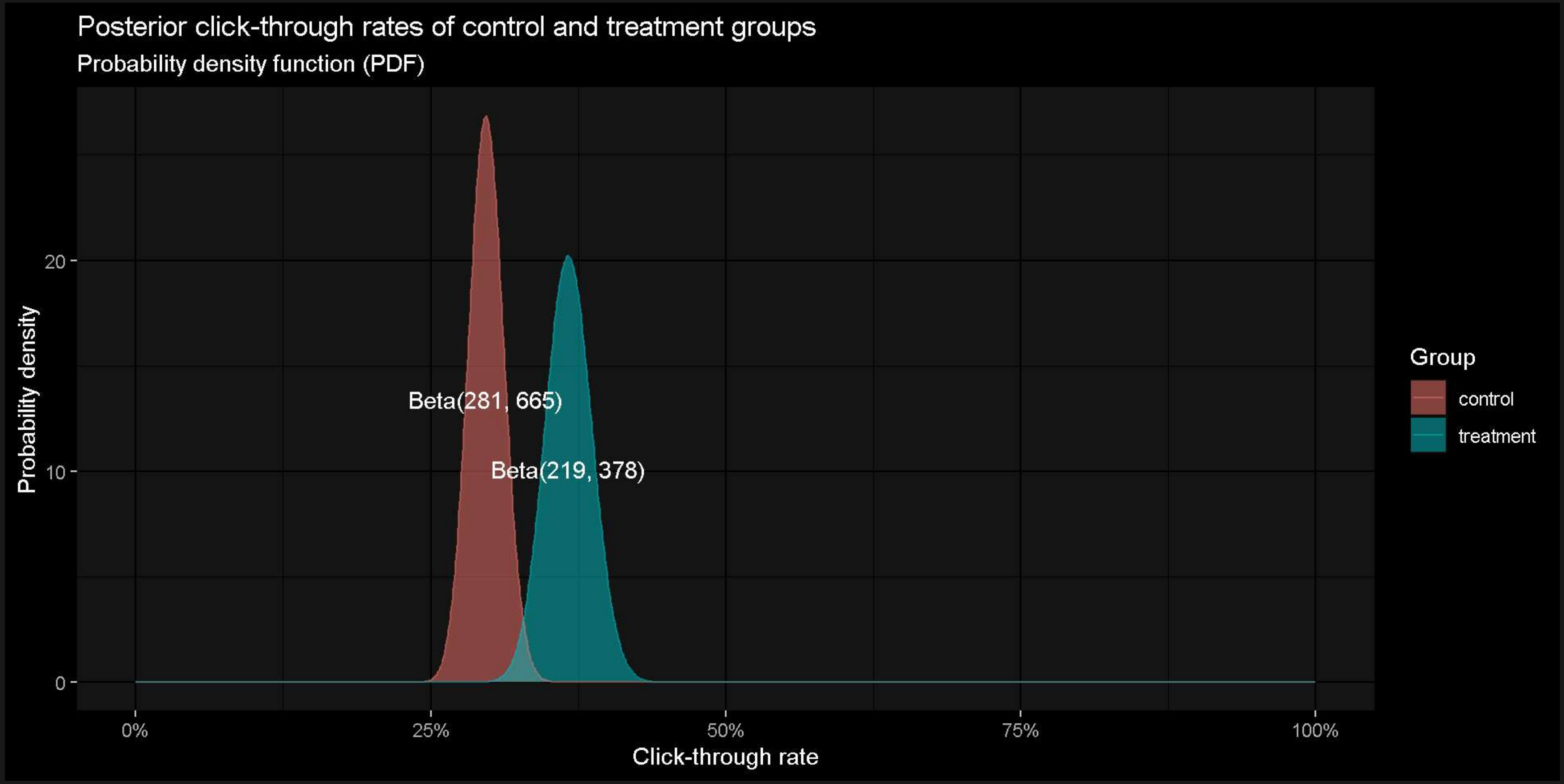
Posterior distribution of each group

We have now updated our beliefs. These posteriors can now be thought of as our new updated priors.



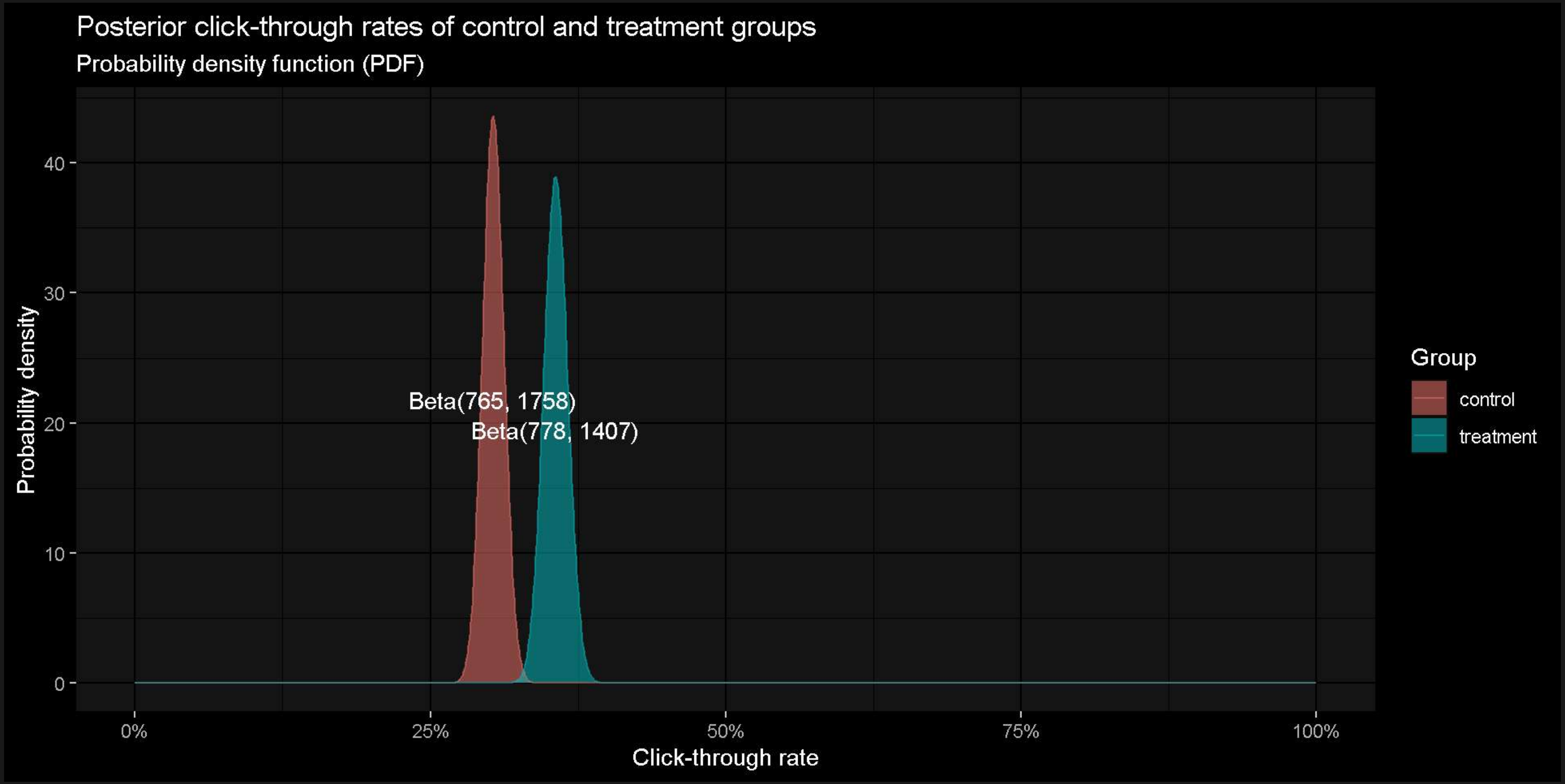
Another six days later…

We've collected more data, so let's again update our priors to form new posteriors for the click-through rates of each group.



Another three weeks later…

We’ve now observed a total sample of 4,213 users and a decision is made to end the experiment.



Posterior analysis

Statistical inferences using the posterior distributions

Monte Carlo simulation

Let’s draw a very large quantity of random samples from our posterior distributions to make inferences about the experiment.

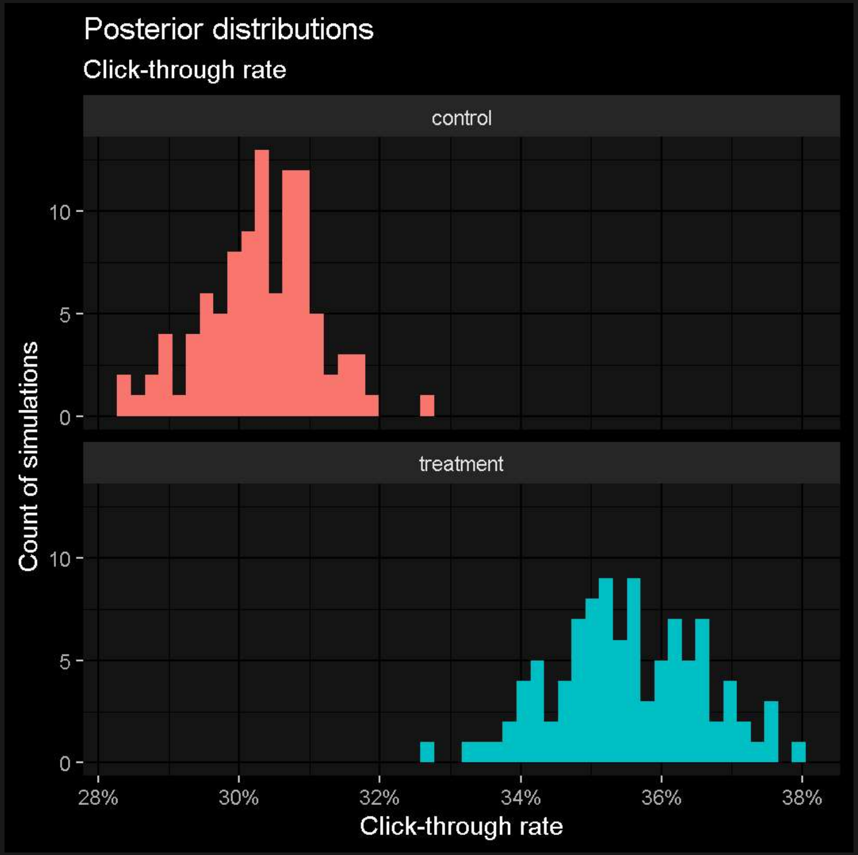
This is called Monte Carlo simulation – named after a casino.

100 simulations

Let's start slowly by drawing 100 random samples from our distributions and plot them using histograms…

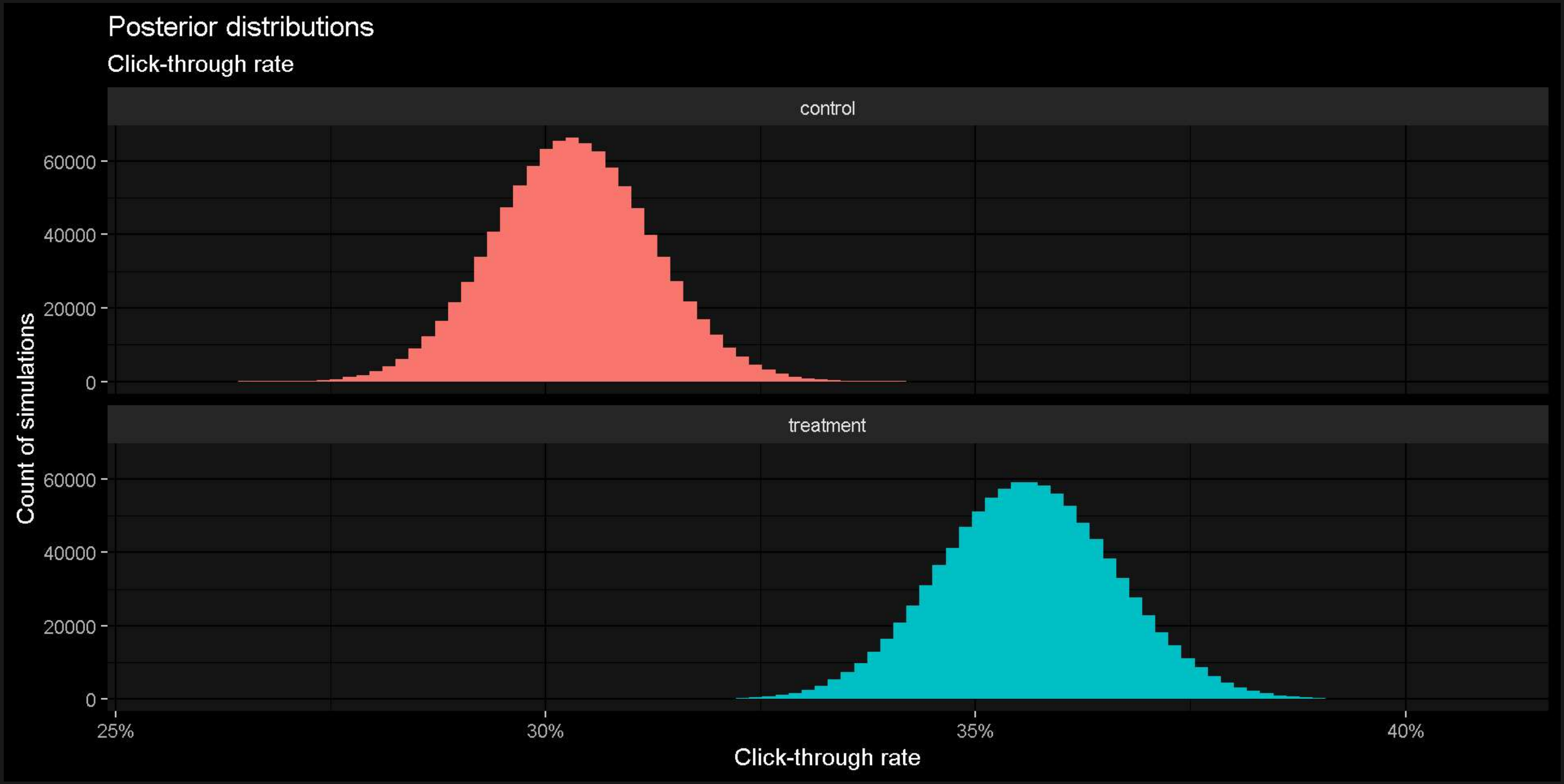
Here's some of our Monte Carlo samples:

control	treatment	uplift	beats_control
0.307	0.341	0.113	TRUE
0.310	0.363	0.170	TRUE
0.294	0.362	0.230	TRUE
0.308	0.366	0.189	TRUE
0.297	0.379	0.276	TRUE
0.308	0.347	0.125	TRUE
0.300	0.370	0.234	TRUE



Let’s now beef it up a bit…

We’ll now draw 1,000,000 samples…



We can now make some inferences

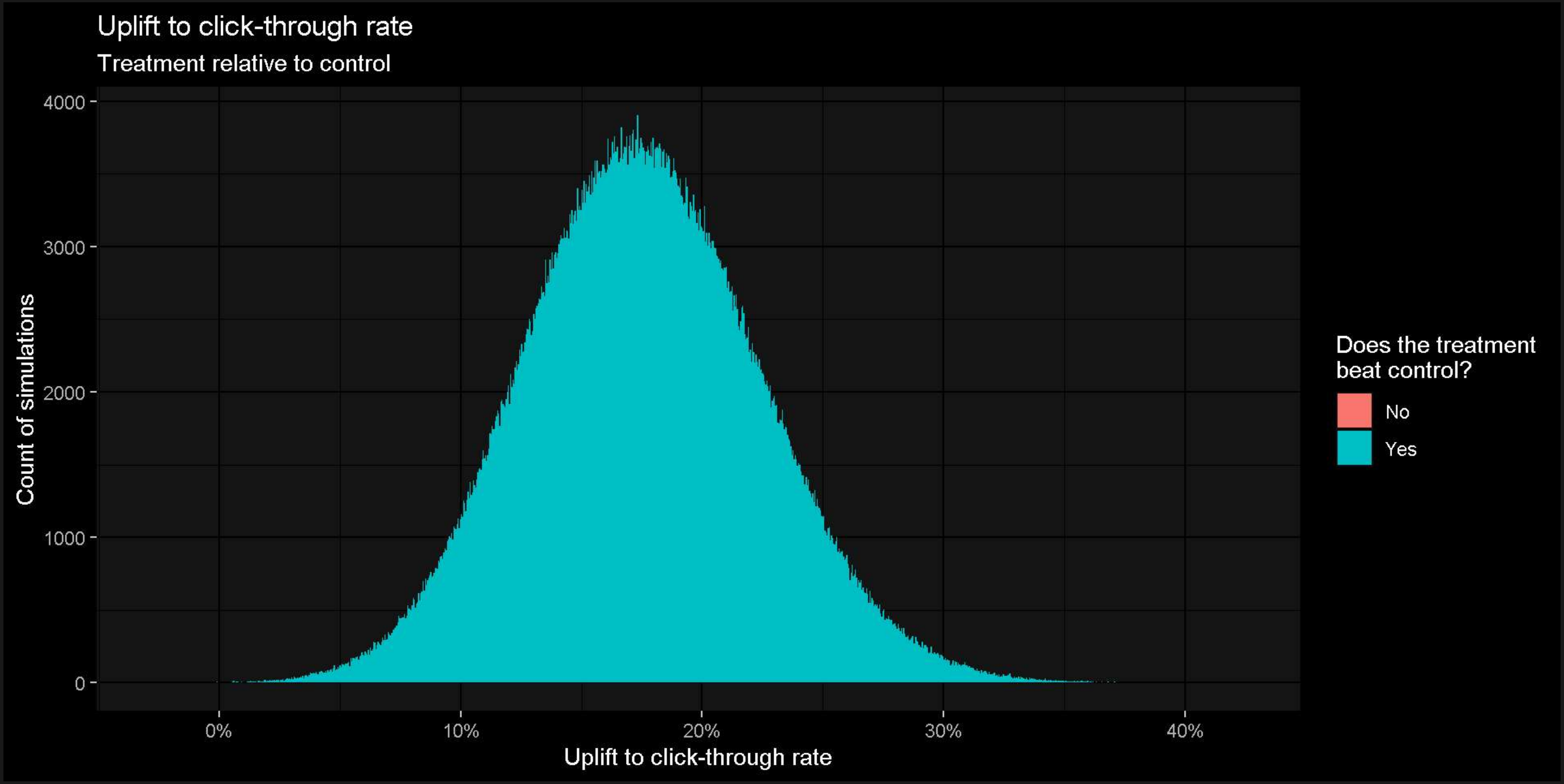
A summary of our 1,000,000 posterior samples for click-through rate:

```
    control          treatment           uplift          beats_control  
 Min.   :0.2579   Min.   :0.3095   Min.   :-0.02791   Mode :logical  
 1st Qu.:0.2970   1st Qu.:0.3491   1st Qu.: 0.14178   FALSE:51       
 Median :0.3032   Median :0.3560   Median : 0.17433   TRUE :999949   
 Mean   :0.3032   Mean   :0.3561   Mean   : 0.17539                  
 3rd Qu.:0.3094   3rd Qu.:0.3629   3rd Qu.: 0.20789                  
 Max.   :0.3467   Max.   :0.4087   Max.   : 0.42535                  
```

- How do these compare to our theoretical CTRs of 32% for control and 35% for treatment, and uplift of 9.4%?

- What is the posterior probability that the CTR of the treatment is greater than that of control? Answer: 99.99%

Posterior distribution of the CTR uplift



When to stop a Bayesian A/B test?

If using *uninformative* priors…

If your original priors are uninformative or too weak, then you face the same risks as with frequentist experiments.

Perform **power analysis** ahead of running the experiment. This is to determine the required sample size before any inferences are made.

Before commencing the experiment, decide on:

- The **minimum detectable effect** size
- The accepted **false positive rate**
- The accepted **false negative rate**

If using *informative* priors…

If your priors are relatively informative and chosen carefully, then this can reduce the chances of false positives and negatives. But:

- Be careful not to bias the results of the experiment.
- Power analysis is still recommended in order to gauge the worse case scenario for how long the experiment might run.

Bayesian inference, with informative priors, can make it possible to end an experiment early.

If deciding to end early…

Ask yourself:

- Has the experiment run for at least a couple of cycles? (e.g. at least two full weeks)
- Have the results stabilised? Is there a clear winner?
- Could it be worth running longer to learn more?
- What are the **risks** of continuing or ending now? What if the results you see are just a fluke and are therefore misguiding you? What is the impact of making the wrong choice? What are the chances?

Summary and some final remarks

Before starting an experiment

Gather prior knowledge and articulate beliefs:

- Establish a baseline - what do you know about the control group?
- What do you expect the effect of the treatment to be? How sure are you?

Express those beliefs and knowledge as distributions - these are your priors for your control and treatment groups.

Running the experiment

- Start the experiment, gather data, and update your priors to form posteriors about the metric of interest
- Draw inferences by running a large number of Monte Carlo simulations using the posterior distributions
- Know when to end the experiment – try to plan for this ahead of running the experiment

Questions?

Further topics that might interest you:

- **Bayesian Generalised Linear Models** to better isolate the effect of the treatment from other predictors (such as seasonality, device types, time of day, etc.)
- **Survival Analysis**, such as **Kaplan Meier**, to analyse lagged conversion outcomes (such as with trial periods) in order to make the most of all of the data you’ve collected.

Tip

These presentation slides and simulations have been produced in RStudio using Quarto. You can download the source code and slides from Github at: https://github.com/jdeboer/measurecamp2022

Extras

Bayes theorem

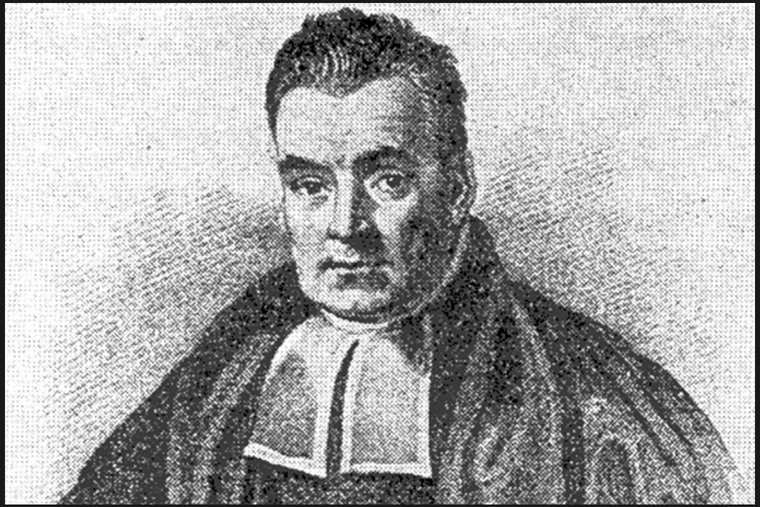
$$P(B \cap A) = P(A \cap B)$$

$$P(B) \times P(A|B) = P(A) \times P(B|A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$$Posterior \propto \mathcal{L}(\theta|data) \times prior$$



Thomas Bayes - 1701 – 1761

Some useful formulas

Let α and β represent the first and second shape parameters of the Beta distribution, respectively.

The mean of this distribution is: $\mu = \frac{\alpha}{\alpha+\beta}$

The standard deviation is: $\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$

Through substitution and rearrangement, you can determine α and β from μ and σ.

$$v = \frac{\mu(1-\mu)}{\sigma^2} - 1$$

$$\alpha = \mu v$$

$$\beta = (1-\mu)v$$

This way, you can determine the shape parameters based on centrality and spread.