

# 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
- Summary and some final remarks
- Thank you!
- When to stop a Bayesian A/B test?
- Extras
- Game of chances

# Setting the scene

# Randomised Control Trials (RCTs)

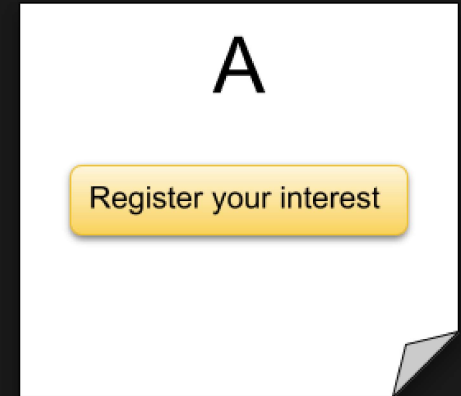
A simplistic example:

- 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.



# Priors and probability distributions

The key to speeding up your experiment

# Prior knowledge and beliefs

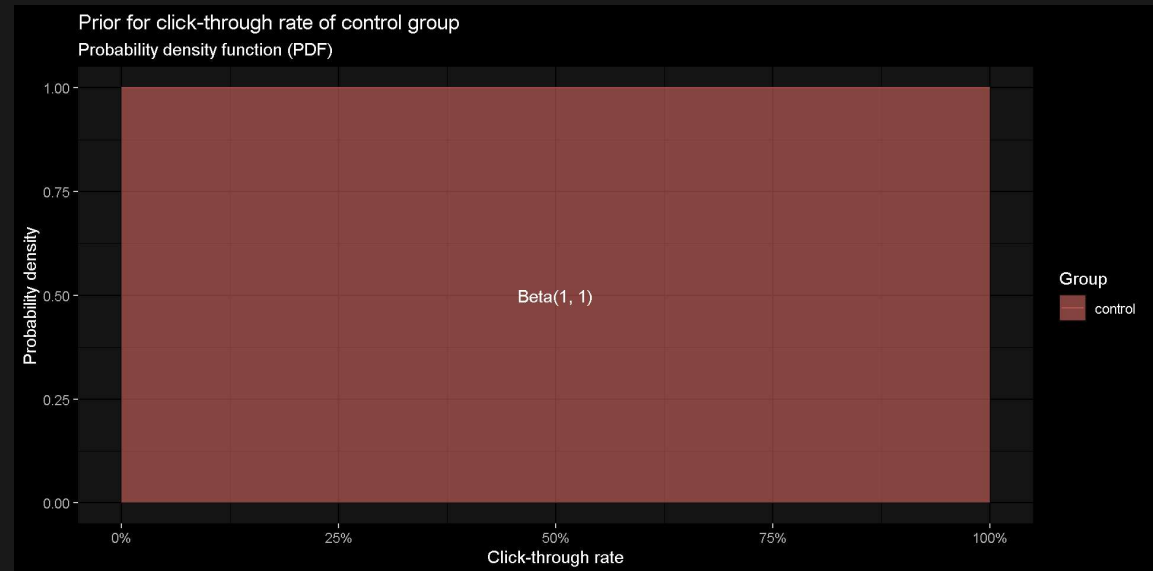
Before running an experiment, we form opinions and gather evidence such as:

- 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.

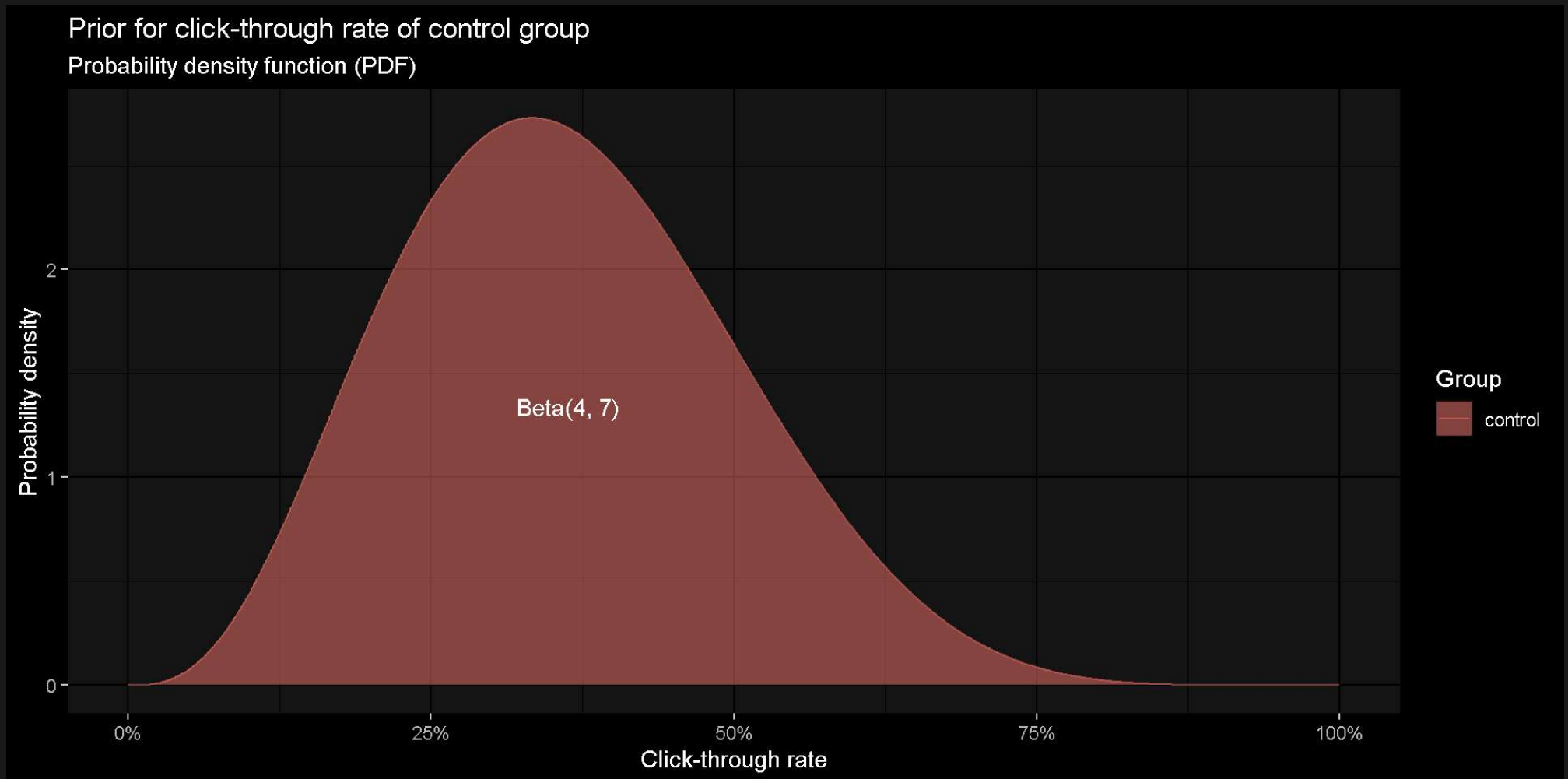


## Tip

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.

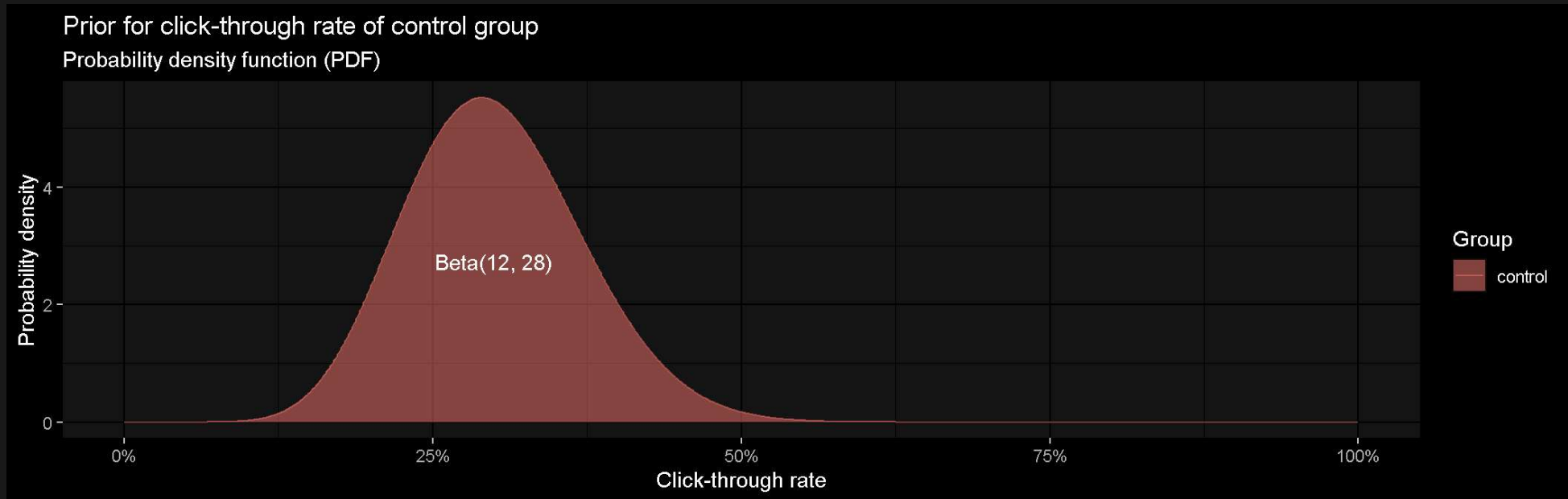


# Something a little more informative



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

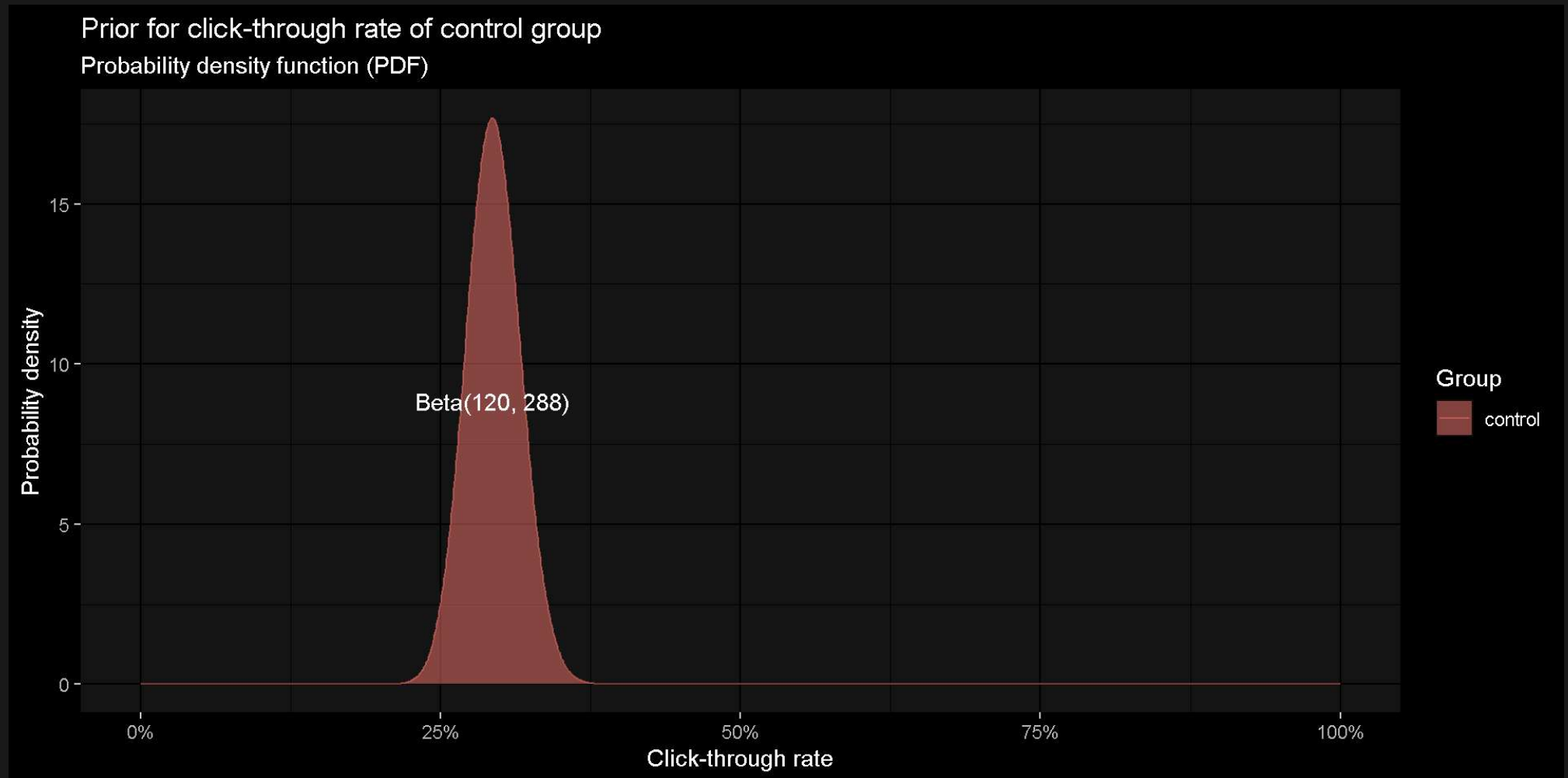
# Something even more informative



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:

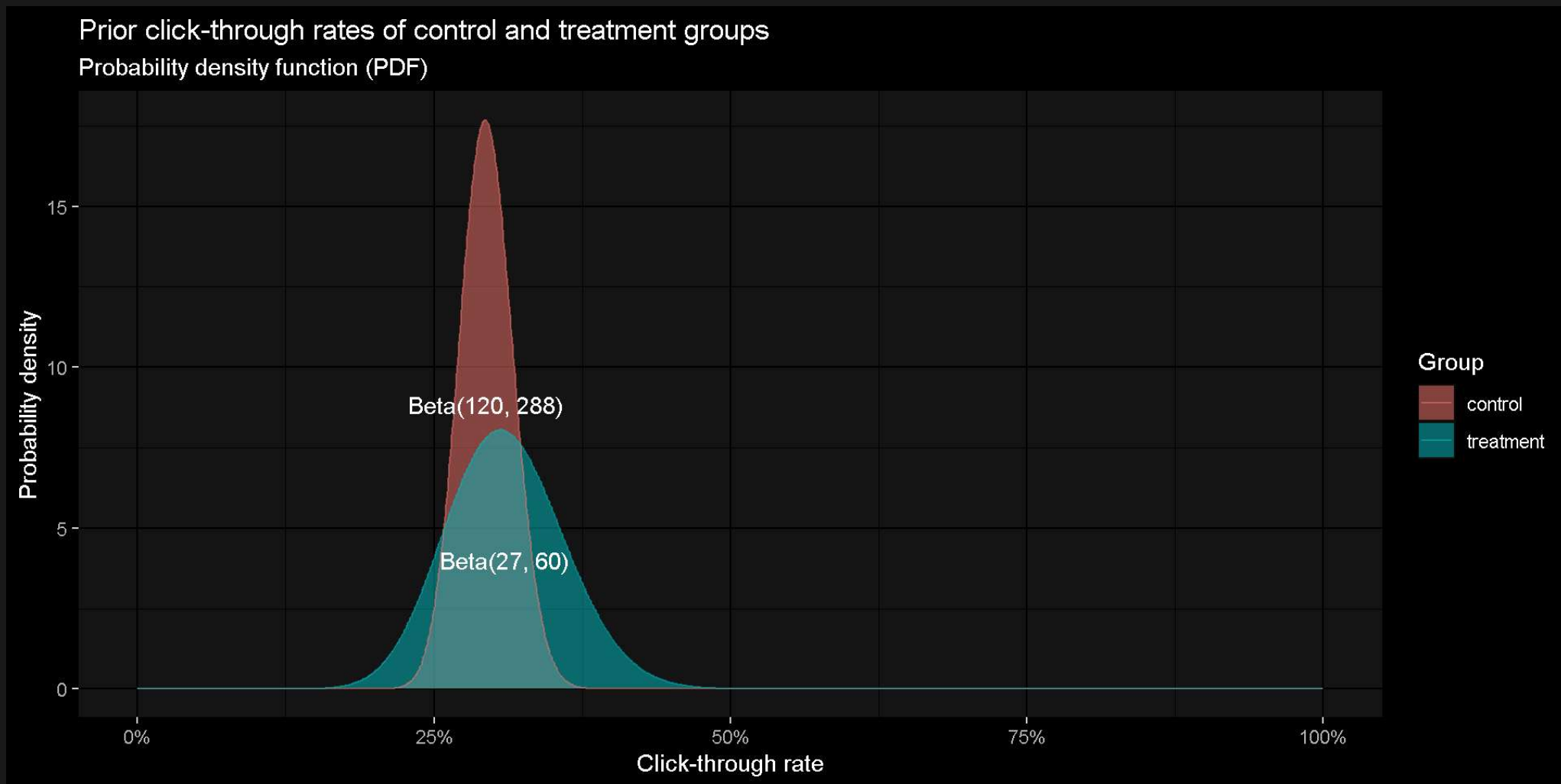


The more confident we are about our beliefs, the narrower the curve.

# 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	86	32	54	37.2%
treatment	81	26	55	32.1%

# 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.

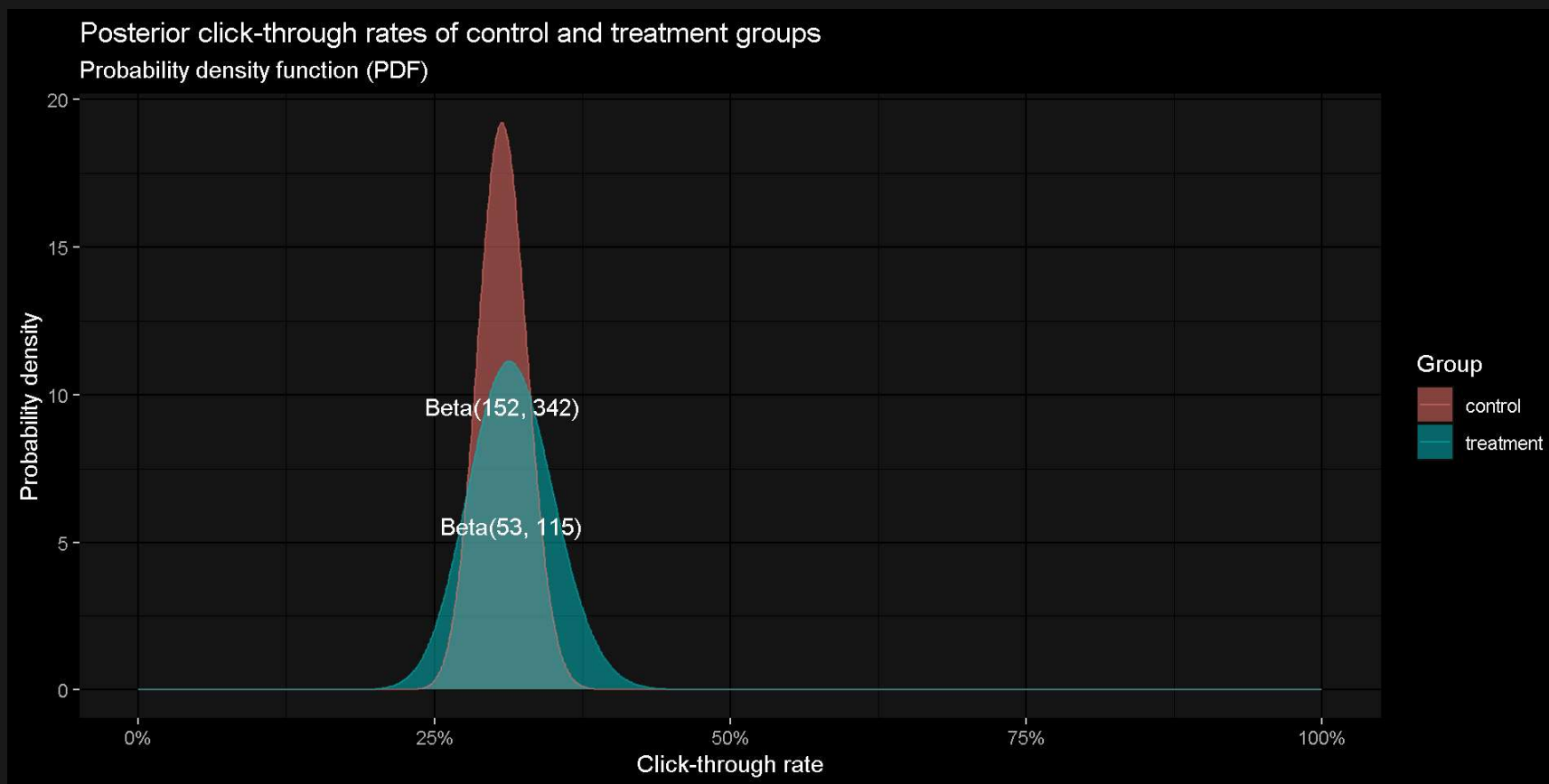
For each experiment group, we derive our posterior shape parameters through simple arithmetic addition:

- 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	32	26
posterior_shape1	152	53
count	control	treatment
prior_shape2	288	60
not_clicked	54	55
posterior_shape2	342	115

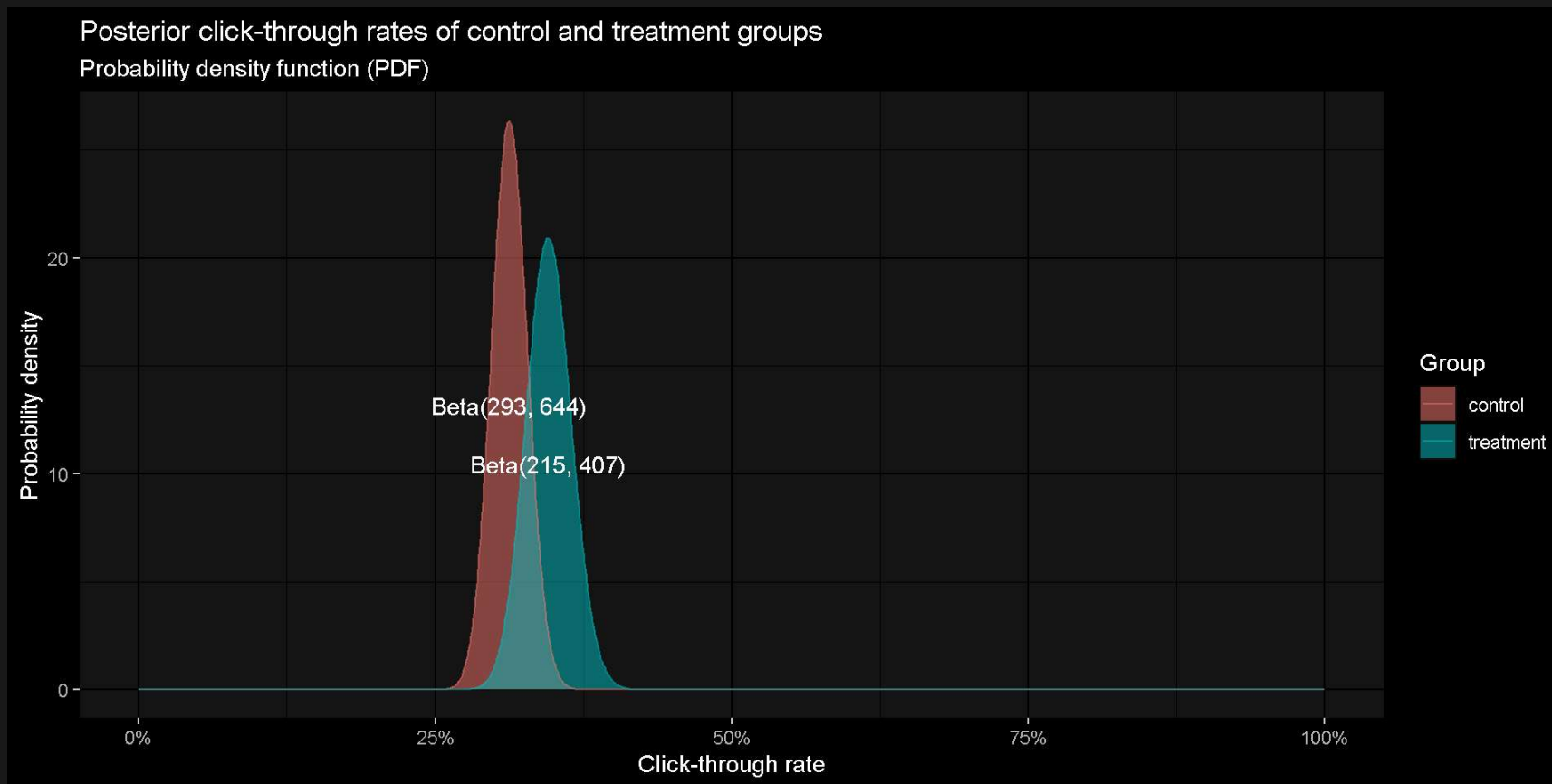
# Posterior distribution of each group

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



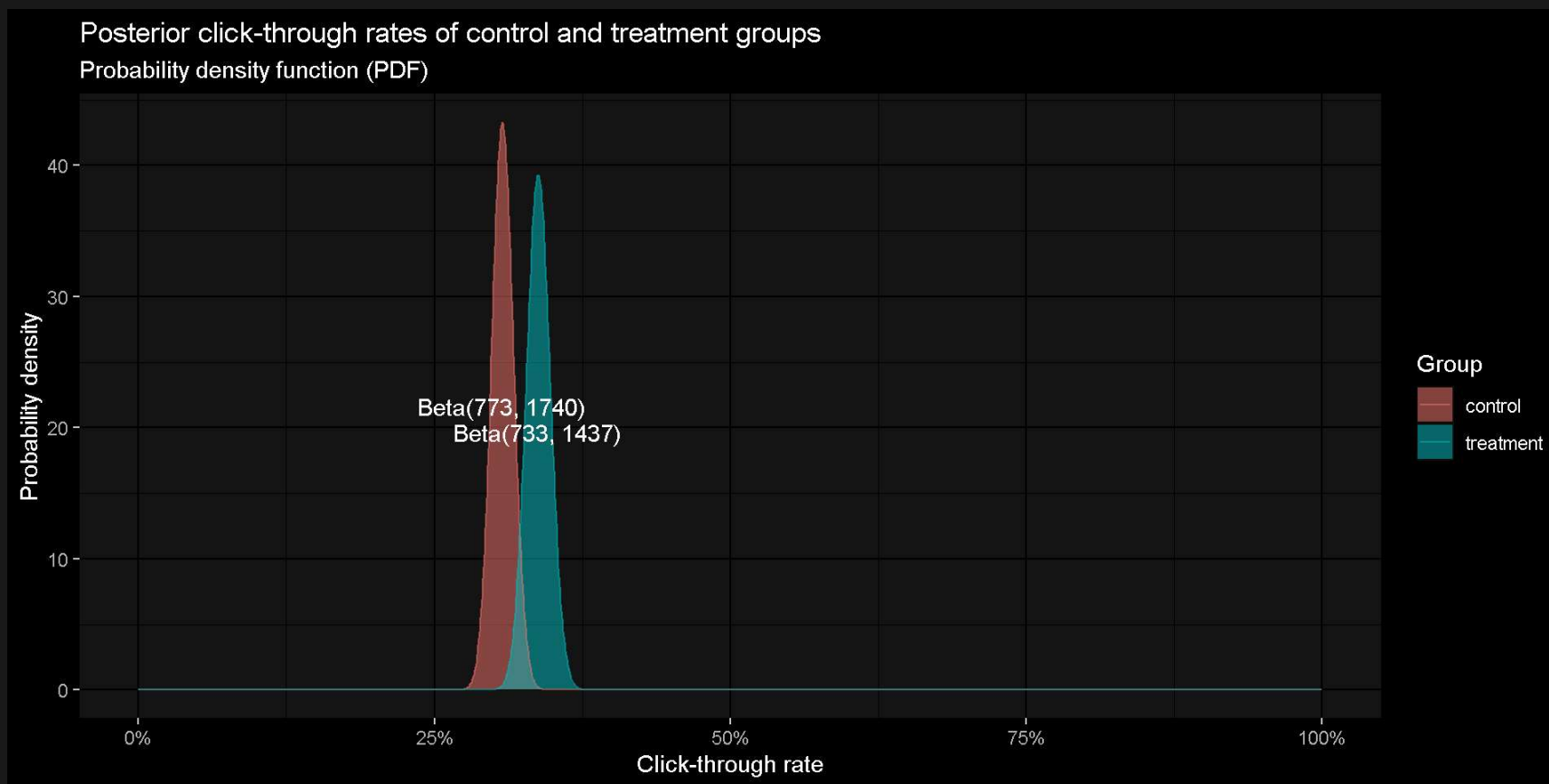
# Another six days later...

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 4,188 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.

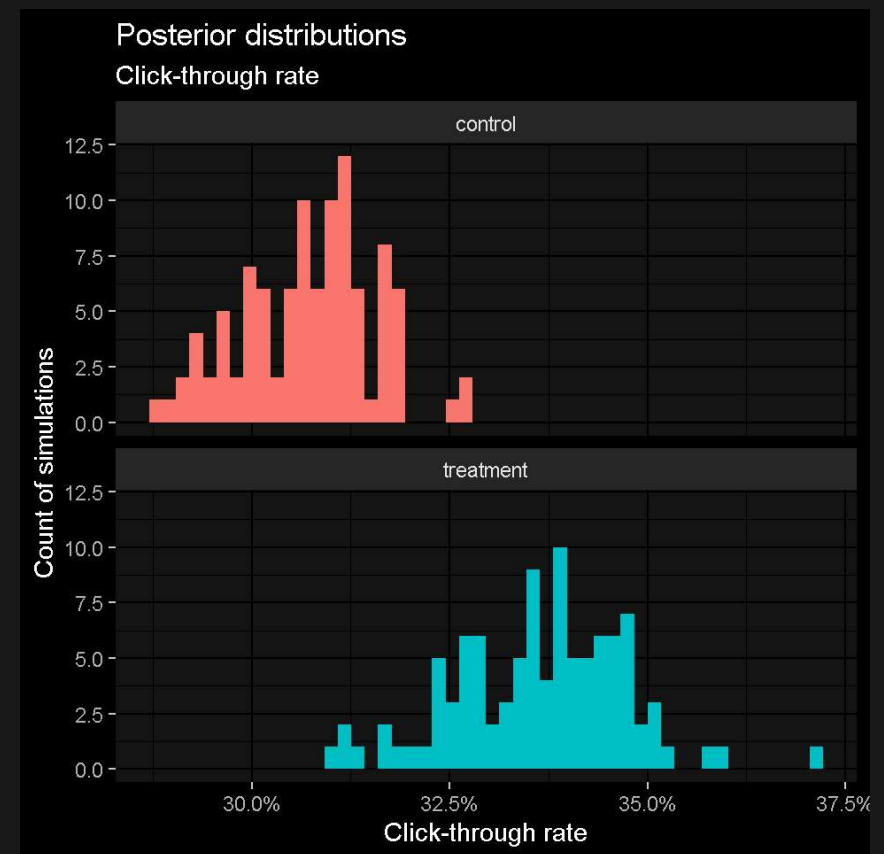


# 100 simulations

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:

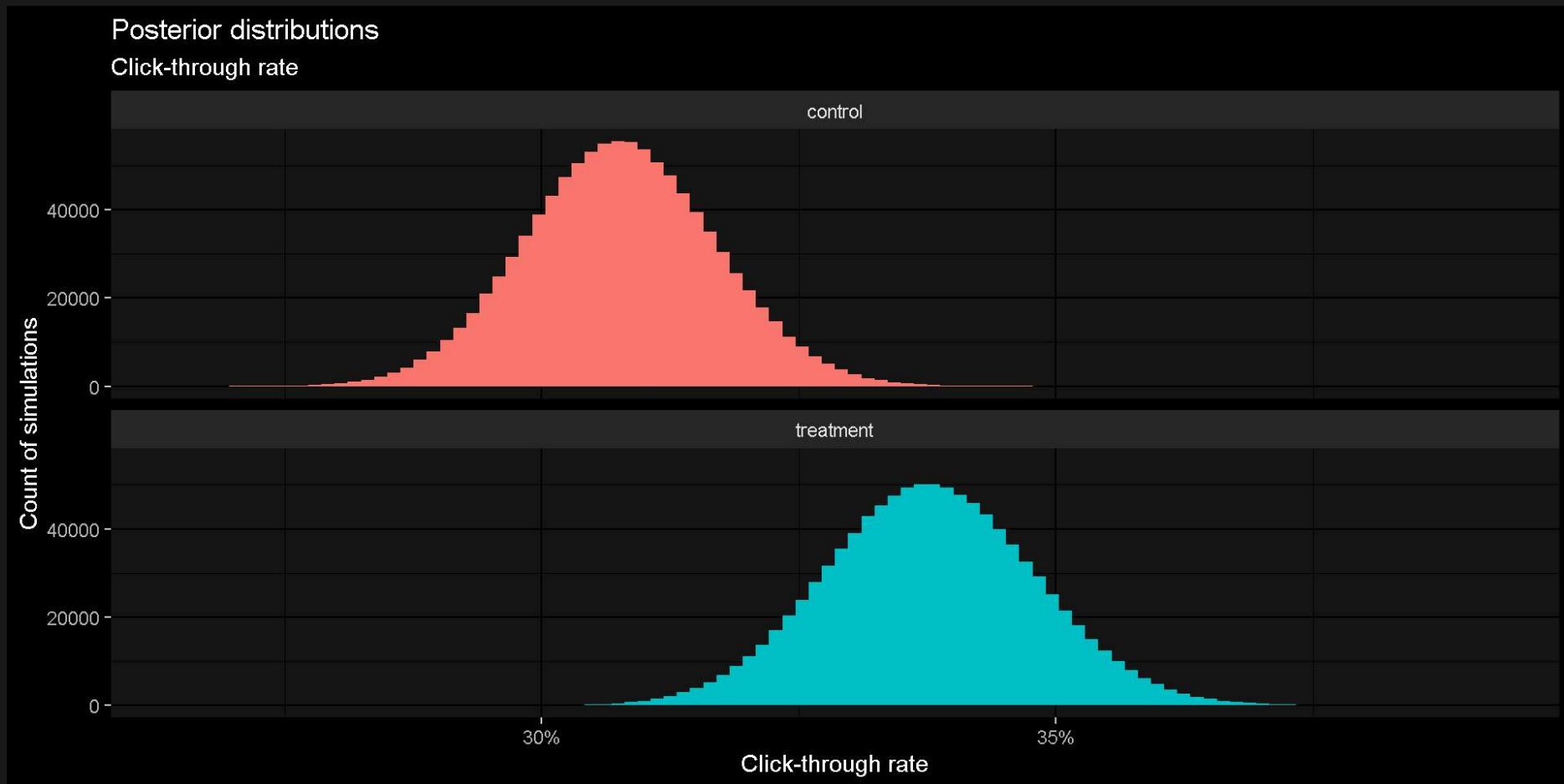
control	treatment	uplift	beats_control
0.316	0.340	0.074	TRUE
0.317	0.358	0.131	TRUE
0.310	0.347	0.119	TRUE
0.294	0.328	0.117	TRUE
0.318	0.341	0.075	TRUE
0.301	0.347	0.154	TRUE
0.309	0.346	0.118	TRUE





# Let's now beef it up a bit...

We'll now draw 1,000,000 samples...



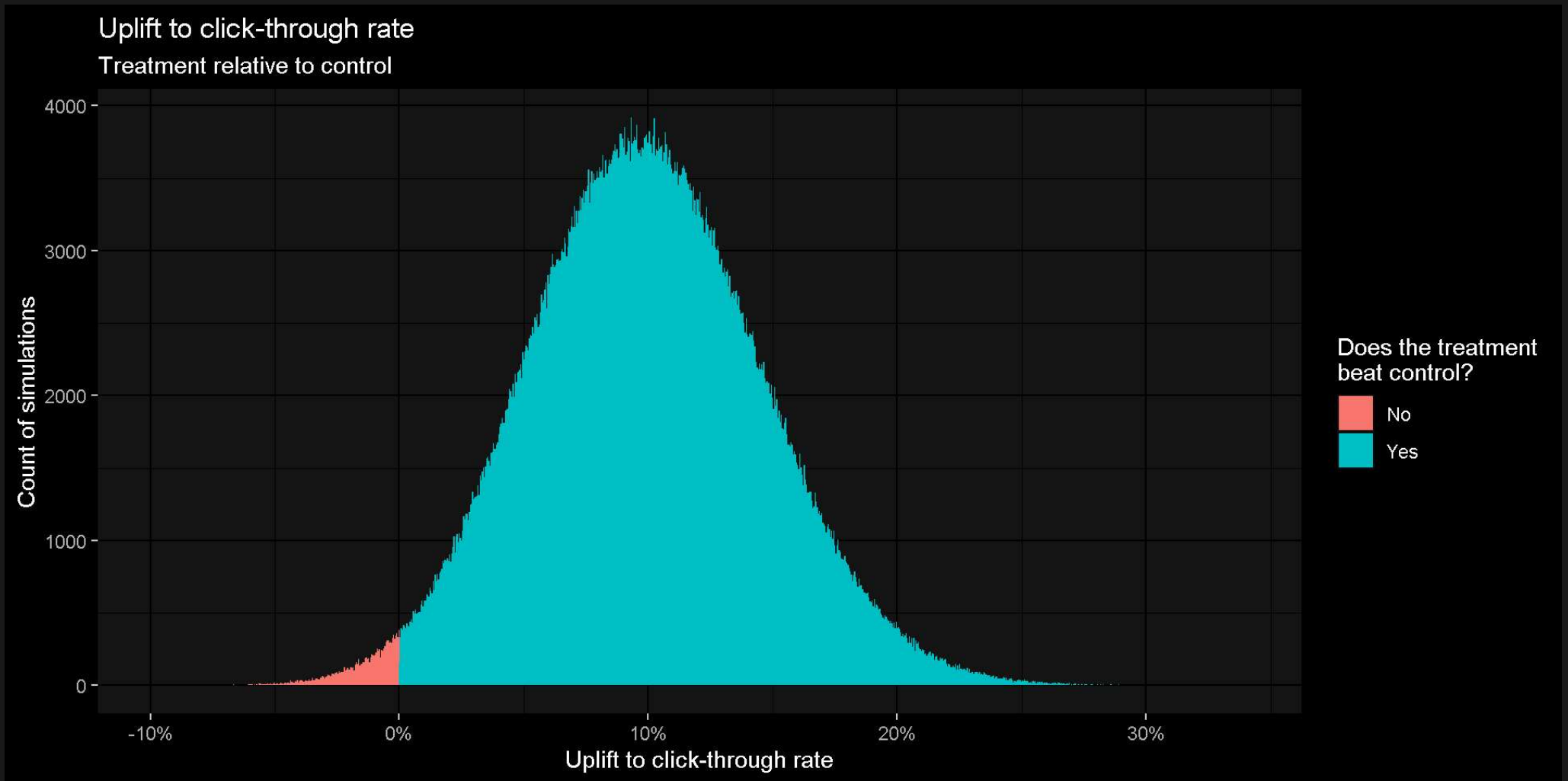
# We can now make some inferences

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

control	treatment	uplift	beats_control
Min. :0.2656	Min. :0.2902	Min. : -0.09909	Mode :logical
1st Qu.:0.3014	1st Qu.:0.3309	1st Qu.: 0.06710	FALSE:13815
Median :0.3075	Median :0.3377	Median : 0.09819	TRUE :986185
Mean :0.3076	Mean :0.3378	Mean : 0.09914	
3rd Qu.:0.3138	3rd Qu.:0.3446	3rd Qu.: 0.12999	
Max. :0.3525	Max. :0.3924	Max. : 0.34028	

- 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: 98.62%

# Posterior distribution of the CTR uplift



# 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

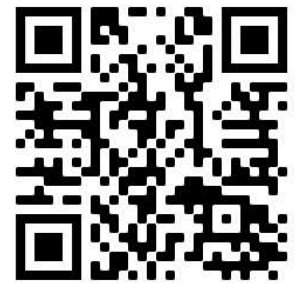
# Thank you!

Further topics that might interest you:

- **Bayesian Generalised Linear Models** to better isolate the effect of the treatment from other predictors.
- **Survival Analysis**, such as **Kaplan Meier**, to analyse lagged conversion outcomes.



These slides and simulations were produced in RStudio using Quarto. Download the source code and slides at:  
<https://github.com/jdeboer/measurecamp2022>



# 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 misleading you? What is the impact of making the wrong choice? What are the chances?

# Extras

# Bayes theorem

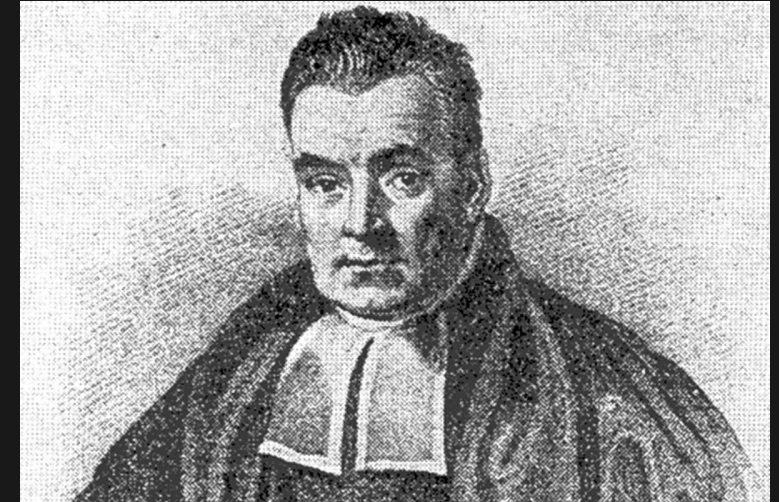
$$P(B \cap A) = P(A \cap B)$$

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

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

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

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



Thomas Bayes - 1701 – 1761

# Some useful formulas

Let  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  from  $\mu$  and  $\sigma$ .

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

$$\alpha = \mu v$$

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

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

# Game of chances

# To play this game:

- There is one host and at least 2 contestants.
- The host will need a uniform random number generator.



# Instructions

1. The host secretly picks a number,  $Y$ , between 0% and 100% and writes it down.
2. The host will then secretly generate a random number,  $X$ , again between 0% and 100%:
  - If  $X$  is less than  $Y$  then the host will mark it as a 'success', otherwise as a 'failure', using a tally board that's visible to all contestants.
3. The objective of the game is for contestants to estimate  $Y$  by asking the host to perform step 2 as many times as they need (to a reasonable limit). The closer their guess is, the better. However, each contestant can only call out their guess once.
4. Once two contestants have called out what they believe  $Y$  is, then the game ends and the host reveals the true answer for  $Y$ .
  - The first contestant to call out their guess wins the game if they are within 5 points of  $Y$ . If not, then the contestant whose guess is closest to  $Y$  wins.

# Lessons of the game

- As the number of successes and failures increases, you get closer to knowing what  $Y$  is.
- The compromise each contestant makes between speed and certainty will influence who wins.



## Tip

Try a modification to the game where  $Y$  is a number related to a topic that the audience will have some prior knowledge about. This way they can incorporate their prior expectations when making a guess.