

# Introduction to Bayesian statistical modelling

A course with R, Stan, and brms

Ladislas Nalborczyk (UNICOG, NeuroSpin, CEA, Gif/Yvette, France)



# Preface 🙌 🙌

This course is largely inspired from the following books:

- McElreath, R. (2016, 2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.
- Kurz, S. (2019). *Statistical Rethinking with brms, ggplot2, and the tidyverse*. Available [online](#).
- Kruschke, J. K. (2015). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press / Elsevier.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, third edition*. London: CRC Press.
- Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. SAGE Publications Ltd.
- Noël, Y. (2015). *Psychologie Statistique*. EDP Sciences.
- Nicenboim, B., Schad, D., & Vasishth, S. (2021). *An Introduction to Bayesian Data Analysis for Cognitive Science*. Available [online](#).

Code and slides are available at the course's website: <https://lnalborczyk.github.io/IBSM2023/>



# Objectives

General objectives:

- Understand the fundamental concepts of Bayesian statistical modelling.
- Be able to understand articles describing Bayesian analyses.
- Bonus: realise that the Bayesian approach is more intuitive than the frequentist approach.

Practical objectives:

- Be able to carry out a complete analysis (i.e., identifying the appropriate model, writing the mathematical model, implementing it in R, interpreting and reporting the results) of a simple dataset.



# Planning

## **Course n°01: Introduction to Bayesian inference, Beta-Binomial model**

Course n°02: Introduction to brms, linear regression

Course n°03: Markov Chain Monte Carlo, generalised linear model

Course n°04: Multilevel models, cognitive models



# A matter of interpretation

What is the probability...

- Of obtaining an odd number with a fair die?
- Of you learning something new during this course?

Do these two questions refer to the same “sort” of probability?



# Classical (or theoretical) interpretation

$$\Pr(\text{odd}) = \frac{\text{number of favorable issues}}{\text{total number of possible issues}} = \frac{3}{6} = \frac{1}{2}$$

Problem: this definition only applies to situations in which there is a **finite** number of **equiprobable** potential outcomes...

Limitation: what is the probability of raining tomorrow?

$$\Pr(\text{rain}) = \frac{\text{rain}}{\{\text{rain, not-rain}\}} = \frac{1}{2}$$



# Frequentist (or empirical) interpretation

$$\Pr(x) = \lim_{n_t \rightarrow \infty} \frac{n_x}{n_t}$$

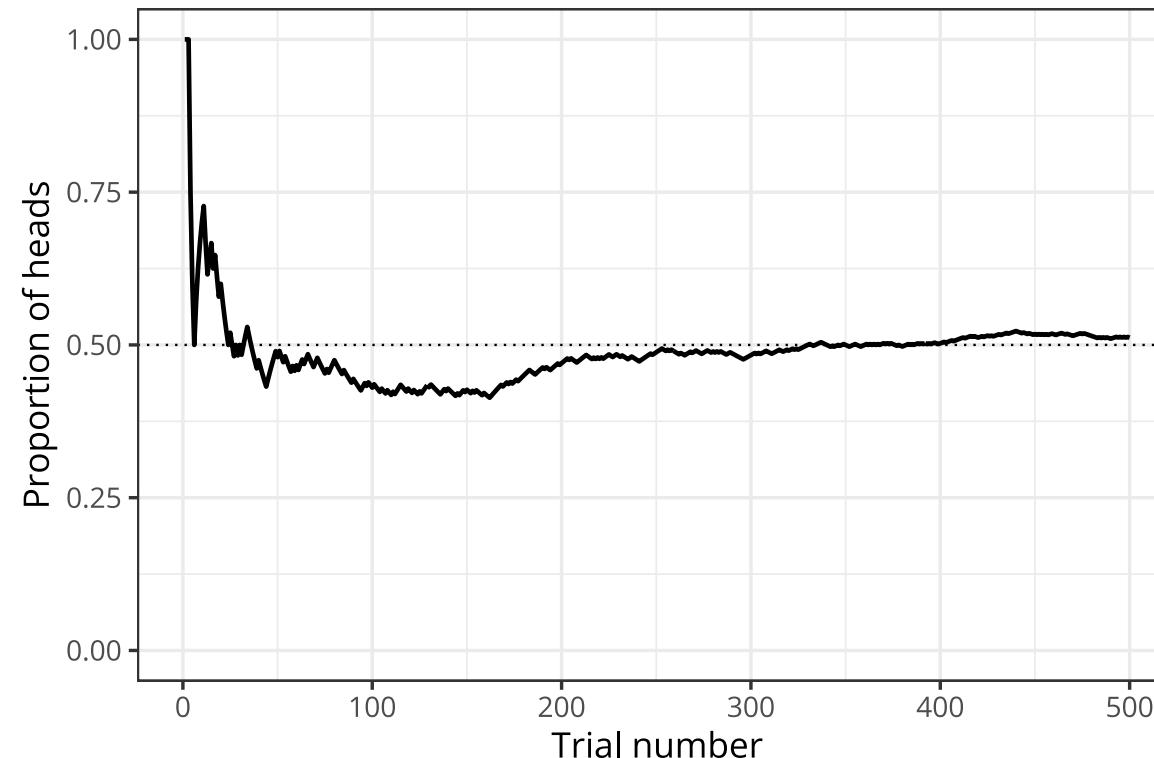
Where  $n_x$  is the number of occurrences of the event  $x$  and  $n_t$  the total number of observations. The **frequentist** interpretation postulates that, in the long-run (i.e., when the number of observations approaches infinity), the relative frequency of an event will converge exactly with what we call “probability”.

Consequence: the concept of probability only applies to collectives, not to single events.



# Frequentist (or empirical) interpretation

```
1 library(tidyverse)
2
3 sample(x = c(0, 1), size = 500, prob = c(0.5, 0.5), replace = TRUE) %>%
4   data.frame() %>%
5   mutate(x = seq_along(.), y = cummean(.)) %>%
6   ggplot(aes(x = x, y = y)) +
7   geom_line(lwd = 1) +
8   geom_hline(yintercept = 0.5, lty = 3) +
9   labs(x = "Trial number", y = "Proportion of heads") +
10  ylim(0, 1)
```



# Limitations of the frequentist interpretation

Which reference class? *What is the probability that I will live until 80 years old? As a man? As a French person?*

What about non-repeatable events? *What is the probability of you learning something new during this course?*

The resolution issue: How many observations do we need to get a good approximation of the underlying probability? A finite class of events of size  $n$  can only produce relative frequencies with precision  $1/n\dots$



# Propensionist interpretation

The frequentist (i.e., long-term) properties of objects (e.g., a coin) are caused by the intrinsic physical properties of the objects. For example, a biased coin will generate a biased relative frequency (and therefore probability) because of its physical properties. For propensionists, probabilities represent these intrinsic characteristics, these **propensities** to generate certain relative frequencies, and not the relative frequencies themselves.

Consequence: these properties are the properties of individual events... and not of sequences! The propensionist interpretation therefore allows us to talk about the probability of single events.



# Logical interpretation

There are 10 students in this room

9 wear a green t-shirt

1 wears a red t-shirt

One person is drawn at random...

---

Conclusion #1: the student drawn wears a t-shirt ✓

---

Conclusion #2: the student drawn wears a green t-shirt ✗

---

Conclusion #3: the student selected at random wears a red t-shirt ✗



# Logical interpretation

The logical interpretation of the concept of probability attempts to generalise logic (true/false) to the probabilistic world. Probability therefore represents the **degree of logical support** that a conclusion has, relative to a set of premises ([Carnap, 1971](#); [Keynes, 1921](#)).

Consequence: all probability is **conditional**.



# Bayesian interpretation

Probability is **a measure of the degree of uncertainty**. An event that is *certain* will therefore have a probability of 1 and an event that is *impossible* will have a probability of 0.

“

So to assign equal probabilities to two events is not in any way an assertion that they must occur equally often in any random experiment [...], it is only a formal way of saying I don't know ([Jaynes, 1986](#)).

To talk about probabilities in this context, we no longer need to refer to the limit of occurrence of an event (frequency).



# Probabilistic interpretations - Summary

- Classical interpretation (Laplace, Bernouilli, Leibniz)
- **Frequentist interpretation** (Venn, Reichenbach, von Mises)
- Propensionist interpretation (Popper, Miller)
- Logical interpretation (Keynes, Carnap)
- **Bayesian interpretation** (Jeffreys, de Finetti, Savage)

For more details, see this article from the Stanford Encyclopedia of Philosophy.



# Probabilistic interpretations - Summary

## Epistemic probability

All probabilities are conditional on available information (e.g., premises or data). Probability is used as a means of quantifying uncertainty.

Logical interpretation, Bayesian interpretation.

## Physical probability

Probabilities depend on a state of the world, on physical characteristics, and are independent of available information (or uncertainty).

Classical interpretation, frequentist interpretation.





# Probability axioms (Kolmogorov, 1933)

A probability is a numerical value assigned to an event  $A$ , understood as a possibility belonging to the set of all possible outcomes  $\Omega$ .

Probabilities have to conform to the following axioms:

- Non-negativity:  $\Pr(A) \geq 0$
- Normalisation:  $\Pr(\Omega) = 1$
- Additivity (for mutually exclusive events):  $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$

The last axiom is also known as the **sum rule** and can be generalised to non-mutually exclusive events:  
 $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$ .



# Sum rule and product rule

**Sum rule** (for two mutually exclusive events):  $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$ .

Think about the probability of getting an odd number with a fair die. We may also write it as  $\Pr(x = 1) + \Pr(x = 3) + \Pr(x = 5) = \frac{3}{6}$ .

**Product rule** (for two independent events):  $\Pr(A_1 \cap A_2) = \Pr(A_1) \times \Pr(A_2)$ .

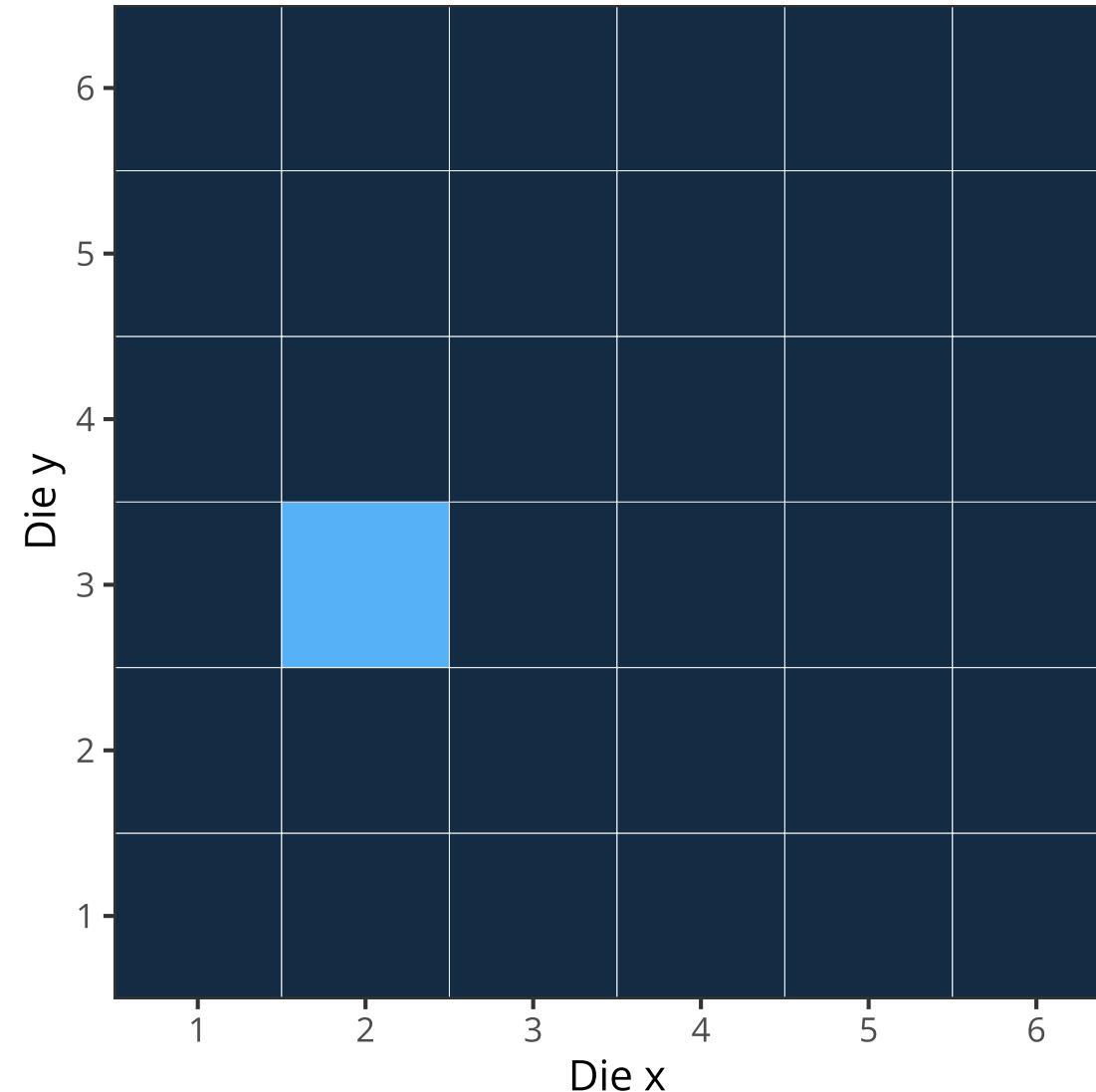
Think about the probability of getting two 6s in a row with a fair die. We may also write it as  $\Pr(x = 6, y = 6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ .

If you understand and remember these two rules, you already know Bayesian statistics!



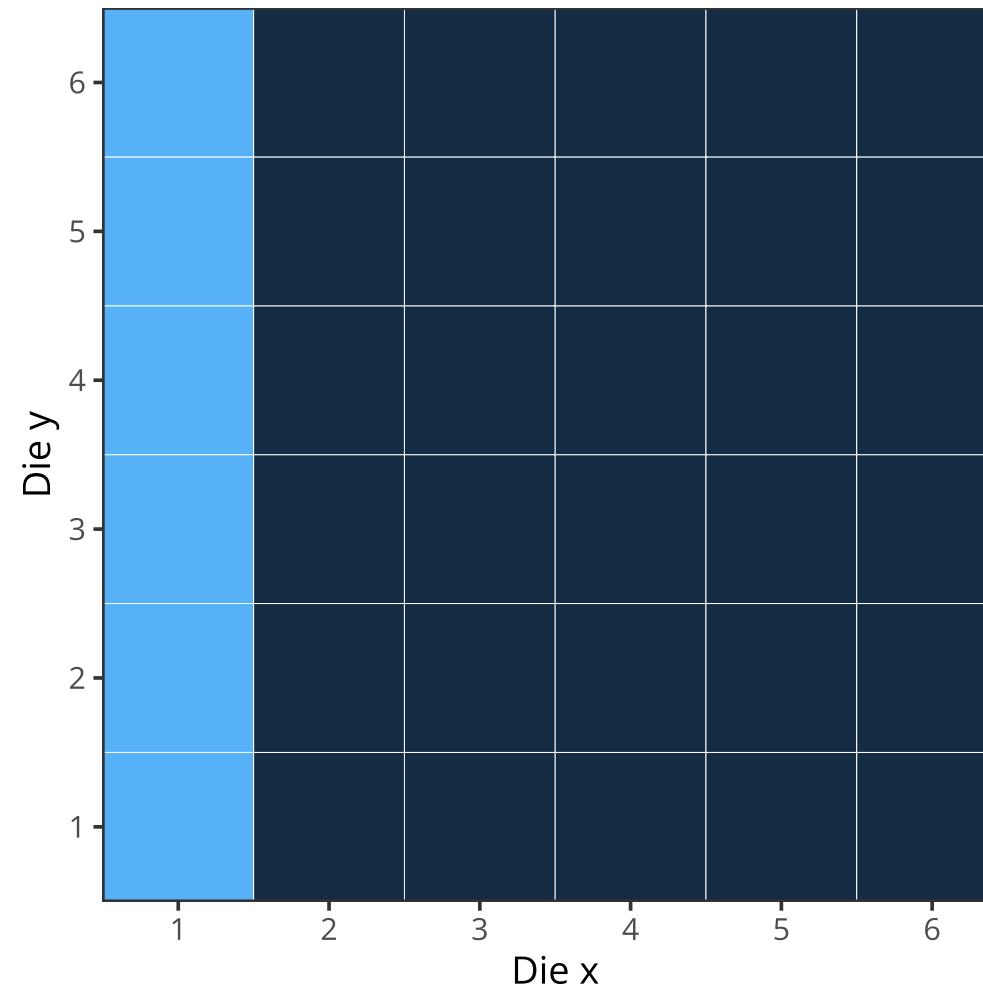
# Joint probability

Probability that die  $x$  is equal to 2 and die  $y$  is equal to 3 is:  $\Pr(x = 2, y = 3) = \Pr(y = 3, x = 2) = \frac{1}{36}$ .



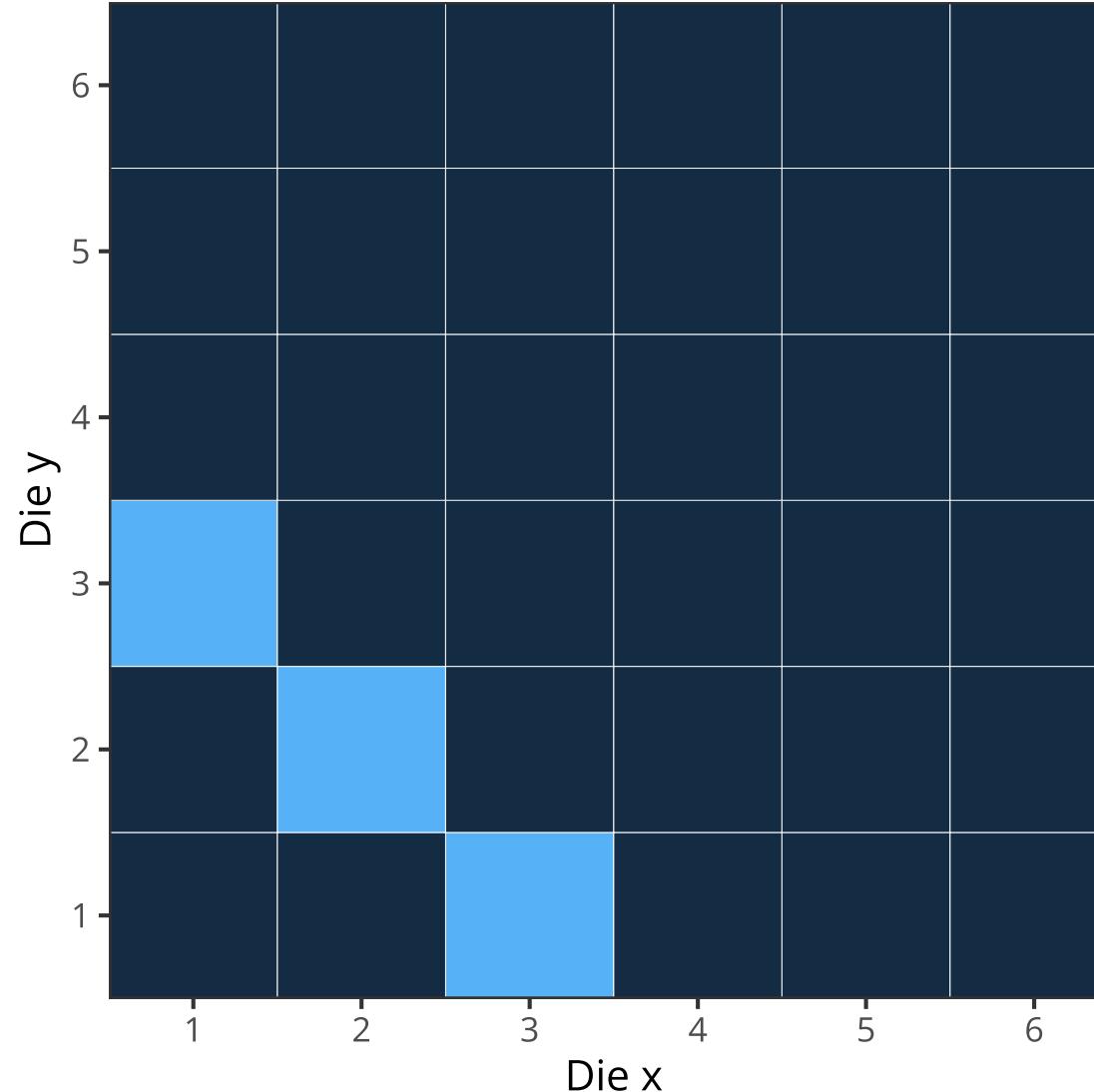
# From the sum rule to marginalisation

With more than one variable, the sum rule tells us how to ignore one. For instance, the probability that the first die shows 1 is:  $\Pr(x = 1) = \Pr(x = 1, y \in \{1, 2, 3, 4, 5, 6\}) = \frac{6}{36}$ . This is called **marginal** because you can write the cumulative probability in the **margin** of a joint probability table.



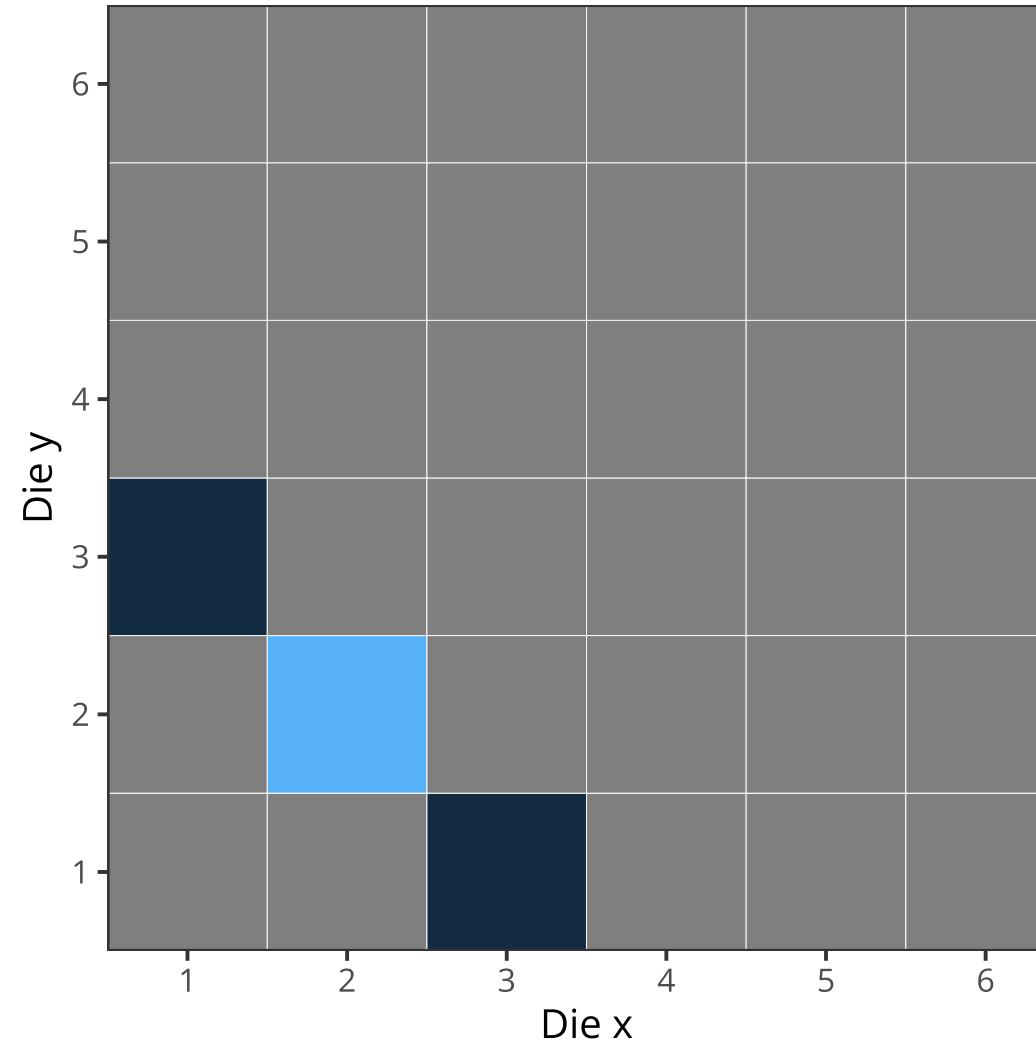
# From the sum rule to marginalisation

Probability that two dice total 4 is:  $\Pr(x + y = 4) = \frac{3}{36}$ .



# Conditional probability

What is the probability that die  $x$  equals some value **given that** the total is 4? For instance, the probability of die  $x$  being equal to 2:  $\Pr(x = 2 \mid x + y = 4) = \frac{1}{3}$ .



# Conditional probability

This conditional probability can be rewritten:  $\Pr(x = 2 \mid x + y = 4) = \frac{\Pr(x=2,x+y=4)}{\Pr(x+y=4)} = \frac{1/36}{3/36} = \frac{1}{3}$ . Note that  $\Pr(x \mid y)$  **is not necessarily equal** (and is generally not equal) to  $\Pr(y \mid x)$ .

For instance: the probability of dying knowing that you have been attacked by a shark is not the same as the probability of having been attacked by a shark knowing that you are dead ([confusion of the inverse](#)). In the same way,  $p(\text{data} \mid \mathcal{H}_0) \neq p(\mathcal{H}_0 \mid \text{data})$ .



# Deriving Bayes theorem

From Kolmogorov's axioms and the previous definitions of joint, marginal, and conditional probabilities, we derive the **product rule** (for dependent events):

$$p(x, y) = p(x \mid y) p(y) = p(y \mid x) p(x)$$

$$p(y \mid x) p(x) = p(x \mid y) p(y)$$

Then divide each side by  $p(x)$ :

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{p(x)}$$

$$p(x \mid y) = \frac{p(y \mid x) p(x)}{p(y)}$$

If we replace  $x$  by hypothesis and  $y$  by data:

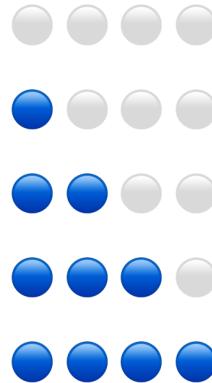
$$\Pr(\text{hypothesis} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{hypothesis}) \times \Pr(\text{hypothesis})}{\text{sum of products}}$$



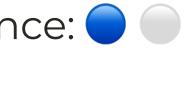
## Exercise - Bag of marbles problem (McElreath, 2020)

Let's imagine we have a bag containing 4 marbles. These marbles can be either white or blue. We know that there are precisely 4 marbles, but we don't know the number of marbles of each colour.

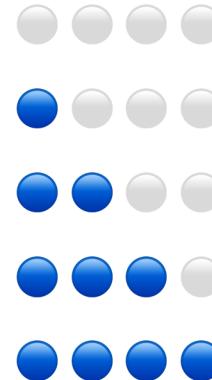
However, we do know that there are five possibilities (which we consider to be our **hypotheses**):



## Exercise - Bag of marbles problem (McElreath, 2020)

The aim is to determine which combination is the most likely, given certain observations. Let's assume that we drawn three marbles in a row, with replacement, and we obtained the following sequence: .

This sequence represents our data. What **inference** can we make about the contents of the bag? In other words, what can we say about the probability of each hypothesis?



02:00  
Eduardo Nalborczyk - IPSM2023



# Enumerating possibilities

Hypothesis: ● ○ ○ ○

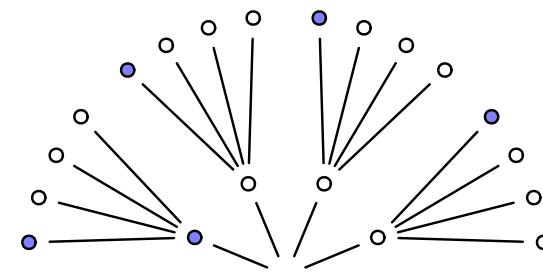
Data: ●



# Enumerating possibilities

Hypothesis: ● ○ ○ ○

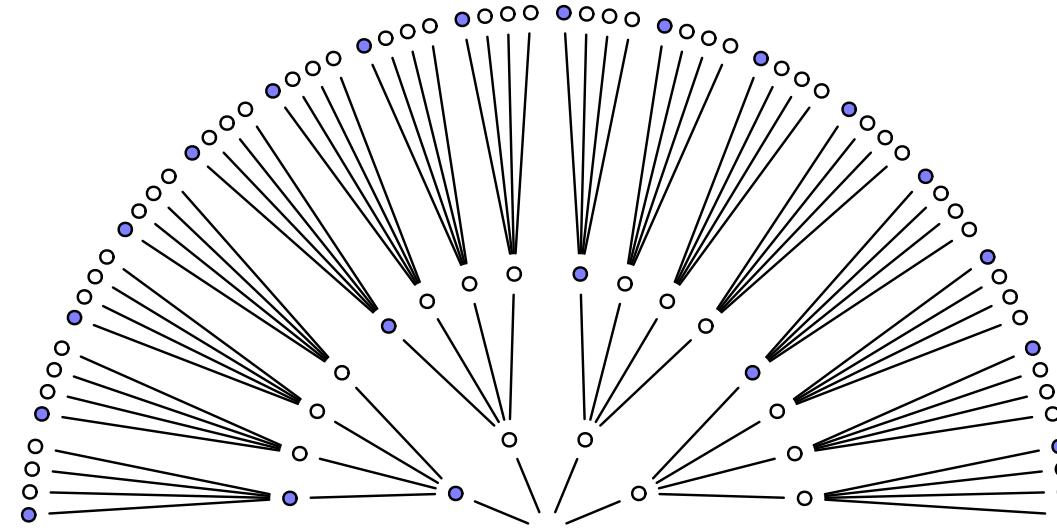
Data: ● ○



# Enumerating possibilities

Hypothesis: ● ○ ○ ○

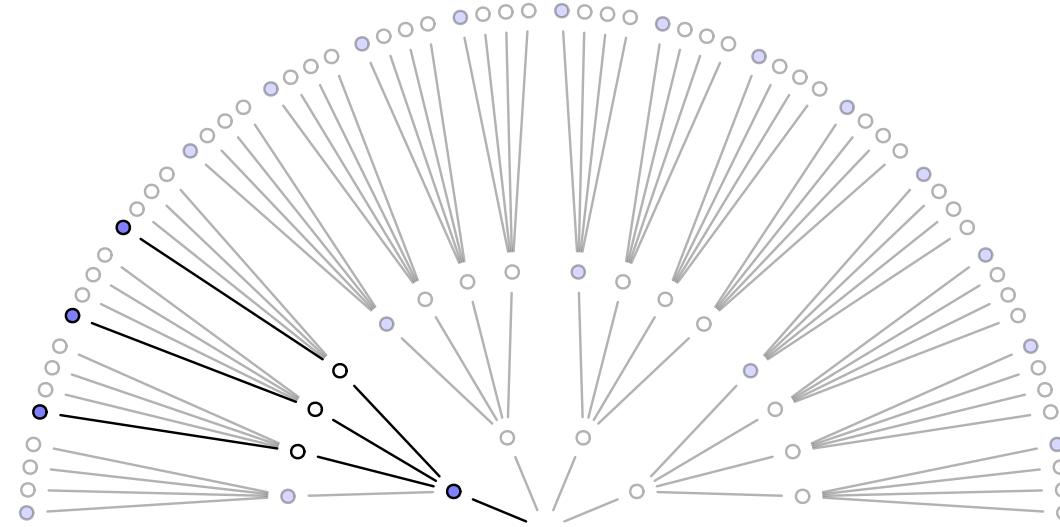
Data: ● ○ ●



# Enumerating possibilities

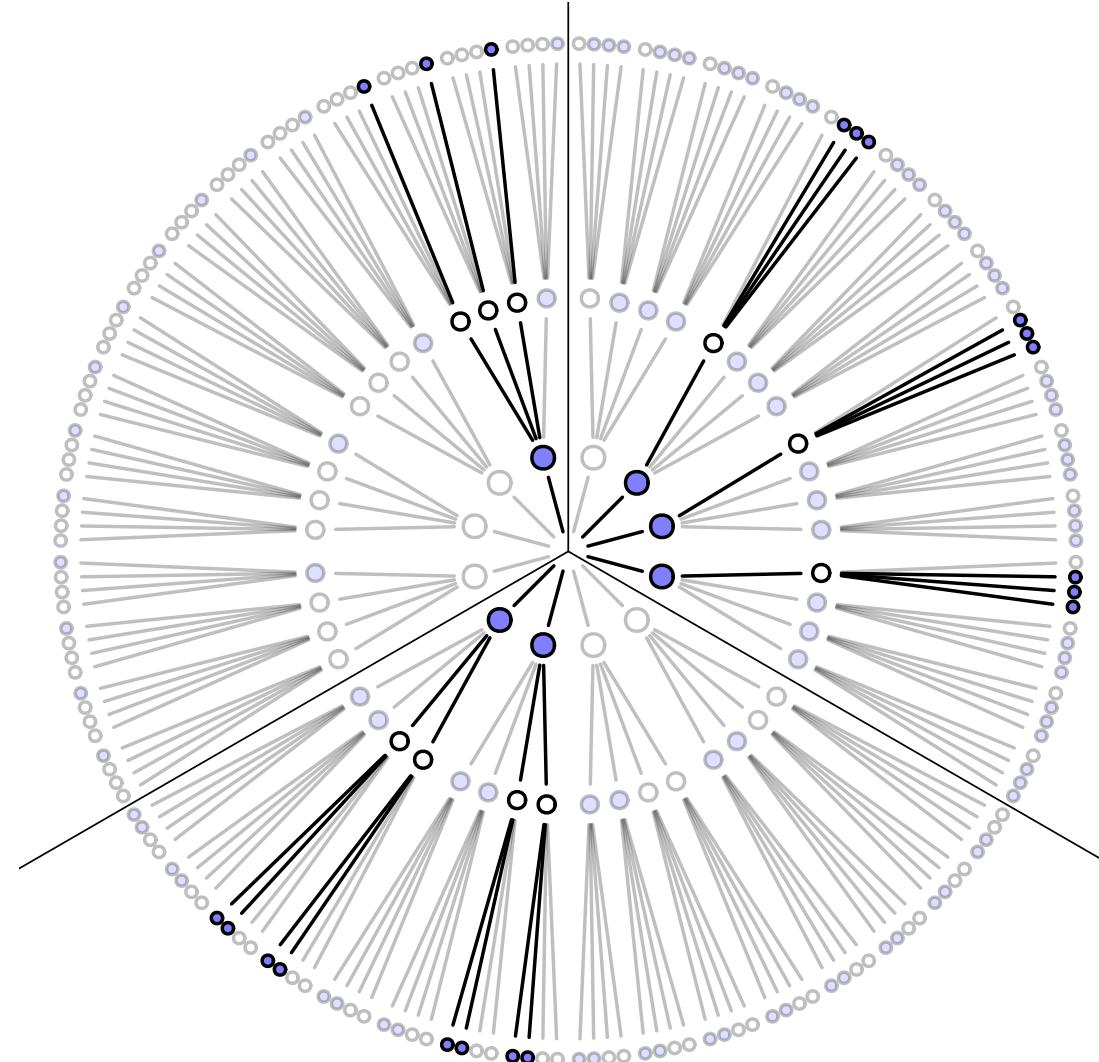
Hypothesis: ● ○ ○ ○

Data: ● ○ ●



# Enumerating possibilities

Under this hypothesis, 3 paths (out of  $4^3 = 64$ ) lead to the observed data. What about the other hypotheses?



# Comparing hypotheses

Given the data, the most *probable* hypothesis is the one that **maximises the number of possible ways** of obtaining the observed data.

Hypothesis	Ways to obtain the data
	$0 \times 4 \times 0 = 0$
	$1 \times 3 \times 1 = 3$
	$2 \times 2 \times 2 = 8$
	$3 \times 1 \times 3 = 9$
	$4 \times 0 \times 4 = 0$



# Evidence accumulation

In the previous case, we considered that all the hypotheses were equally probable a priori (according to the principle of indifference). However, we could have a priori information from our knowledge (of the characteristics of the bags of marbles, for example) or from previous data.

Let's assume we draw a new marble from the bag. How do we incorporate this new observation?



# Evidence accumulation

All we have to do is apply the same strategy as before, and update the last count by multiplying it by the new data. *Yesterday's posterior is today's prior* ([Lindley, 2000](#)).

Hypothesis	Ways to produce	Previous count	New count
● ● ● ●	0	0	$0 \times 0 = 0$
● ● ● ○	1	3	$3 \times 1 = 3$
● ● ○ ○	2	8	$8 \times 2 = 16$
● ● ○ ○	3	9	$9 \times 3 = 27$
● ● ● ●	4	0	$0 \times 4 = 0$



# Incorporating prior knowledge

Now let's suppose that an employee at the marble factory tells us that blue marbles are rare... This employee tells us that for every bag containing 3 blue marbles, they make two bags containing only two, and three bags containing only one. He also tells us that every bag contains at least one blue marble and one white marble...

Hypothesis	Previous count	Factory prior	New count
白白白白	0	0	$0 \times 0 = 0$
蓝白白白	3	3	$3 \times 3 = 9$
蓝蓝白白	16	2	$16 \times 2 = 32$
蓝蓝蓝白	27	1	$27 \times 1 = 27$
蓝蓝蓝蓝	0	0	$0 \times 0 = 0$



# From enumerations to probabilities

The probability of a hypothesis after observing certain data is proportional to the number of ways in which this hypothesis can produce the observed data, multiplied by its a priori probability.

$$\Pr(\text{hypothesis} \mid \text{data}) \propto \Pr(\text{data} \mid \text{hypothesis}) \times \Pr(\text{hypothesis})$$

To convert *plausibilities* to *probabilities*, all we have to do is standardise these plausibilities so that the sum of the plausibilities of all possible hypotheses is equal to 1.

$$\Pr(\text{hypothesis} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{hypothesis}) \times \Pr(\text{hypothesis})}{\text{sum of products}}$$



# From enumerations to probabilities

We define  $p$  as the proportion of blue marbles in the bag.

Hypothesis	$p$	Ways to produce the data	Probability
白白白白	0	0	0
蓝白白白	0.25	3	0.15
蓝蓝白白	0.5	8	0.40
蓝蓝蓝白	0.75	9	0.45
蓝蓝蓝蓝	1	0	0

```
1 ways <- c(0, 3, 8, 9, 0)
2 ways / sum(ways)
```

```
[1] 0.00 0.15 0.40 0.45 0.00
```



# Notations

- $\theta$  is a parameter or vector of parameters (e.g., the proportion of blue marbles).
- $p(x | \theta)$  the conditional probability of the data  $x$  given parameter  $\theta$  [ $p(x | \theta = \theta)$ ].
- $p(x | \theta)$  once the value of  $x$  is known, it is seen as the likelihood function of the parameter  $\theta$ . Note that this is not a valid probability distribution [ $p(x = x | \theta)$ ].
- $p(\theta)$  the prior probability of  $\theta$ .
- $p(\theta | x)$  the posterior probability of  $\theta$  (knowing  $x$ ).
- $p(x)$  the marginal probability of  $x$  (on  $\theta$ ) or “marginal likelihood”.

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)} = \frac{p(x | \theta)p(\theta)}{\sum_{\theta} p(x | \theta)p(\theta)} = \frac{p(x | \theta)p(\theta)}{\int_{\theta} p(x | \theta)p(\theta)dx} \propto p(x | \theta)p(\theta)$$



# Bayesian inference

In this framework, for each problem, we will follow 3 steps:

- Build the model (the story of the data): likelihood + prior.
- Update with data, compute (or approximate) the posterior.
- Evaluate the model, quality of fit, sensitivity, summarise results, readjust.

“

Bayesian inference is really just counting and comparing of possibilities [...] in order to make good inference about what actually happened, it helps to consider everything that could have happened ([McElreath, 2016](#)).

“

The first idea is that Bayesian inference is **reallocation of credibility across possibilities**. The second foundational idea is that the possibilities, over which we allocate credibility, are **parameter values** in meaningful mathematical models ([Kruschke, 2015](#)).



## Example, medical diagnosis (Gigerenzer et al., 2007)

- In women aged 40-50, with no family history and no symptoms, the probability of developing breast cancer is around 0.008.
- Known properties of mammography:
  - If a woman has breast cancer, the probability of having a positive result is 0.90 (true positive).
  - If a woman does not have breast cancer, the probability of having a positive result is 0.07 (false positive).
- Suppose a woman has a mammogram and the test is positive. What should be **inferred**? What is the probability that this woman has breast cancer?



# Maximum likelihood estimation

- A general approach to parameter estimation.
- The parameters **govern** the data, the data **depend** on the parameters.
  - Knowing certain parameter values, we can calculate the **conditional probability** of the observed data.
  - The result of the mammogram (i.e., the data) depends on the presence/absence of breast cancer (i.e., the parameter).
- The **maximum likelihood** approach asks the question: “Which values of the parameter make the observed data the most probable?
- Specify the conditional probability of the data  $p(x | \theta)$ .
- When we consider it as a function of  $\theta$ , we talk about **likelihood**:  $\mathcal{L}(\theta | x) = p(X = x | \theta)$ .
- The maximum likelihood approach therefore consists of maximising this function, using the (known) values of  $x$ .



# Conditional probability

- If a woman has breast cancer, the probability of obtaining a positive result is .90.
  - $\Pr(\text{Mam}=+ \mid \text{Cancer}=+) = 0.90$
  - $\Pr(\text{Mam}=- \mid \text{Cancer}=+) = 0.10$
- If a woman does not have breast cancer, the probability of obtaining a positive result is .07.
  - $\Pr(\text{Mam}=+ \mid \text{Cancer}=-) = 0.07$
  - $\Pr(\text{Mam}=- \mid \text{Cancer}=-) = 0.93$



# Medical diagnosis, maximum likelihood

- A woman gets a mammogram, the result is positive...
  - $\Pr(\text{Mam}=+ | \text{Cancer}=+) = 0.90$
  - $\Pr(\text{Mum}=+ | \text{Cancer}=-) = 0.07$
- Maximum likelihood: what is the value of Cancer that **maximises** Mam=+?

This approach leads to the conclusion that cancer is present (because it maximises the probability of a positive mammogram)...



Wait a minute...



Ladislas Nalborczyk - IBSM2023

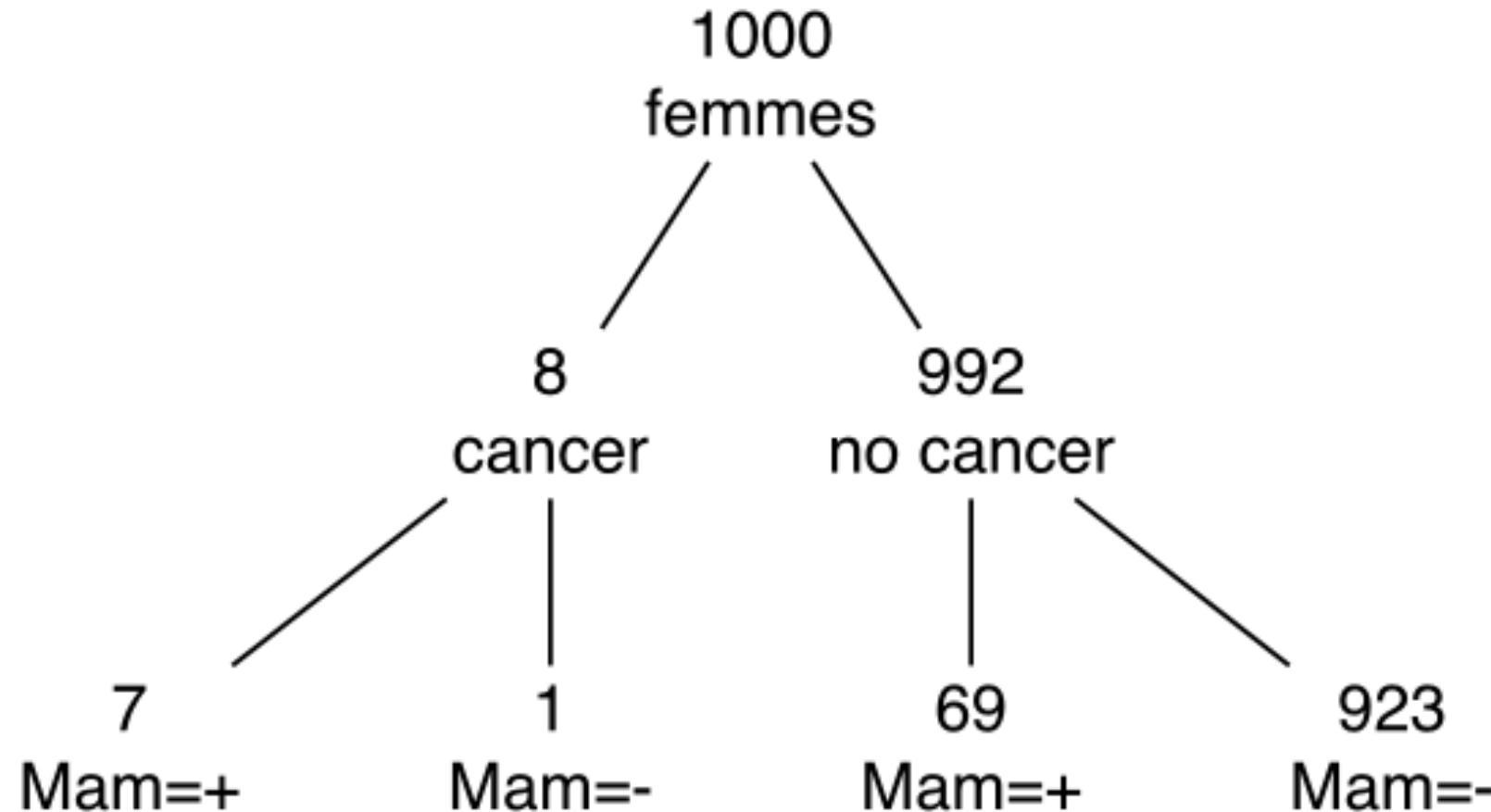


# Medical diagnosis, working with frequencies

- Consider 1000 women aged between 40 and 50, with no family history of cancer and no symptoms.
  - 8 out of 1000 women have cancer
- A mammogram is performed
  - Of the 8 women with cancer, 7 will have a positive result
  - Of the remaining 992 women, 69 will have a positive result
- One woman has a mammogram, the result is positive
- What should we infer?



# Medical diagnosis, working with frequencies



$$\Pr(\text{Cancer}=+ \mid \text{Mam}=+) = \frac{7}{7 + 69} = \frac{7}{76} \approx 0.09$$



# Medical diagnosis, Bayes' theorem

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

$p(\theta)$  represents the prior probability of  $\theta$ : what we know about  $\theta$  before observing the data. In this case:  $\text{Pr}(\text{Cancer}=+) = 0.008$  and  $\text{Pr}(\text{Cancer}=-) = 0.992$ .

```
1 prior <- c(0.008, 0.992)
```



# Medical diagnosis, Bayes' theorem

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

$p(x | \theta)$  represents the conditional probability of the data  $x$  knowing the parameter  $\theta$ , also known as the likelihood function of the parameter  $\theta$ .

```
1 like <- rbind(c(0.9, 0.1), c(0.07, 0.93)) %>% data.frame
2 colnames(like) <- c("Mam+", "Mam-")
3 rownames(like) <- c("Cancer+", "Cancer-")
4 like
```

	Mam+	Mam-
Cancer+	0.90	0.10
Cancer-	0.07	0.93



# Medical diagnosis, Bayes' theorem

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

$p(x)$  the marginal probability of  $x$  (over  $\theta$ ). This is a constant used to normalise the distribution (the “sum of products” from the previous example).

$$p(x) = \sum_{\theta} p(x | \theta)p(\theta)$$

```
1 (marginal <- sum(like$"Mam+" * prior) )
```



```
[1] 0.07664
```



# Medical diagnosis, Bayes' theorem

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

$p(\theta | x)$  the posterior probability of  $\theta$  given  $x$ , that is, what we know about  $\theta$  after seeing  $x$ .

```
1 (posterior <- (like$"Mam+" * prior ) / marginal )  
[1] 0.09394572 0.90605428
```



# Bayesian inference as probabilistic knowledge updating

Before the mammogram, the probability of a woman drawn at random having breast cancer was  $\text{Pr}(\text{Cancer}=+) = 0.008$  (prior). After a positive result, this probability became  $\text{Pr}(\text{Cancer}=+ | \text{Mam}=+) = 0.09$  (posterior). These probabilities are expressions of our *knowledge*. After a positive mammogram, we still think it's "very unlikely" to have cancer, but this probability has changed considerably compared with "before the test".

"

A Bayesianly justifiable analysis is one that treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes' theorem ([Rubin, 1984](#)).



# Beta-Binomial model



# Bernoulli law

Applies to all situations where the data generation process can only result in two mutually exclusive outcomes (e.g., a coin toss). At each trial, if we assume that  $\Pr(\text{heads}) = \theta$ , then  $\Pr(\text{tails}) = 1 - \theta$ .

Since Bernoulli, we know how to compute the probability of the result of a coin toss, as long as we know the coin's bias  $\theta$ . Let's assume that  $Y = 0$  when you get tails, and that  $Y = 1$  when you get heads. Then  $Y$  is distributed according to a Bernoulli distribution:

$$p(y | \theta) = \Pr(Y = y | \theta) = \theta^y(1 - \theta)^{(1-y)}$$

If we replace  $y$  by 0 or 1, we come back to our previous observations:

$$\Pr(Y = 1 | \theta) = \theta^1(1 - \theta)^{(1-1)} = \theta \times 1 = \theta$$

$$\Pr(Y = 0 | \theta) = \theta^0(1 - \theta)^{(1-0)} = 1 \times (1 - \theta) = 1 - \theta$$



# Bernoulli process

If we have a series of independent and identically distributed throws  $\{Y_i\}$  (i.e., each throw has a Bernoulli distribution of probability distribution with probability  $\theta$ ), the set of throws can be described by a **binomial distribution**.

For example, suppose we have the following sequence of five throws: Tails, Tails, Tails, Heads, Heads. We can recode this sequence into  $\{0, 0, 0, 1, 1\}$ .

Reminder: The probability of each 1 is  $\theta$  and the probability of each 0 is  $1 - \theta$ .

What is the probability of getting 2 heads out of 5 throws?



# Bernoulli process

Knowing that the trials are independent of each other, the probability of obtaining this sequence is  $(1 - \theta) \times (1 - \theta) \times (1 - \theta) \times \theta \times \theta$ , that is:  $\theta^2(1 - \theta)^3$ .

We can generalise this result for a sequence of  $n$  throws and  $y$  “successes”:

$$\theta^y(1 - \theta)^{n-y}$$

So far, we have only considered a single sequence resulting in 2 successes for 5 throws, but there are many sequences that can result in 2 successes for 5 throws (e.g.,  $\{0, 0, 1, 0, 1\}$ ,  $\{0, 1, 1, 0, 0\}$ )...



# Binomial coefficient

The **binomial coefficient** allows computing the number of possible combinations resulting in  $y$  successes for  $n$  throws in the following way (read “ $y$  among  $n$ ” or “number of combinations of  $y$  among  $n$ ”)<sup>1</sup>:

$$\binom{n}{y} = C_y^n = \frac{n!}{y!(n-y)!}$$

For instance for  $y = 1$  and  $n = 3$ , we know there are 3 possible combinations:  $\{0, 0, 1\}$ ,  $\{0, 1, 0\}$ ,  $\{1, 0, 0\}$ . We can check this by applying the formula above.

$$\binom{3}{1} = C_1^3 = \frac{3!}{1!(3-1)!} = \frac{3 \times 2 \times 1}{1 \times 2 \times 1} = \frac{6}{2} = 3$$

```
1 # computing the total number of possible configurations in R
2 choose(n = 3, k = 1)
```

```
[1] 3
```

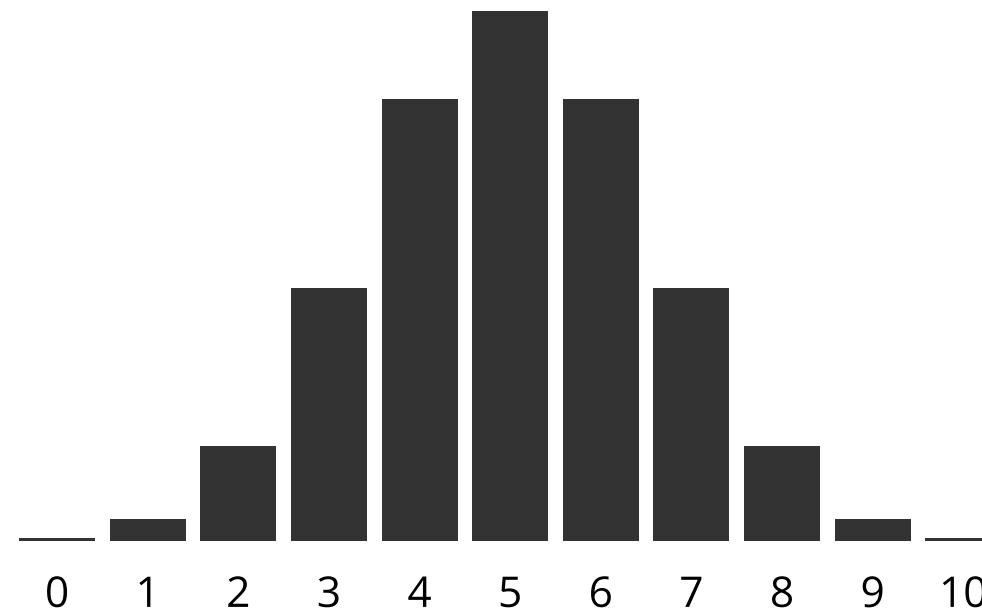
1. The factorial of a non-negative integer  $n$ , denoted by  $n!$ , is the product of all positive integers less than



# Binomial distribution

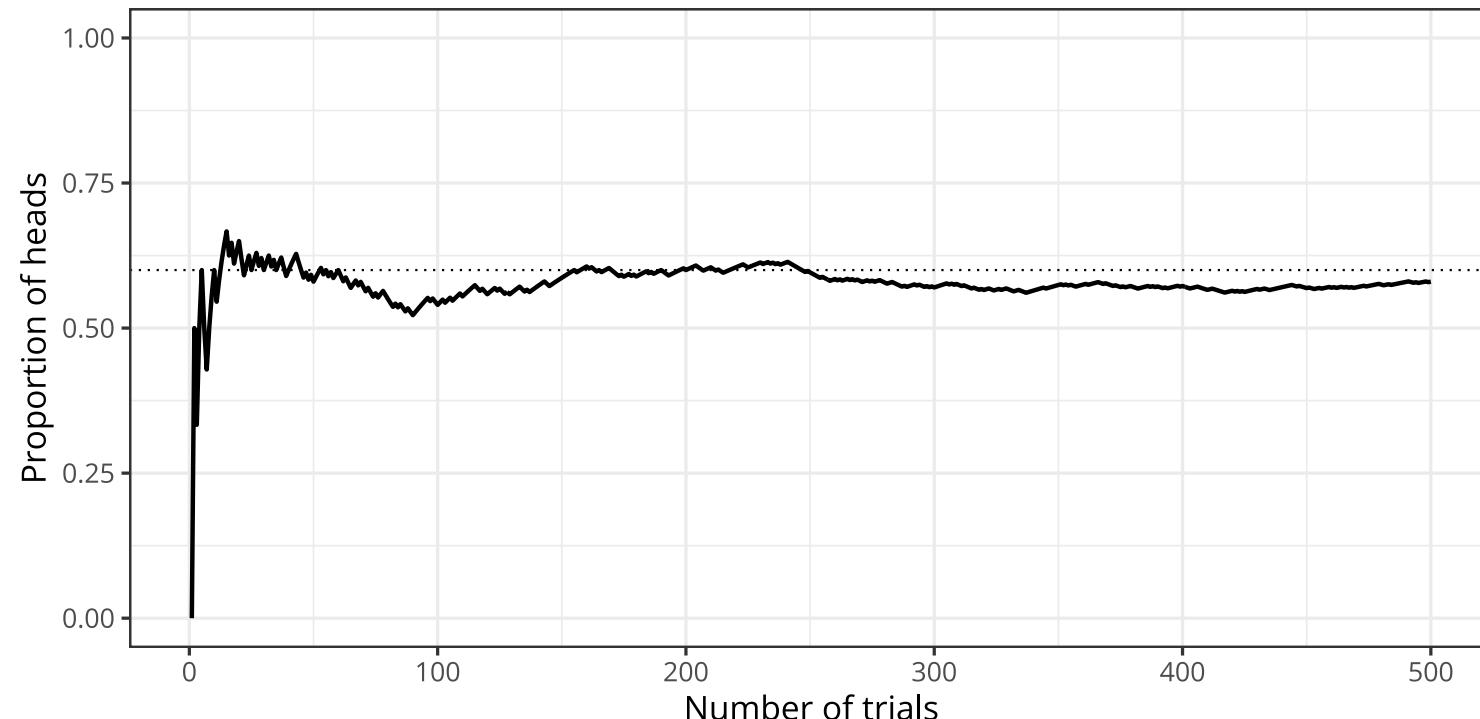
$$p(y | \theta) = \Pr(Y = y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$

The binomial distribution allows us to calculate the probability of obtaining  $y$  successes on  $n$  trials, for a given  $\theta$ . Example of the binomial distribution for an unbiased coin ( $\theta = 0.5$ ), indicating the probability of obtaining  $n$  faces in 10 throws (in R: `dbinom(x = 0:10, size = 10, prob = 0.5)`).



# Sampling binary data

```
1 library(tidyverse)
2 set.seed(666) # for reproducibility
3
4 sample(x = c(0, 1), size = 500, prob = c(0.4, 0.6), replace = TRUE) %>% # theta = 0.6
5     data.frame() %>%
6     mutate(x = seq_along(.), y = cummean(.)) %>%
7     ggplot(aes(x = x, y = y)) +
8     geom_line(lwd = 1) +
9     geom_hline(yintercept = 0.6, lty = 3) +
10    labs(x = "Number of trials", y = "Proportion of heads") +
11    ylim(0, 1)
```



# Defining the model (likelihood)

Likelihood function:

- We consider  $y$  to be the number of successes.
- We consider the number of observations  $n$  to be a **constant**.
- We consider  $\theta$  to be the **parameter** of our model (i.e., the probability of success).

The likelihood function is written as:

$$\mathcal{L}(\theta | y, n) = p(y | \theta, n) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \propto \theta^y (1 - \theta)^{n-y}$$



# Likelihood versus probability

A coin with a bias  $\theta$  is tossed again (where  $\theta$  represents the probability of getting heads). This coin is tossed twice and we obtain a Heads and a Tails.

The probability of observing a Heads on two coin tosses **as a function of** different values of  $\theta$  can be calculated as follows:

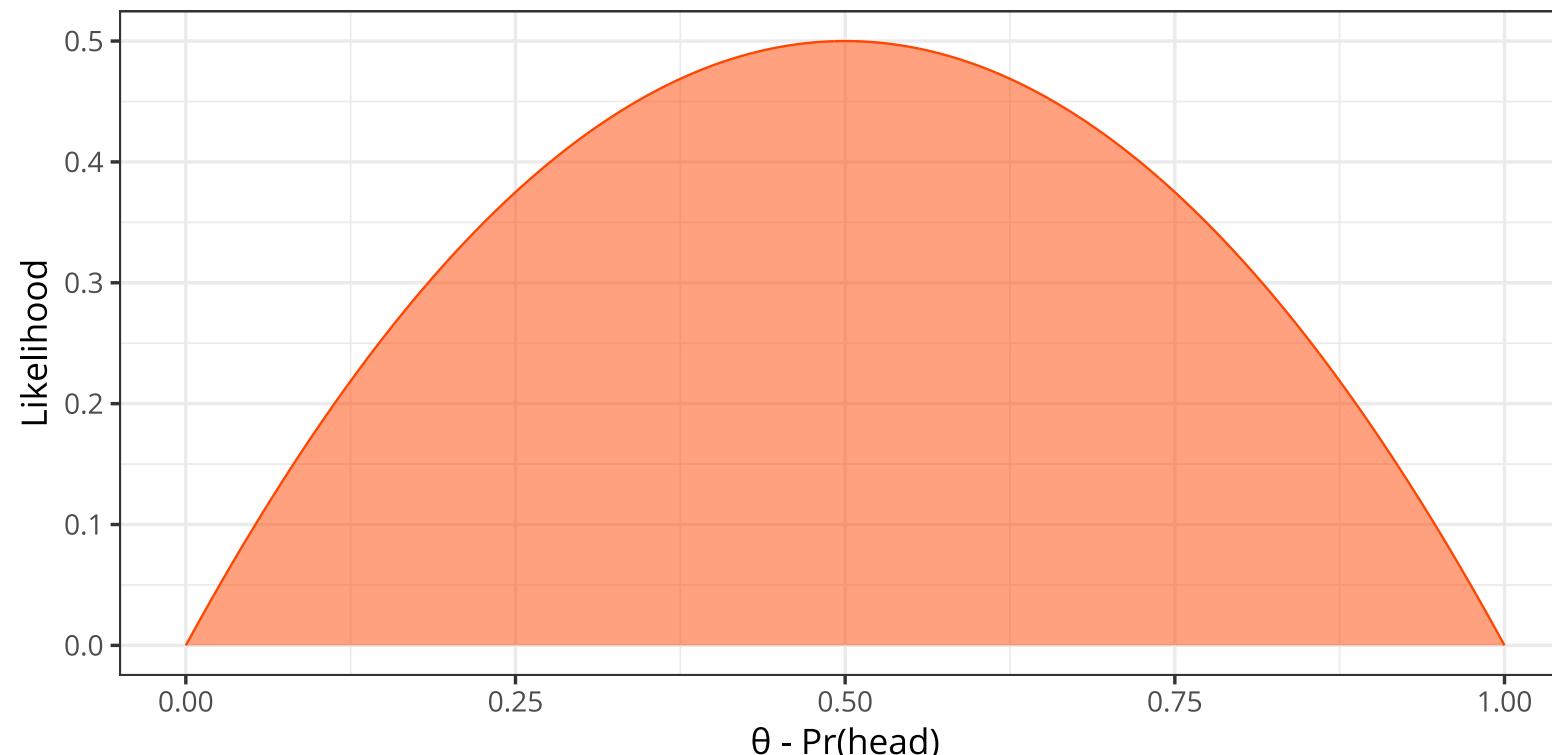
$$\begin{aligned}\Pr(H, T \mid \theta) + \Pr(T, H \mid \theta) &= 2 \times \Pr(T \mid \theta) \times \Pr(H \mid \theta) \\ &= \theta(1 - \theta) + \theta(1 - \theta) \\ &= 2\theta(1 - \theta)\end{aligned}$$

This probability is defined for a fixed data set and a variable value of  $\theta$ . This function can be represented visually.



# Likelihood versus probability

```
1 # Graphical representation of the likelihood function for y = 1 and n = 2
2
3 y <- 1 # number of heads
4 n <- 2 # number of trials
5
6 data.frame(theta = seq(from = 0, to = 1, length.out = 1e3) ) %>%
7   mutate(likelihood = dbinom(x = y, size = n, prob = theta) ) %>%
8   ggplot(aes(x = theta, y = likelihood) ) +
9   geom_area(color = "orangered", fill = "orangered", alpha = 0.5) +
10  labs(x = expression(paste(theta, " - Pr(head)")), y = "Likelihood")
```



# Likelihood versus probability

If we calculate the area under the curve of this function, we get:

$$\int_0^1 2\theta(1 - \theta)d\theta = \frac{1}{3}$$

```
1 f <- function(theta) {2 * theta * (1 - theta) }
2 integrate(f = f, lower = 0, upper = 1)
```

0.3333333 with absolute error < 3.7e-15

When we vary  $\theta$ , the likelihood function **is not** a valid probability distribution (i.e., its integral is not equal to 1). We use the term **likelihood** to distinguish this type of function from probability density functions. We use the following notation to emphasise the fact that the likelihood function is a function of  $\theta$ , and that the data are fixed:  $\mathcal{L}(\theta | \text{data}) = p(\text{data} | \theta)$ .



# Likelihood versus probability

Likelihood versus probability for two coin tosses

theta	Number of Heads (y)			Total
	0	1	2	
0	1.00	0.00	0.00	1
0.2	0.64	0.32	0.04	1
0.4	0.36	0.48	0.16	1
0.6	0.16	0.48	0.36	1
0.8	0.04	0.32	0.64	1
1	0.00	0.00	1.00	1
Total	2.20	1.60	2.20	

Note that the likelihood of  $\theta$  for a particular data item is equal to the probability of the data for this value of  $\theta$ . However, the *distribution* of these likelihoods (in columns) is not a probability distribution. In a usual Bayesian analysis, **the data are considered fixed** and the value of  $\theta$  is considered a **random variable**.



# Defining the model (prior)

How can we define a prior for  $\theta$ ?

**Semantic aspect** → *the prior should be able to represent:*

- An absence of information
- Knowledge of previous observations concerning this coin
- A level of uncertainty concerning these previous observations

**Mathematical aspect** → *for a fully analytical solution:*

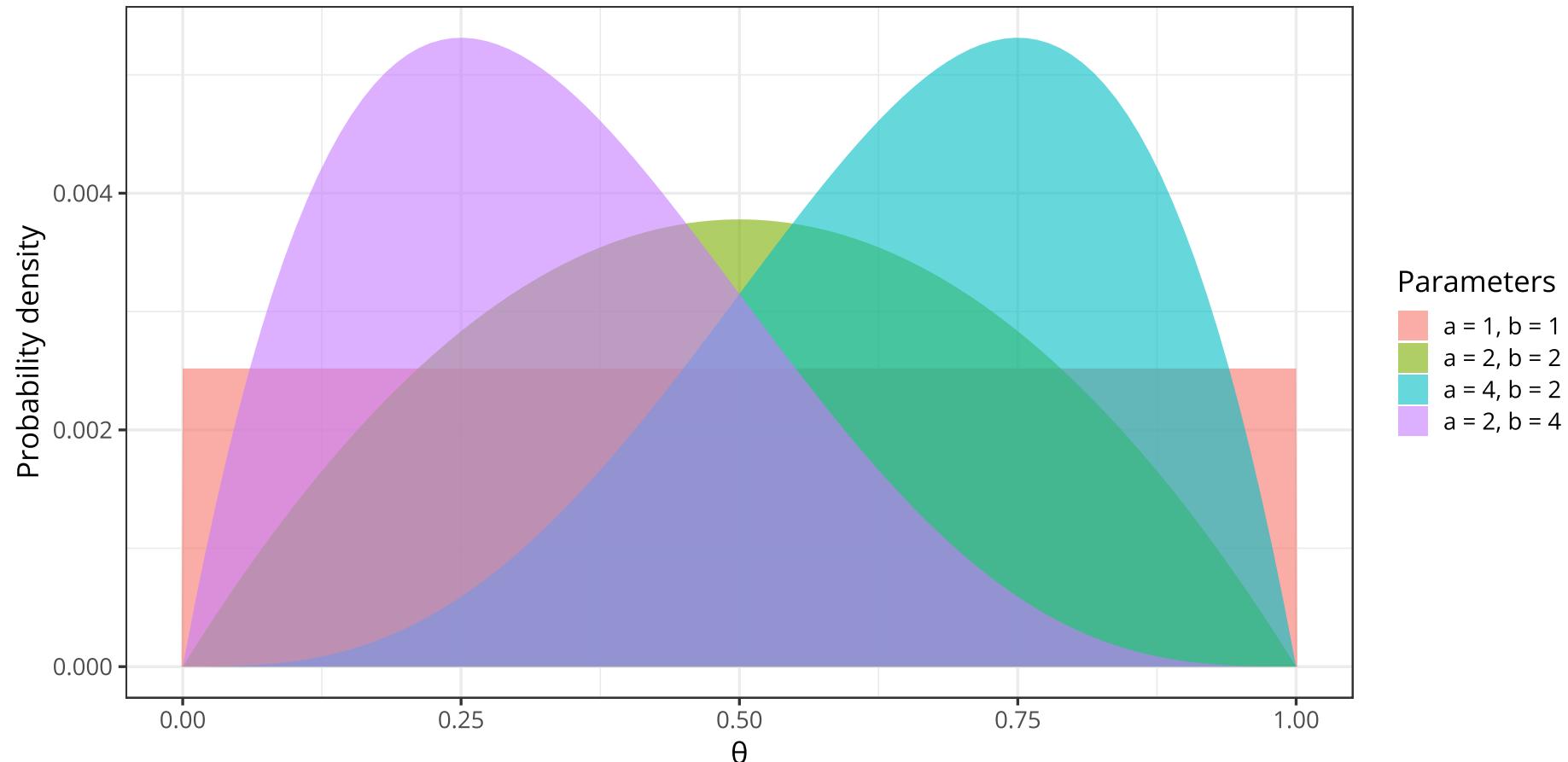
- The prior and posterior distributions must have the same form
- The marginal likelihood must be calculable analytically



# Beta distribution

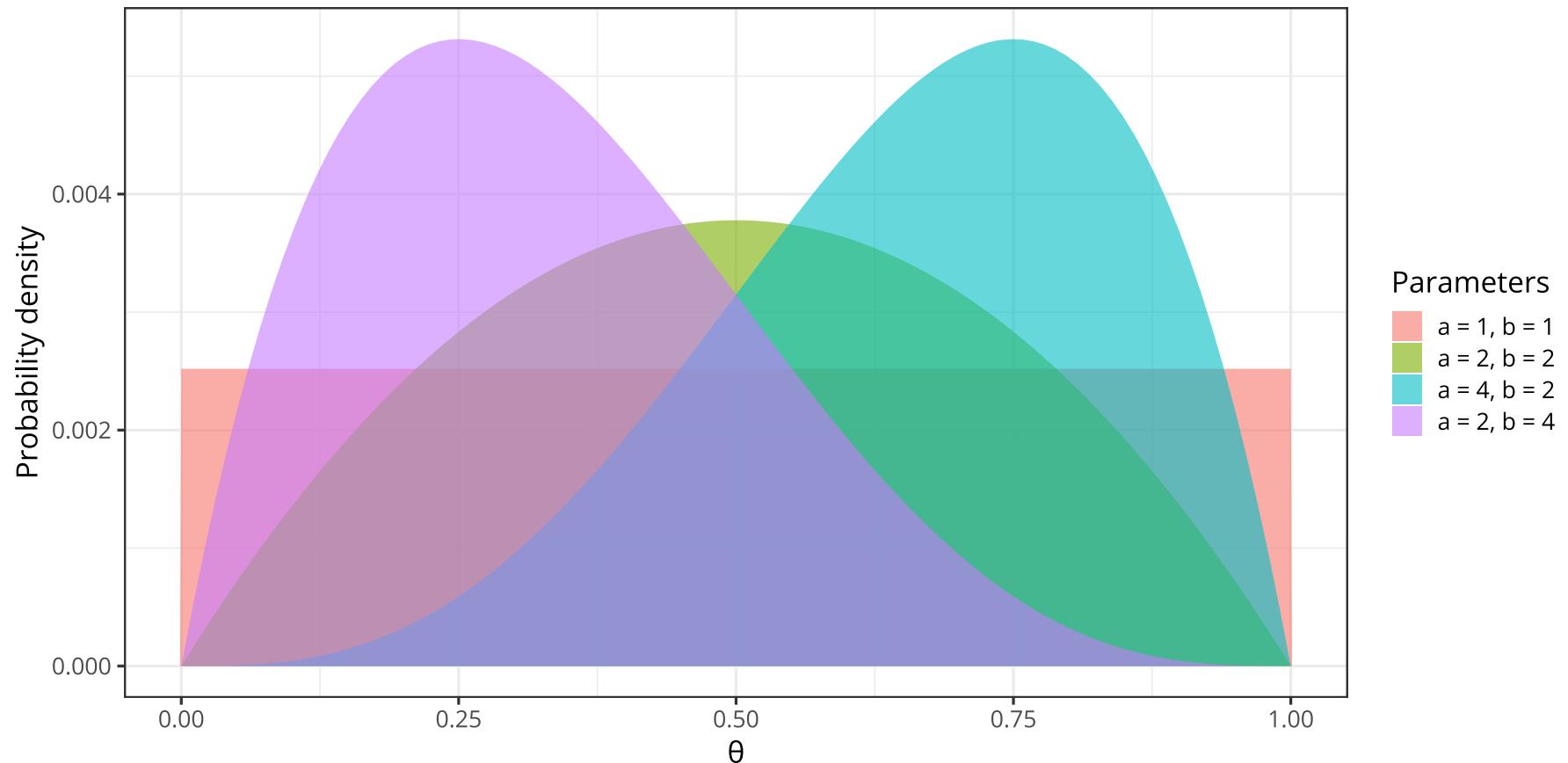
$$\begin{aligned}
 p(\theta | a, b) &= \text{Beta}(\theta | a, b) \\
 &= \theta^{a-1}(1-\theta)^{b-1}/B(a, b) \\
 &\propto \theta^{a-1}(1-\theta)^{b-1}
 \end{aligned}$$

where  $a$  and  $b$  are two parameters such that  $a \geq 0, b \geq 0$ , and  $B(a, b)$  is a normalisation constant.



# Interpretation of parameters

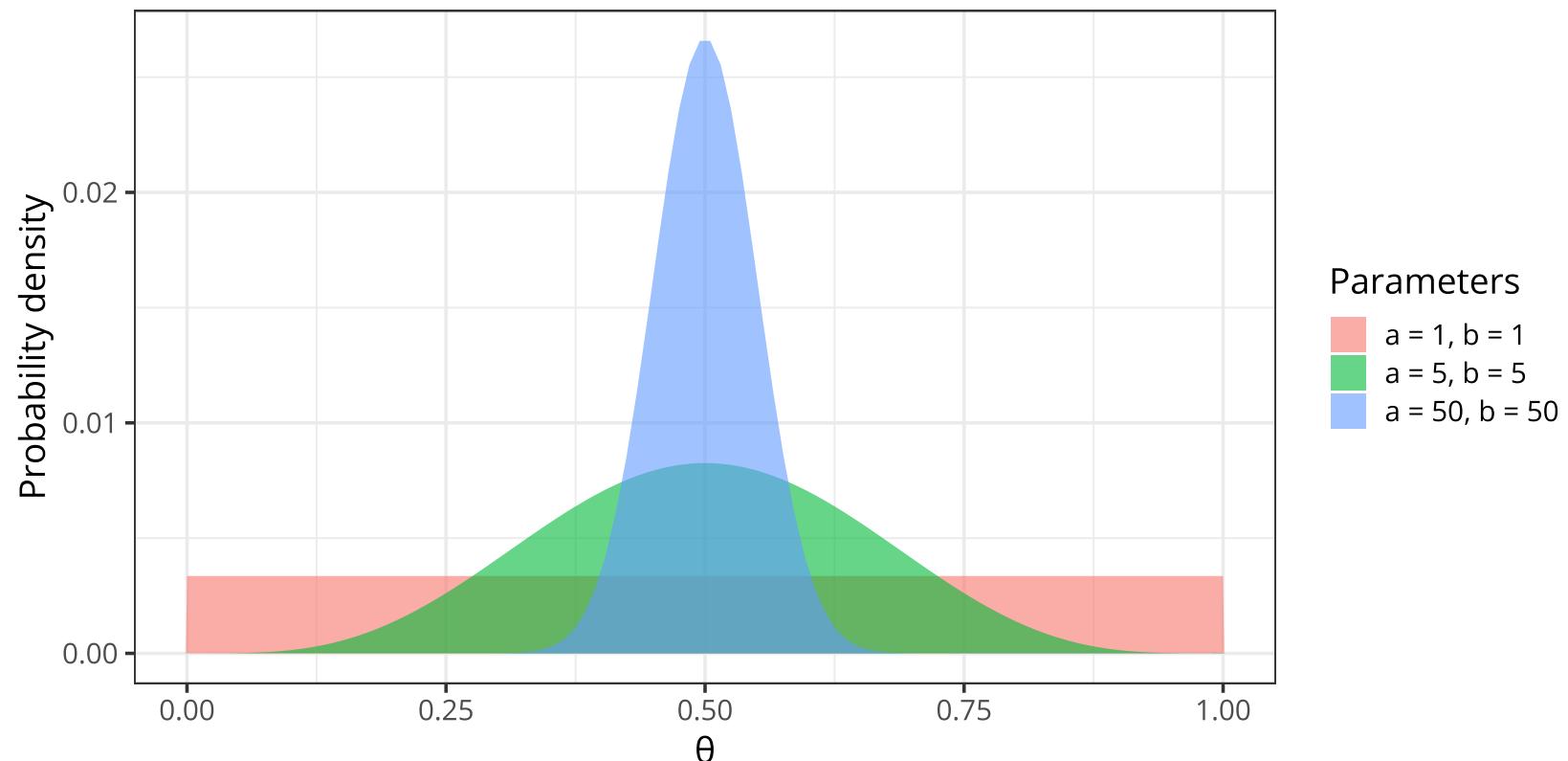
- The absence of prior knowledge can be expressed by setting  $a = b = 1$  (orange distribution).
- A prior in favour of an absence of bias can be expressed by setting  $a = b > 1$  (green distribution).
- A bias in favour of *Heads* can be expressed by setting  $a > b$  (blue distribution).
- A bias in favour of *Tails* can be expressed by setting  $a < b$  (purple distribution).



# Interpretation of parameters

The level of certainty increases with the sum  $\kappa = a + b$ .

- No idea where the coin comes from:  $a = b = 1 \rightarrow \text{flat prior}$ .
- While waiting for the experiment to begin, the coin was tossed 10 times and we observed 5 “Heads”:  $a = b = 5 \rightarrow \text{weakly informative prior}$ .
- The coin comes from the Bank of France:  $a = b = 50 \rightarrow \text{strongly informative prior}$ .



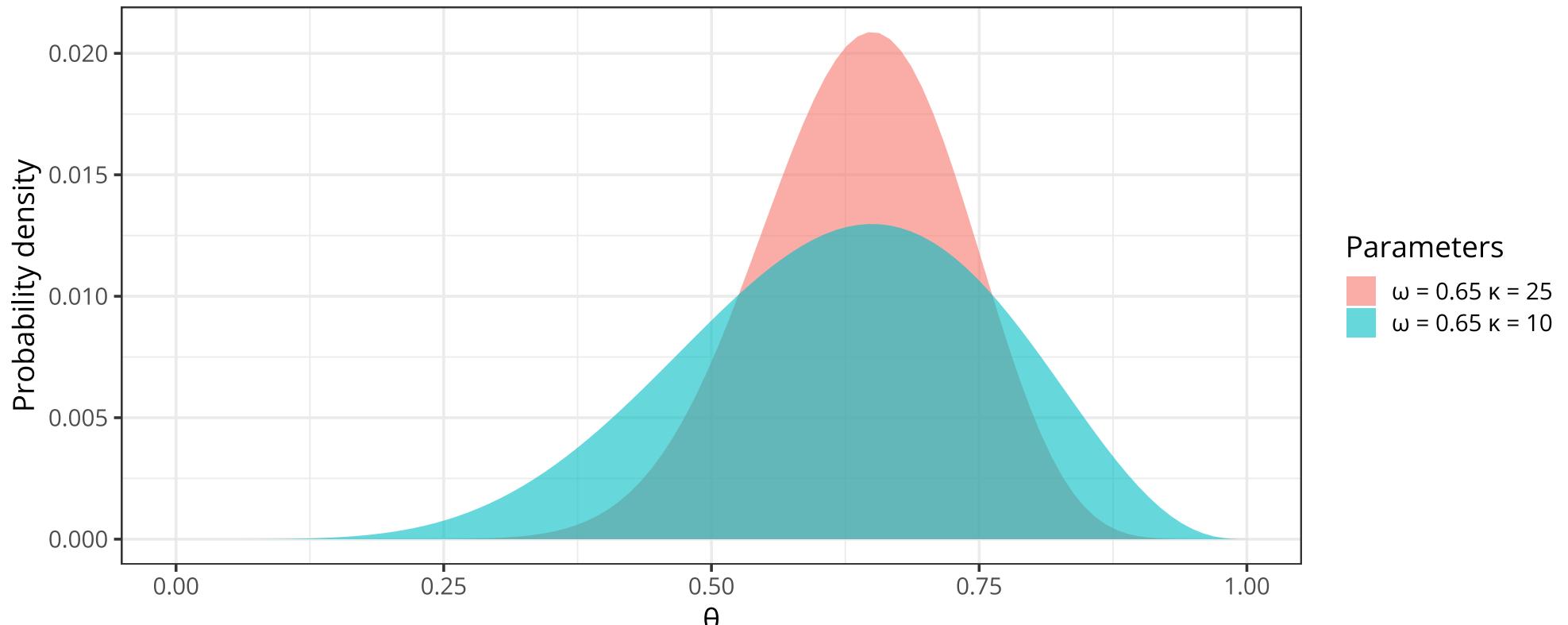
# Interpretation of parameters

Suppose we have an estimate of the most likely value of the parameter  $\theta$ . We can re-parametrise the Beta distribution as a function of the mode  $\omega$  and the level of certainty  $\kappa$ :

$$\begin{aligned} a &= \omega(\kappa - 2) + 1 \\ b &= (1 - \omega)(\kappa - 2) + 1 \quad \text{for } \kappa > 2 \end{aligned}$$

If  $\omega = 0.65$  and  $\kappa = 25$ , then  $p(\theta) = \text{Beta}(\theta | 15.95, 9.05)$ .

If  $\omega = 0.65$  and  $\kappa = 10$  then  $p(\theta) = \text{Beta}(\theta | 6.2, 3.8)$ .





# Conjugate prior

Formally, if  $\mathcal{F}$  is a class of sampling distributions  $p(y | \theta)$ , and  $\mathcal{P}$  is a class of prior distributions for  $\theta$ , then the class  $\mathcal{P}$  is **conjugate** to  $\mathcal{F}$  if:

$$p(\theta | y) \in \mathcal{P} \text{ for all } p(\cdot | \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

(p.35, [Gelman et al., 2013](#)). In other words, a prior is called a *conjugate prior* if, when converted to a posterior by being multiplied by the likelihood, it keeps the same form. In our case, the Beta prior is a conjugate prior for the Binomial likelihood, because the posterior is a Beta distribution as well.



# Analytical derivation of the posterior distribution

Assume a prior defined by:  $p(\theta | a, b) = \text{Beta}(a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$

Given a likelihood function associated with  $y$  “Heads” for  $n$  throws:

$$p(y | n, \theta) = \text{Bin}(y | n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

Then (omitting the normalisation constants):

$$\begin{aligned} p(\theta | y, n) &\propto p(y | n, \theta) p(\theta) && \text{Bayes theorem} \\ &\propto \text{Bin}(y | n, \theta) \text{Beta}(\theta | a, b) \\ &\propto \theta^y (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} && \text{Application of previous formulas} \\ &\propto \theta^{y+a-1} (1-\theta)^{n-y+b-1} && \text{Grouping powers of identical terms} \end{aligned}$$

Here, we have ignored the constants which do not depend on  $\theta$  (i.e., the number of combinations in the binomial likelihood function and the Beta function  $B(a, b)$  in the Beta prior).<sup>1</sup> Taking them into account, we obtain a Beta posterior distribution of the following form:

$$p(\theta | y, n) = \text{Beta}(y + a, n - y + b)$$



# An example to help you digest

We observe  $y = 7$  correct answers out of  $n = 10$  questions. We choose a prior  $\text{Beta}(1, 1)$ , that is, a uniform prior on  $[0, 1]$ . This prior is equivalent to a priori knowledge of 0 successes and 0 failures (i.e., a flat prior).

The posterior distribution is given by:

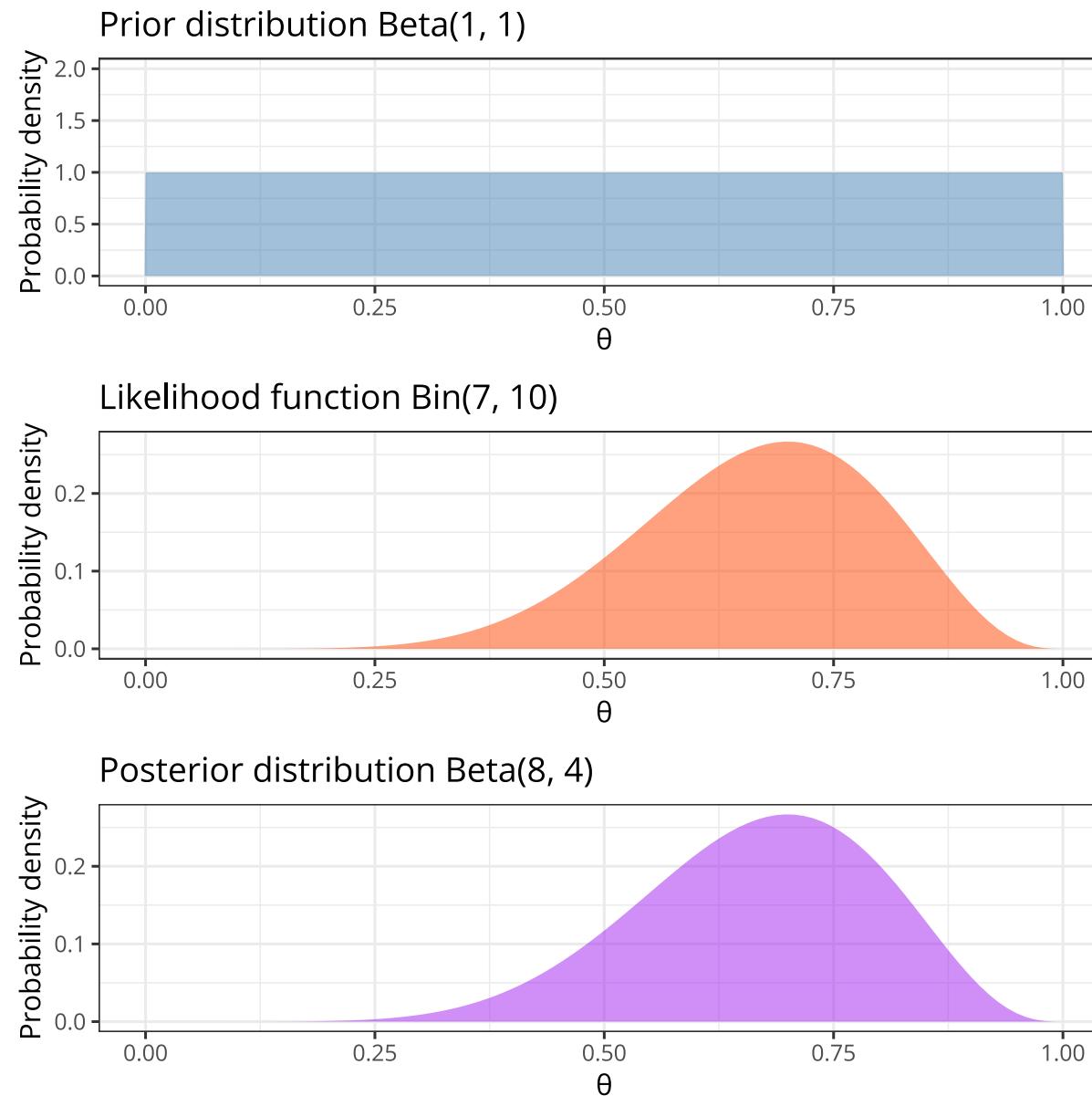
$$\begin{aligned} p(\theta | y, n) &\propto p(y | n, \theta) p(\theta) \\ &\propto \text{Bin}(7 | 10, \theta) \text{Beta}(\theta | 1, 1) \\ &= \text{Beta}(y + a, n - y + b) \\ &= \text{Beta}(8, 4) \end{aligned}$$

The mean of the posterior distribution is given by:

$$\underbrace{\frac{y + a}{n + a + b}}_{\text{posterior}} = \underbrace{\frac{y}{n}}_{\text{data}} \underbrace{\frac{n}{n + a + b}}_{\text{weight}} + \underbrace{\frac{a}{a + b}}_{\text{prior}} \underbrace{\frac{a + b}{n + a + b}}_{\text{weight}}$$

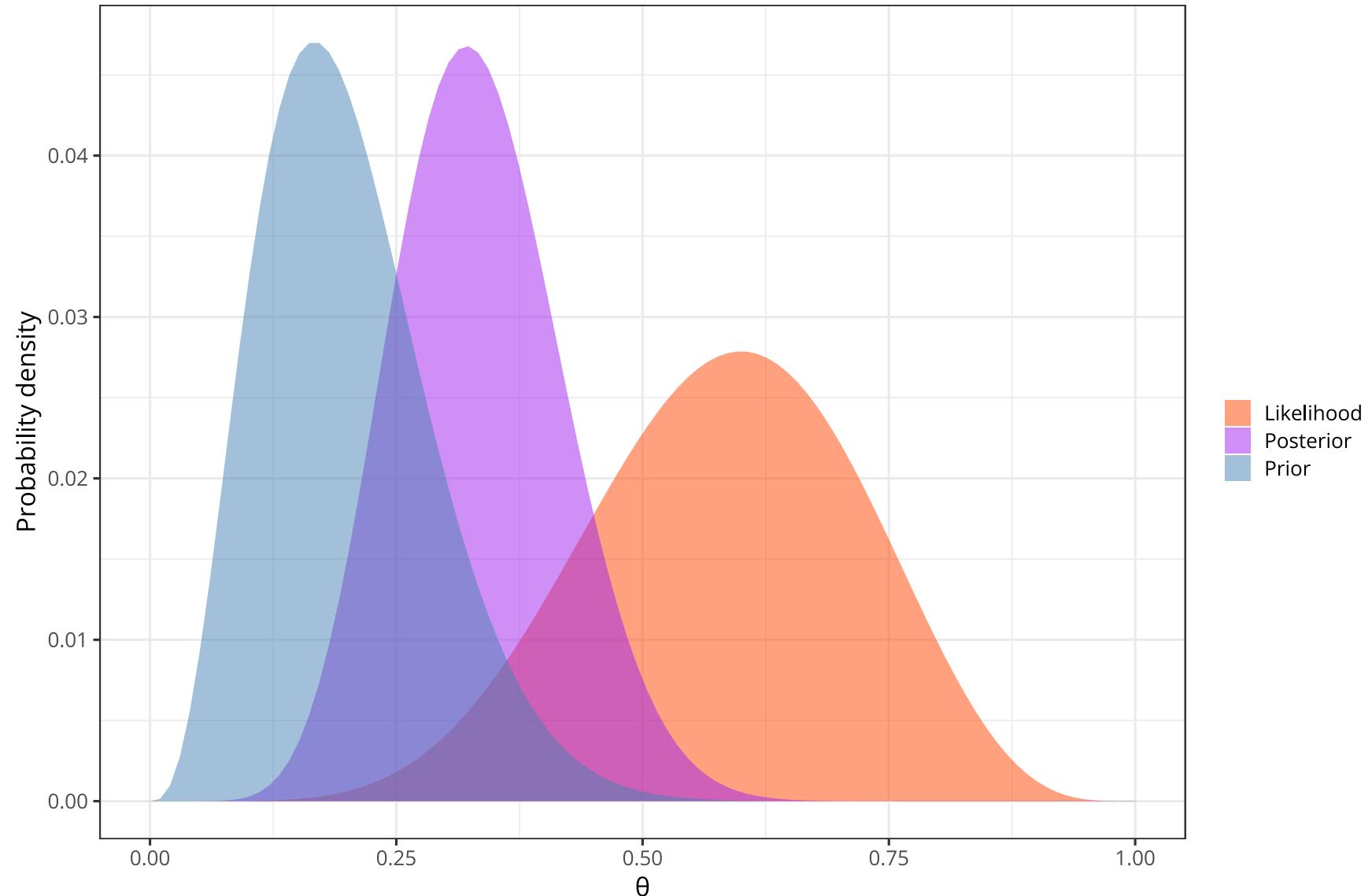


# An example to help you digest



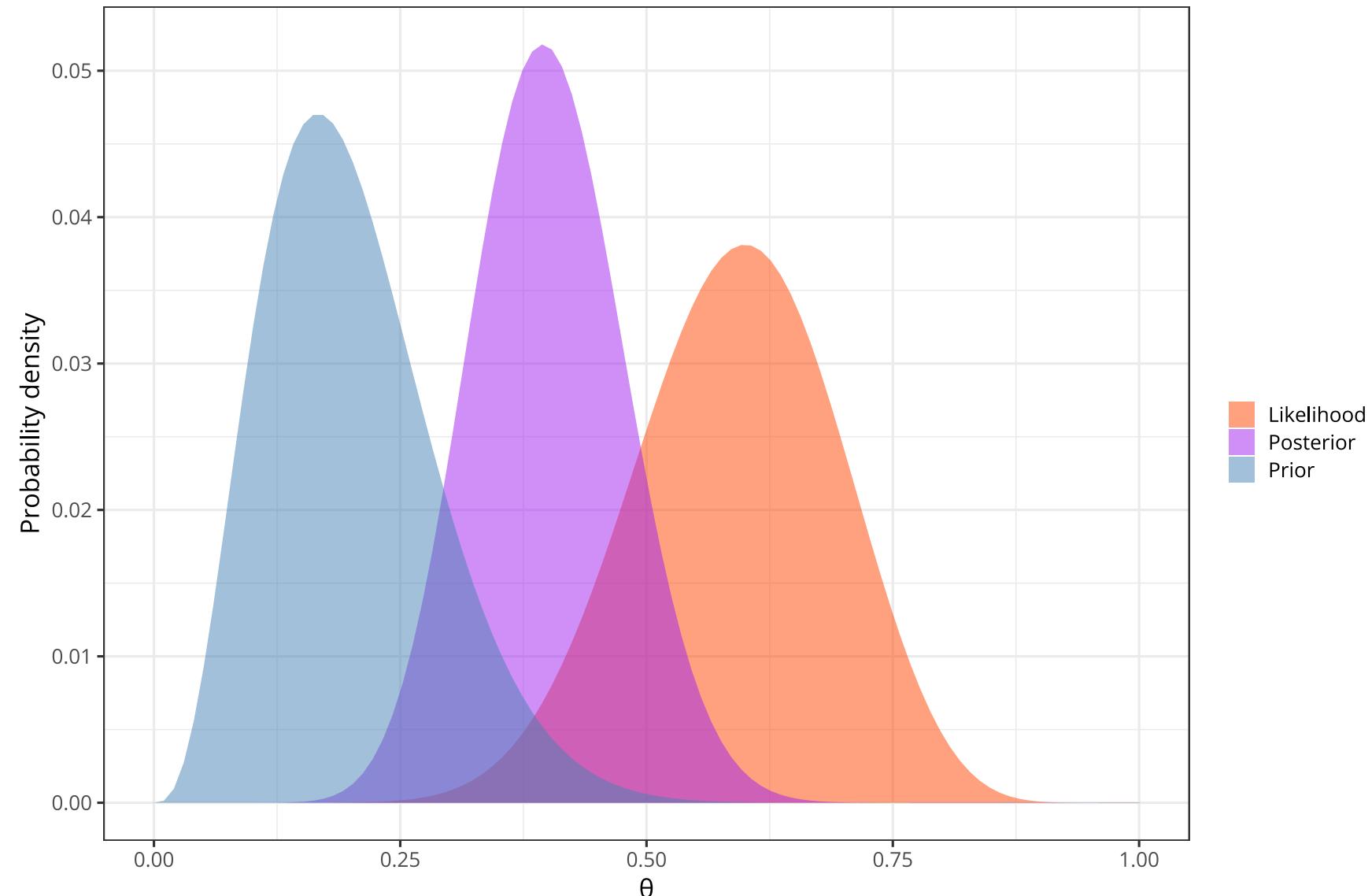
# Influence of prior on posterior distribution

Case where  $n < a + b$ , ( $n = 10$ ,  $a = 4$ ,  $b = 16$ ).



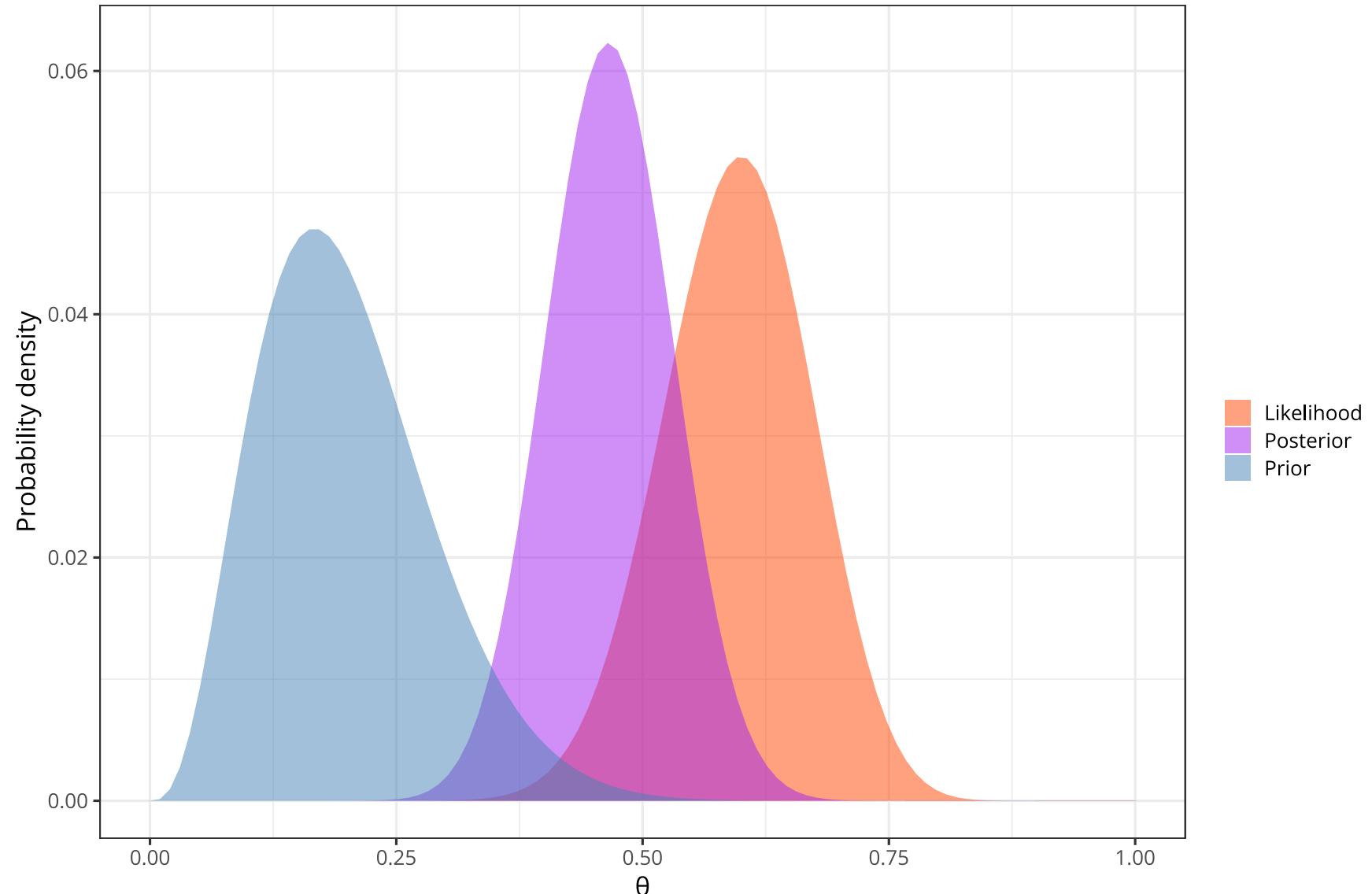
# Influence of prior on posterior distribution

Case where  $n = a + b$ , ( $n = 20$ ,  $a = 4$ ,  $b = 16$ ).



# Influence of prior on posterior distribution

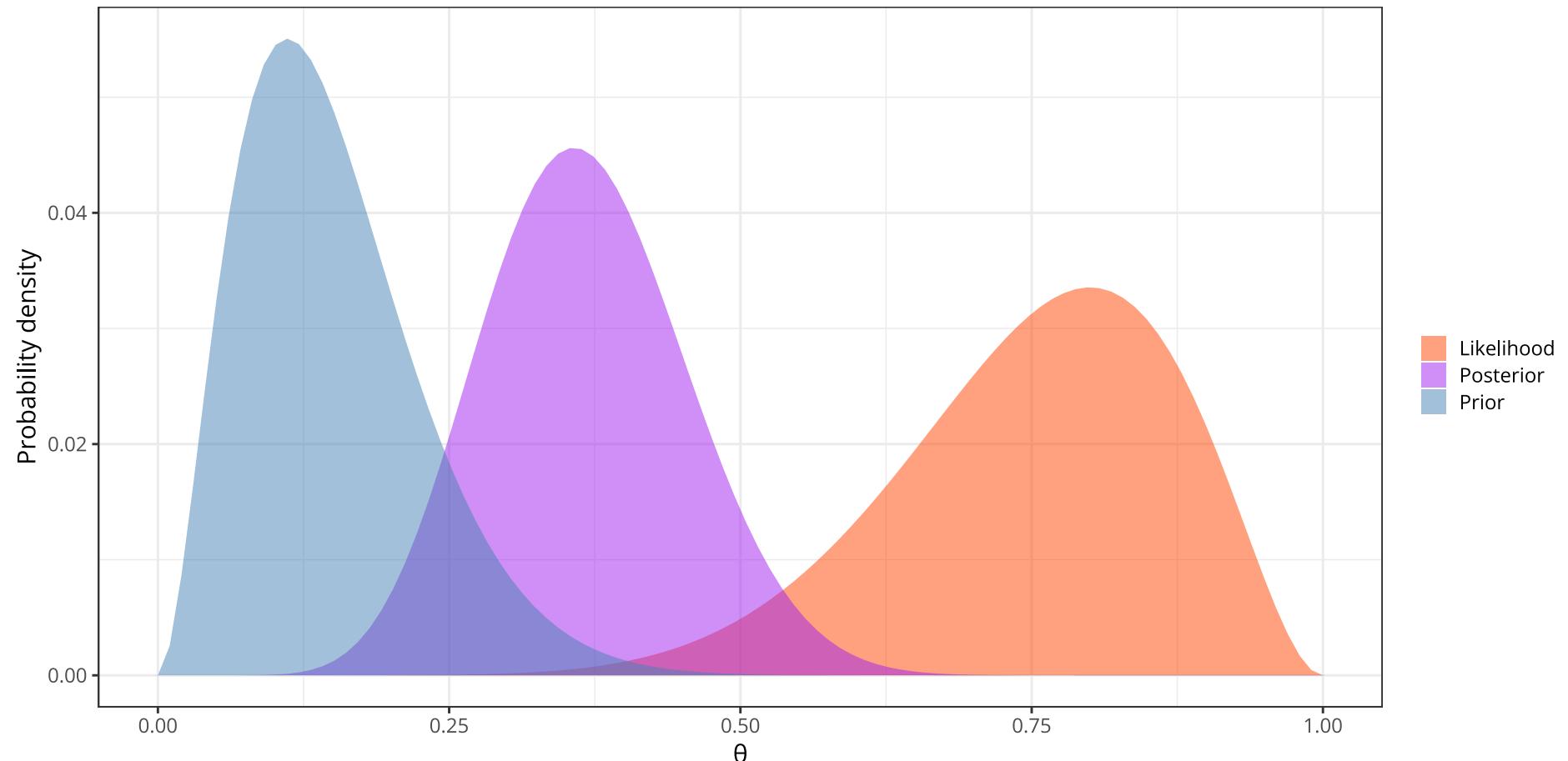
Case where  $n > a + b$ , ( $n = 40$ ,  $a = 4$ ,  $b = 16$ ).



# Key points to remember

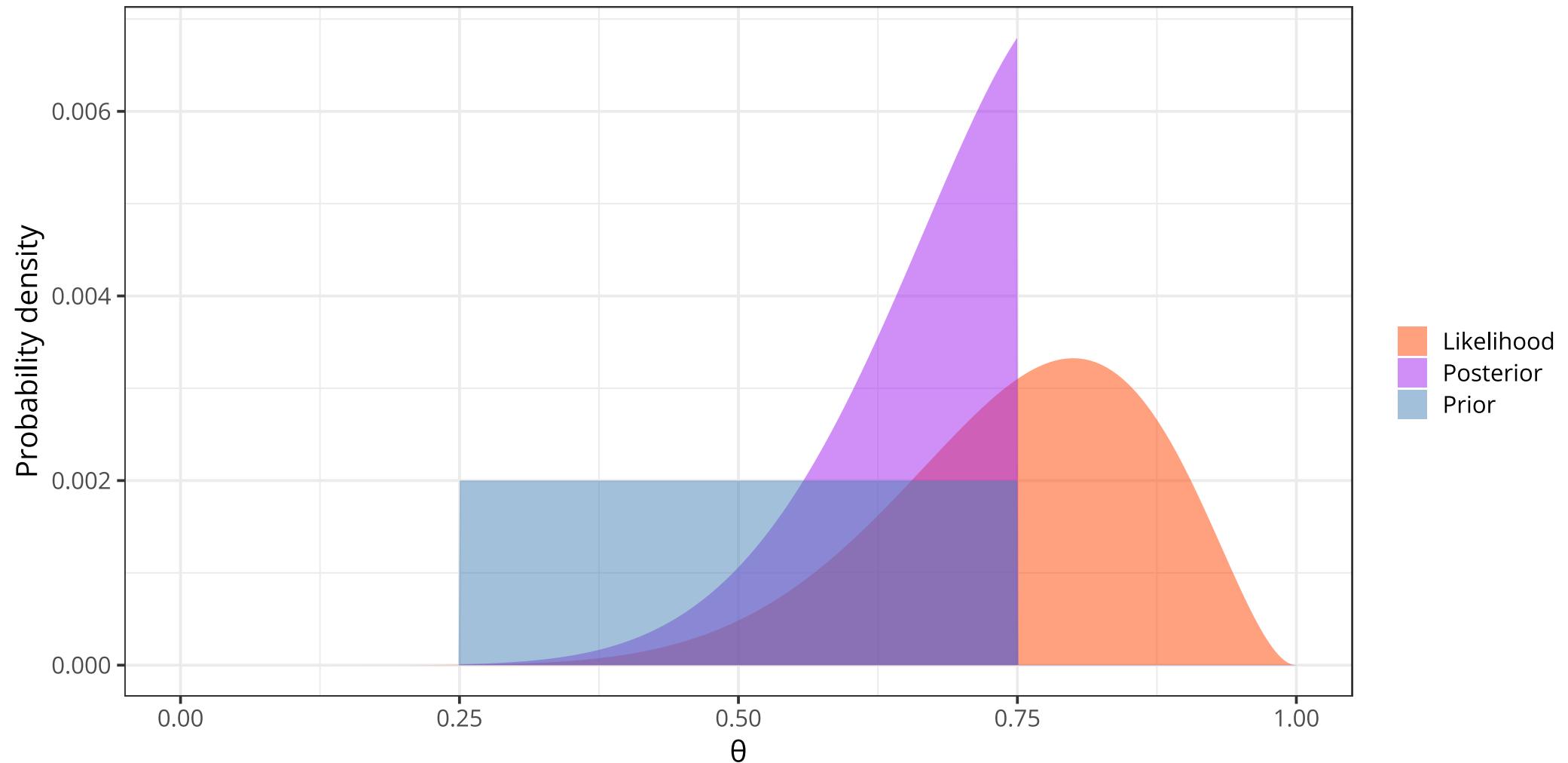
“

The posterior distribution is always a compromise between the prior distribution and the likelihood function ([Kruschke, 2015](#)).



# Key points to remember

The more data we have, the less influence the prior has in estimating the posterior distribution (and vice versa). **Warning:** When the prior assigns a probability of 0 to certain values of  $\theta$ , the model is unable to learn (these values are then considered “impossible”).



# Marginal likelihood

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}} \propto \text{Likelihood} \times \text{Prior}$$

$$p(\theta | \text{data}) = \frac{p(\text{data} | \theta) \times p(\theta)}{p(\text{data})} \propto p(\text{data} | \theta) \times p(\theta)$$

If we zoom in on the marginal likelihood (also known as **evidence**)...

$$p(\text{data}) = \int p(\text{data}, \theta) d\theta \quad \text{Marginalising over } \theta$$

$$p(\text{data}) = \int p(\text{data} | \theta) \times p(\theta) d\theta \quad \text{Applying the product rule}$$



# Marginal likelihood

New issue:  $p(\text{data})$  is obtained by calculating the sum (for discrete variables) or the integral (for continuous variables) of the joint density  $p(\text{data}, \theta)$  over all possible values of  $\theta$ . This can become tricky when the model includes many continuous parameters...

Let's consider a model with two discrete parameters. The marginal likelihood is obtained as:

$$p(\text{data}) = \sum_{\theta_1} \sum_{\theta_2} p(\text{data}, \theta_1, \theta_2)$$

Let's now consider a model with two continuous parameters. The marginal likelihood is obtained as:

$$p(\text{data}) = \int_{\theta_1} \int_{\theta_2} p(\text{data}, \theta_1, \theta_2) d\theta_1 d\theta_2$$



# Marginal likelihood

There are three ways to get around this problem:

- Analytical solution —→ Using a conjugate prior (e.g., the Beta-Binomial model).
- Discretised solution —→ Computing the posterior on a finite set of points (grid method).
- Approximate solution —→ The parameter space is “cleverly” sampled (e.g., MCMC methods, cf. Course n°03).



# Discrete distributions

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters <sup>[note 1]</sup>	Posterior predictive <sup>[note 2]</sup>
Bernoulli	$p$ (probability)	Beta	$\alpha, \beta$	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures <sup>[note 1]</sup>	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	$p$ (probability)	Beta	$\alpha, \beta$	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures <sup>[note 1]</sup>	BetaBin $(\tilde{x} \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, $r$	$p$ (probability)	Beta	$\alpha, \beta$	$\alpha + \sum_{i=1}^n x_i, \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures <sup>[note 1]</sup> (i.e., $\frac{\beta - 1}{r}$ experiments, assuming $r$ stays fixed)	
Poisson	$\lambda$ (rate)	Gamma	$k, \theta$	$k + \sum_{i=1}^n x_i, \frac{\theta}{n\theta + 1}$	$k$ total occurrences in $\frac{1}{\theta}$ intervals	NB $(\tilde{x} k', \theta')$ (negative binomial)
			$\alpha, \beta^{[note 3]}$	$\alpha + \sum_{i=1}^n x_i, \beta + n$	$\alpha$ total occurrences in $\beta$ intervals	NB $(\tilde{x} \alpha', \frac{1}{1 + \beta'})$ (negative binomial)
Categorical	$\mathbf{p}$ (probability vector), $k$ (number of categories; i.e., size of $\mathbf{p}$ )	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \dots, c_k)$ , where $c_i$ is the number of observations in category $i$	$\alpha_i - 1$ occurrences of category $i^{[note 1]}$	$p(\tilde{x} = i) = \frac{\alpha_i'}{\sum_i \alpha_i'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\mathbf{p}$ (probability vector), $k$ (number of categories; i.e., size of $\mathbf{p}$ )	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category $i^{[note 1]}$	DirMult $(\tilde{\mathbf{x}} \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, $N$	$M$ (number of target members)	Beta-binomial <sup>[4]</sup>	$n = N, \alpha, \beta$	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures <sup>[note 1]</sup>	
Geometric	$p_0$ (probability)	Beta	$\alpha, \beta$	$\alpha + n, \beta + \sum_{i=1}^n x_i$	$\alpha - 1$ experiments, $\beta - 1$ total failures <sup>[note 1]</sup>	



# Continuous distributions

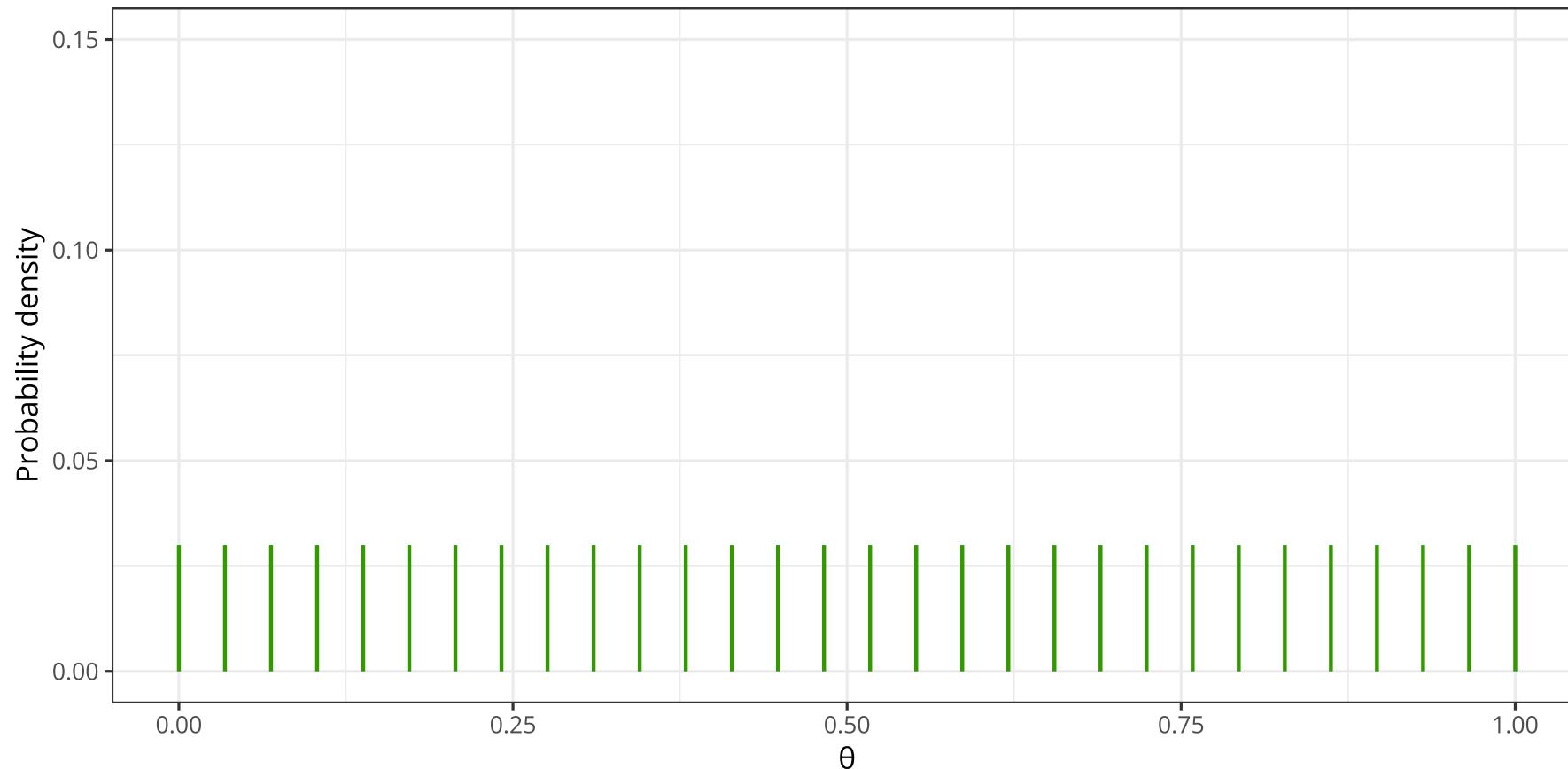
Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive <sup>[note 4]</sup>
Normal with known variance $\sigma^2$	$\mu$ (mean)	Normal	$\mu_0, \sigma_0^2$	$\left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) / \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right),$ $\left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean $\mu_0$	$\mathcal{N}(\tilde{x} \mu'_0, \sigma'^2_0 + \sigma^2)$ <sup>[5]</sup>
Normal with known precision $\tau$	$\mu$ (mean)	Normal	$\mu_0, \tau_0$	$\left( \tau_0 \mu_0 + \tau \sum_{i=1}^n x_i \right) / (\tau_0 + n\tau), \tau_0 + n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) $\tau_0$ and with sample mean $\mu_0$	$\mathcal{N}(\tilde{x} \mu'_0, \frac{1}{\tau'_0} + \frac{1}{\tau})$ <sup>[5]</sup>
Normal with known mean $\mu$	$\sigma^2$ (variance)	Inverse gamma	$\alpha, \beta$ <sup>[note 5]</sup>	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$	variance was estimated from $2\alpha$ observations with sample variance $\beta/\alpha$ (i.e. with sum of squared deviations $2\beta$ , where deviations are from known mean $\mu$ )	$t_{2\alpha'}(\tilde{x} \mu, \sigma^2 = \beta'/\alpha')$ <sup>[5]</sup>
Normal with known mean $\mu$	$\sigma^2$ (variance)	Scaled inverse chi-squared	$\nu, \sigma_0^2$	$\nu + n, \frac{\nu\sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{\nu + n}$	variance was estimated from $\nu$ observations with sample variance $\sigma_0^2$	$t_\nu(\tilde{x} \mu, \sigma'^2_0)$ <sup>[5]</sup>
Normal with known mean $\mu$	$\tau$ (precision)	Gamma	$\alpha, \beta$ <sup>[note 3]</sup>	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$	precision was estimated from $2\alpha$ observations with sample variance $\beta/\alpha$ (i.e. with sum of squared deviations $2\beta$ , where deviations are from known mean $\mu$ )	$t_{2\alpha'}(\tilde{x} \mu, \sigma^2 = \beta'/\alpha')$ <sup>[5]</sup>
Normal <sup>[note 6]</sup>	$\mu$ and $\sigma^2$ Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from $\nu$ observations with sample mean $\mu_0$ ; variance was estimated from $2\alpha$ observations with sample mean $\mu_0$ and sum of squared deviations $2\beta$	$t_{2\alpha'}(\tilde{x} \mu', \frac{\beta'(\nu' + 1)}{\nu'\alpha'})$ <sup>[5]</sup>

Problem: This solution is very restrictive. Ideally, the model (likelihood + prior) model should be defined on the basis of the interpretation of the parameters of these distributions, and not to facilitate the calculations...



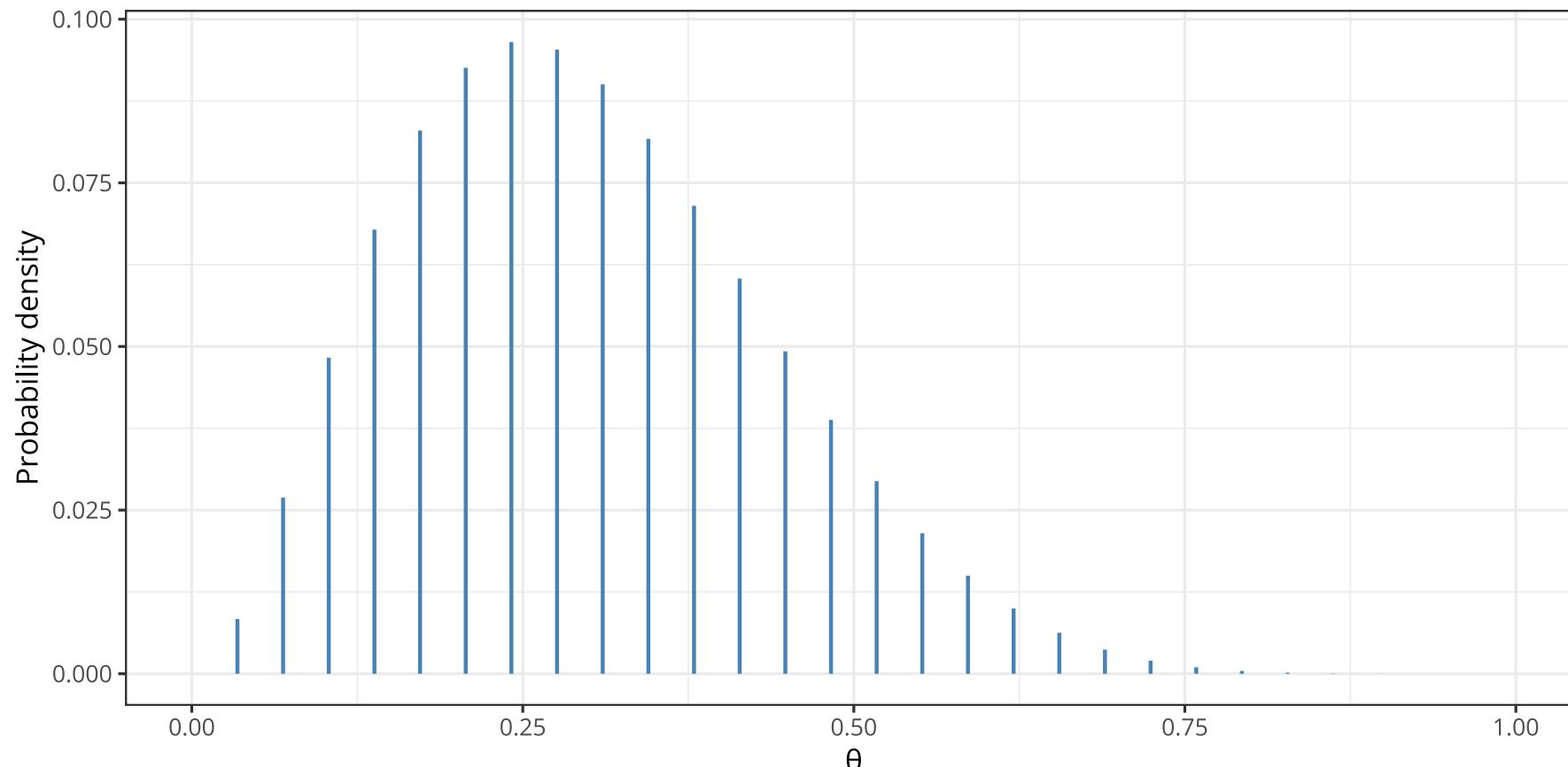
# Posterior distribution, grid method

- **Define the grid**
- Calculate the prior probability for each grid value
- Calculate the likelihood for each grid value
- Calculate the product of prior x likelihood for each grid value, then normalise



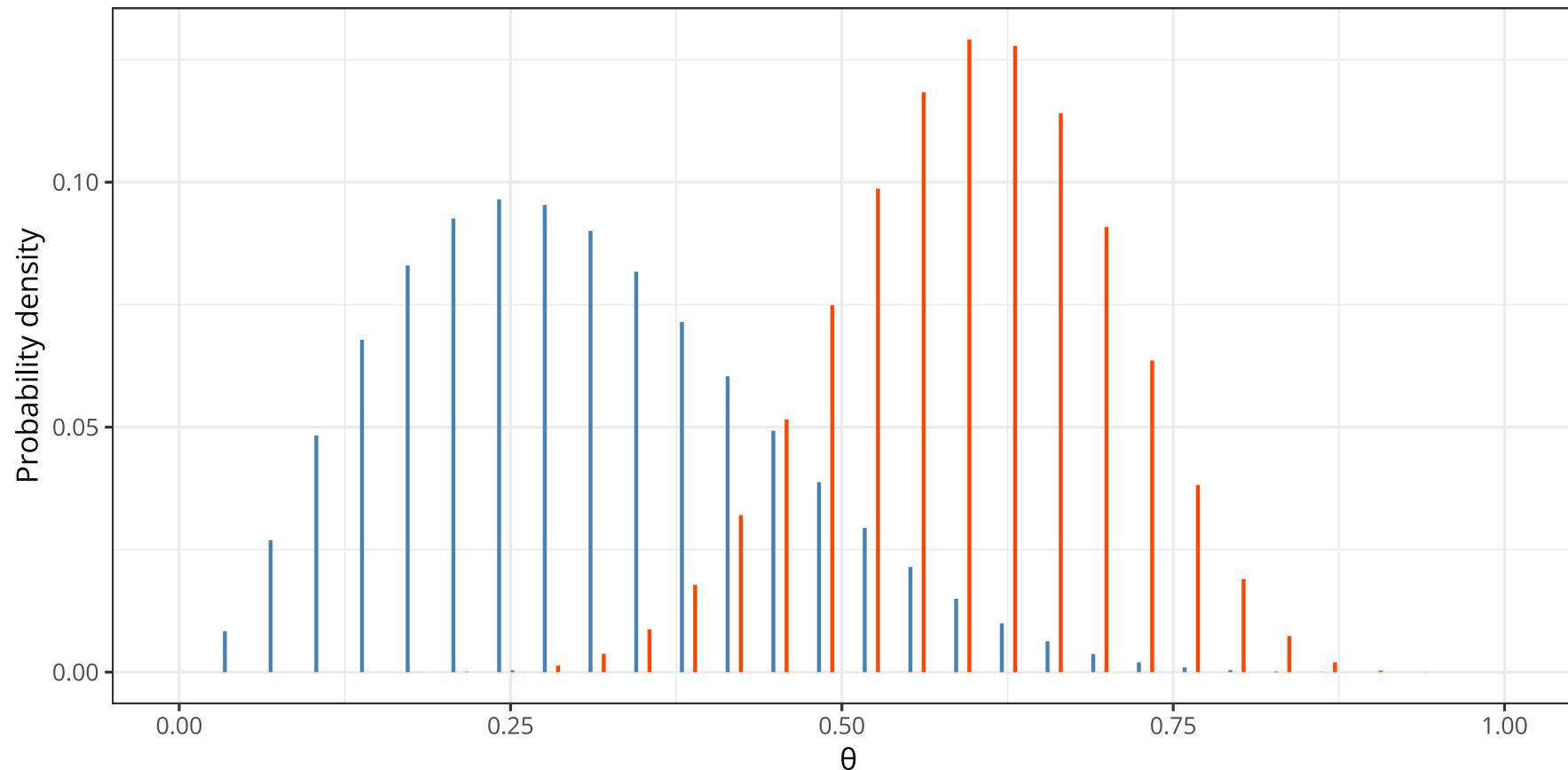
# Posterior distribution, grid method

- Define the grid
- **Calculate the prior probability for each grid value**
- Calculate the likelihood for each grid value
- Calculate the product of prior x likelihood for each grid value, then normalise



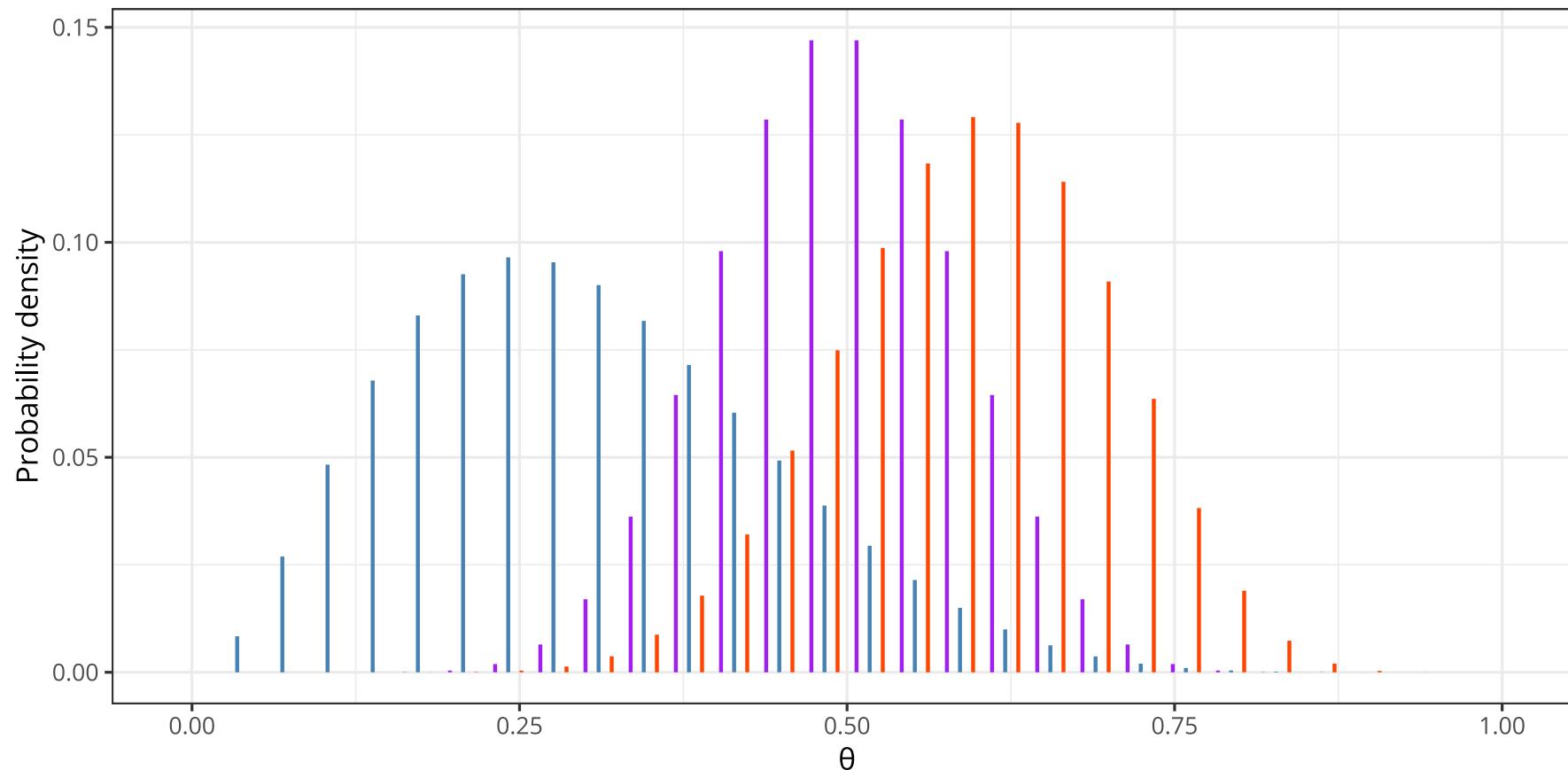
# Posterior distribution, grid method

- Define the grid
- Calculate the prior probability for each grid value
- **Calculate the likelihood for each grid value**
- Calculate the product of prior  $\times$  likelihood for each grid value, then normalise



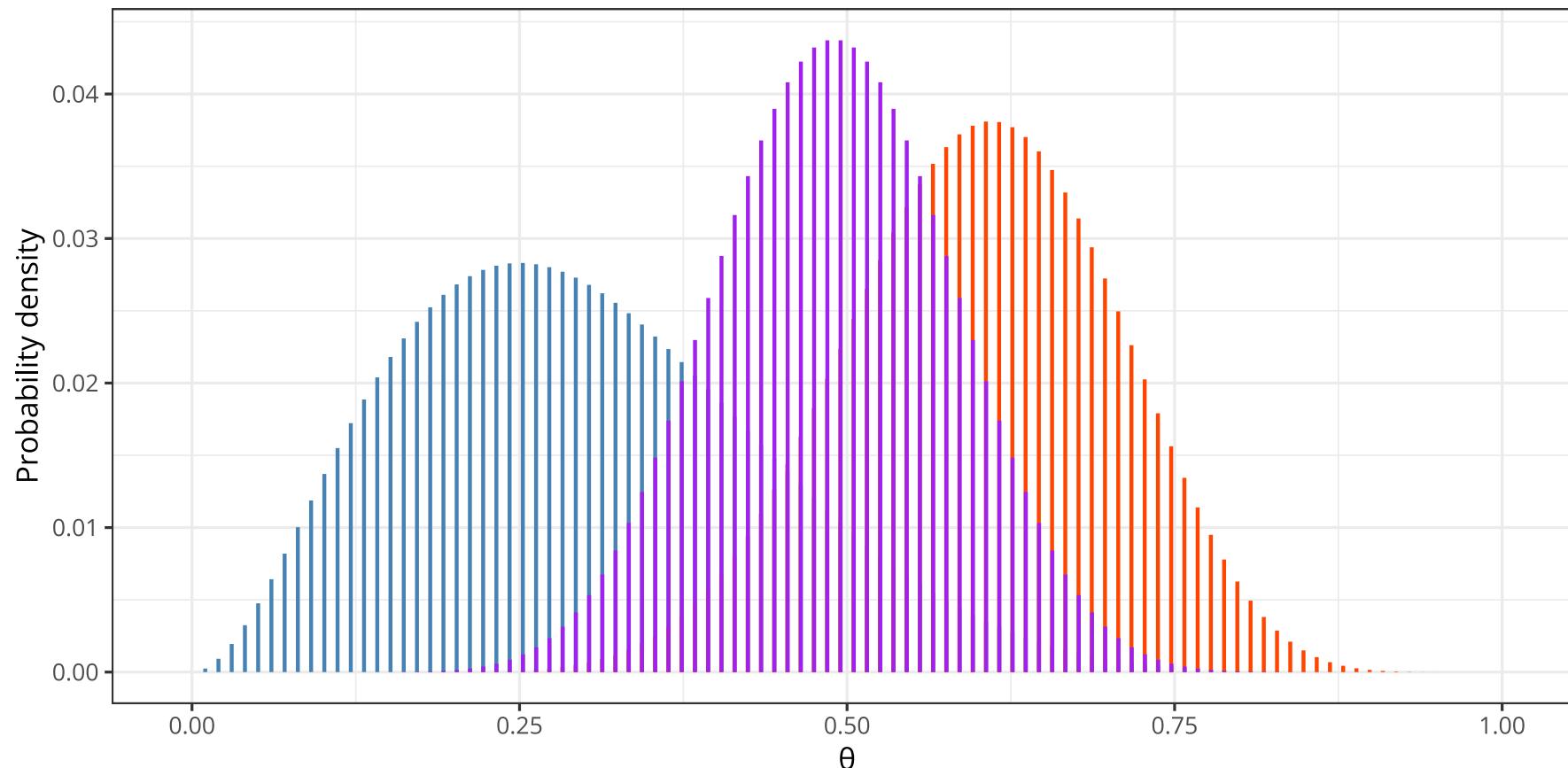
# Posterior distribution, grid method

- Define the grid
- Calculate the prior probability for each grid value
- Calculate the likelihood for each grid value
- **Calculate the product of prior x likelihood for each grid value, then normalise**



# Posterior distribution, grid method

- Define the grid
- Calculate the prior probability for each grid value
- Calculate the likelihood for each grid value
- **Calculate the product of prior x likelihood for each grid value, then normalise**



# Posterior distribution, grid method

Problem with the number of parameters... Refining the grid increases the calculation time:

- 3 parameters with a  $10^3$  grid =  $10^9$  calculation points
- 10 parameters with a grid of  $10^3$  nodes =  $10^{30}$  calculation points

The best supercomputer (Frontier) performs around  $1.194 \times 10^{18}$  operations per second. If we consider that it would have to perform 4 operations per node of the grid, it would take more time to go through the grid than the estimated age of the universe (approximately  $(4.36 \pm 0.012) \times 10^{17}$  seconds)...



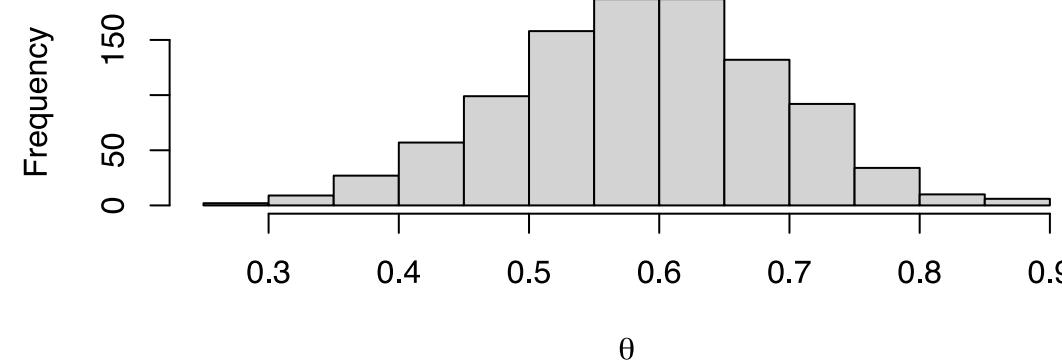
# Another solution: Sampling the posterior distribution

To sample (cleverly) a posterior distribution, we can use different implementations of MCMC methods (e.g., Metropolis-Hastings, Hamiltonian Monte Carlo) which we will discuss in Course n°03. In the meantime, we will work with samples from the posterior distribution i) to get used to results from MCMC methods and ii) because it is simpler to compute summary statistics (e.g., mean or credible intervals) on samples rather than by computing integrals.

```

1 p_grid <- seq(from = 0, to = 1, length.out = 1000) # creates a grid
2 prior <- rep(1, 1000) # uniform prior
3 likelihood <- dbinom(x = 12, size = 20, prob = p_grid) # computes likelihood
4 posterior <- (likelihood * prior) / sum(likelihood * prior) # computes posterior
5 samples <- sample(x = p_grid, size = 1e3, prob = posterior, replace = TRUE) # sampling
6 hist(samples, main = "", xlab = expression(theta)) # histogram

```



# Posterior distribution, summary

Analytical solution for the Beta-Binomial model:

```
1 a <- b <- 1 # parameters of the Beta prior
2 n <- 9 # number of observations
3 y <- 6 # number of successes
4 p_grid <- seq(from = 0, to = 1, length.out = 1000)
5 posterior <- dbeta(p_grid, y + a, n - y + b) # plot(posterior)
```

Grid method:

```
1 p_grid <- seq(from = 0, to = 1, length.out = 1000)
2 prior <- rep(1, 1000) # uniform prior
3 likelihood <- dbinom(x = y, size = n, prob = p_grid)
4 posterior <- (likelihood * prior) / sum(likelihood * prior) # plot(posterior)
```

Sampling the posterior distribution to describe it:

```
1 samples <- sample(x = p_grid, size = 1e4, prob = posterior, replace = TRUE) # hist(samples)
```



# Posterior distribution, summary

## Analytical solution

- The posterior distribution is described explicitly
- The model is strongly constrained (i.e., we have to pick conjugate priors)

## Grid method

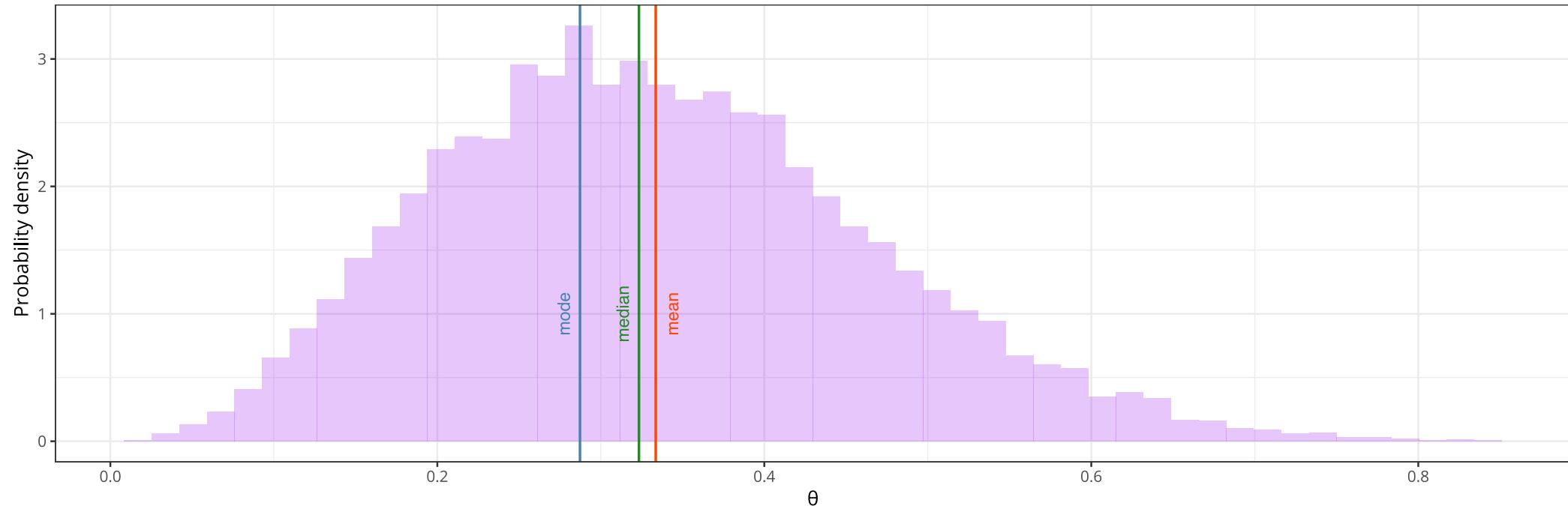
- The posterior distribution is only computed for a finite set of values
- The finer the grid, the better the estimate of the posterior distribution
- There is an “accuracy - calculation time” trade-off



# Using samples to summarise the posterior distribution

Estimation of the central tendency: From a set of samples of samples from a posterior distribution, we can calculate the mean mean, mode, and median. For example, for a uniform prior, 10 coin tosses and 3 heads.

```
1 mode_posterior <- find_mode(samples) # in blue
2 mean_posterior <- mean(samples) # in orange
3 median_posterior <- median(samples) # in green
```



# Using samples to summarise the posterior distribution

What is the probability that the bias of the coin  $\theta$  is greater than 0.5?

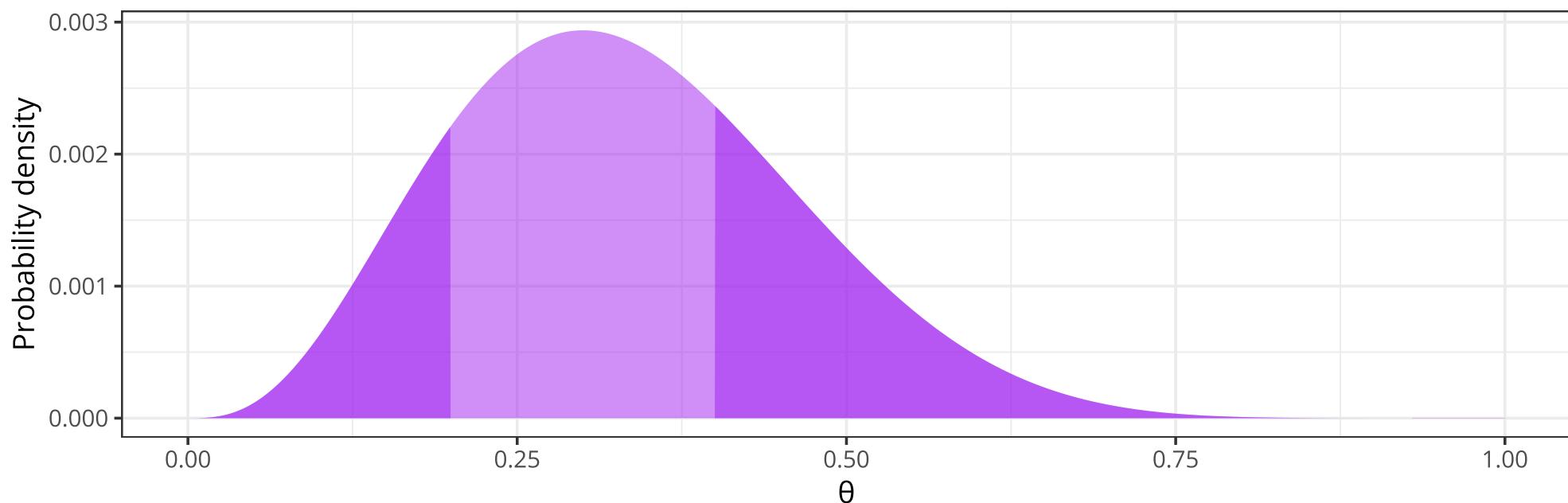
```
1 sum(samples > 0.5) / length(samples) # equivalent to mean(samples > 0.5)
```

```
[1] 0.112
```

What is the probability that the bias of the coin  $\theta$  is between 0.2 and 0.4?

```
1 sum(samples > 0.2 & samples < 0.4) / length(samples)
```

```
[1] 0.5482
```



# Highest density interval (HDI)

Properties of the highest density interval:

- The HDI indicates the most likely values for the parameter (given the data and the priors)
- The narrower the HDI, the greater the degree of certainty
- The width of the HDI decreases as the number of measurements increases

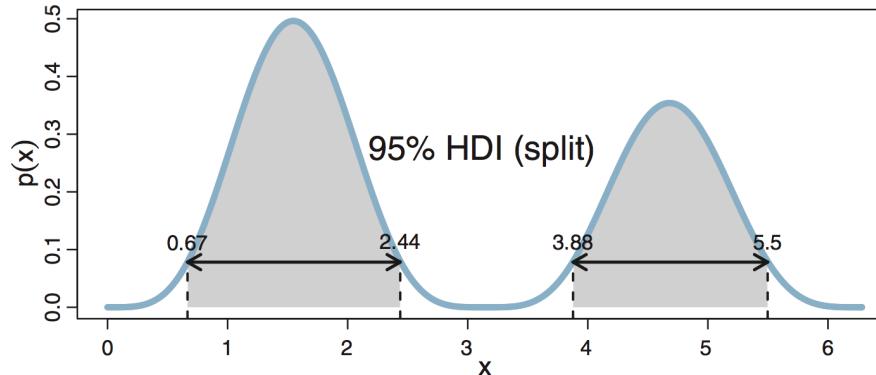
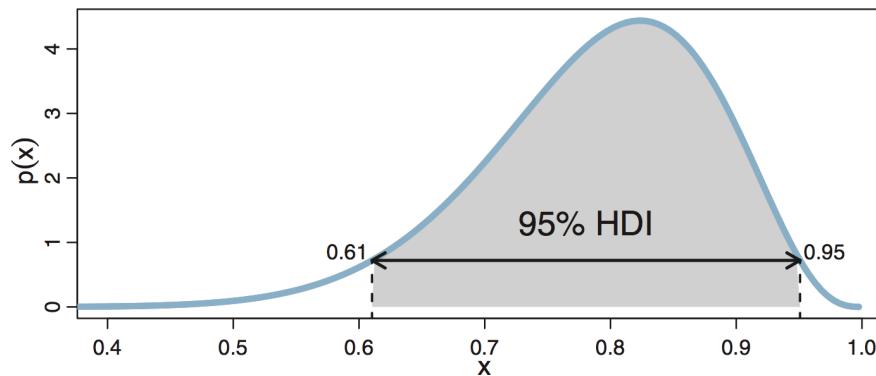
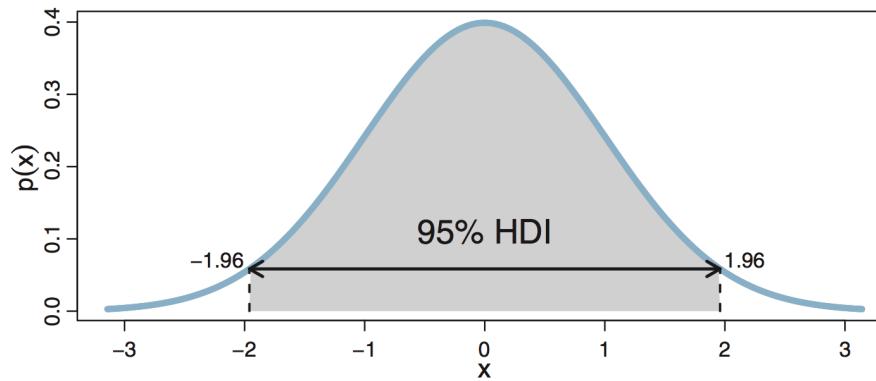
“

Definition: the values of the parameter  $\theta$  contained in an HDI at 89% are such that  $p(\theta) > W$  where  $W$  satisfies the following condition:

$$\int_{\theta : p(\theta) > W} p(\theta) d\theta = 0.89.$$

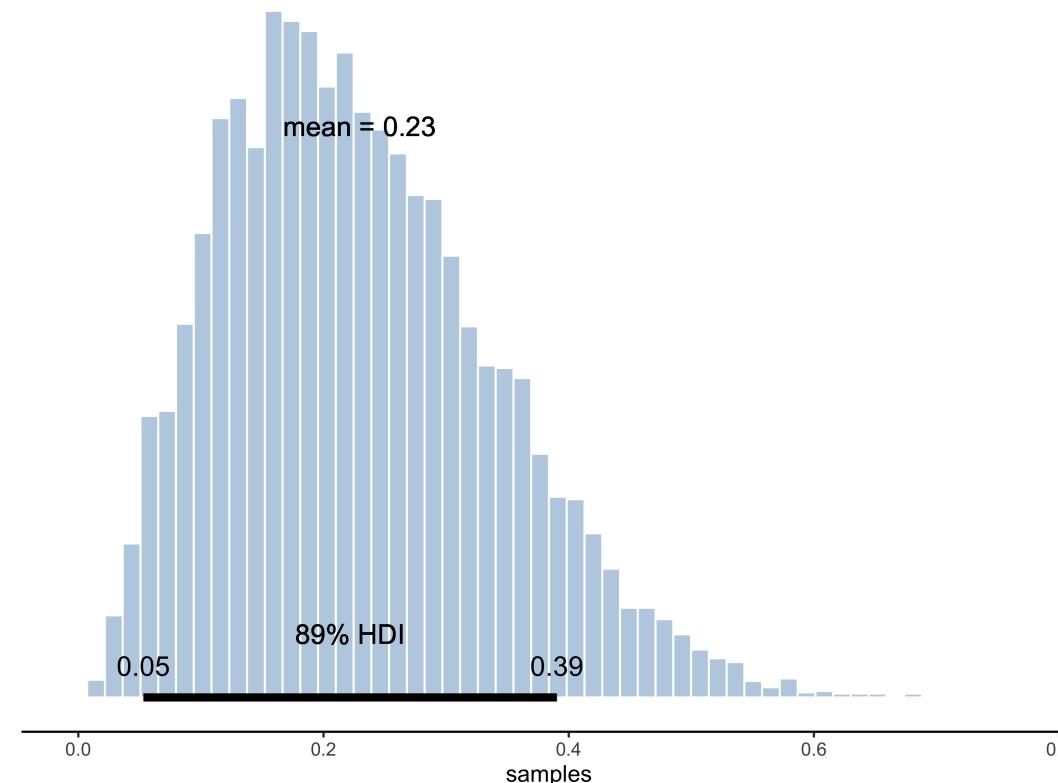


# Highest density interval (HDI)



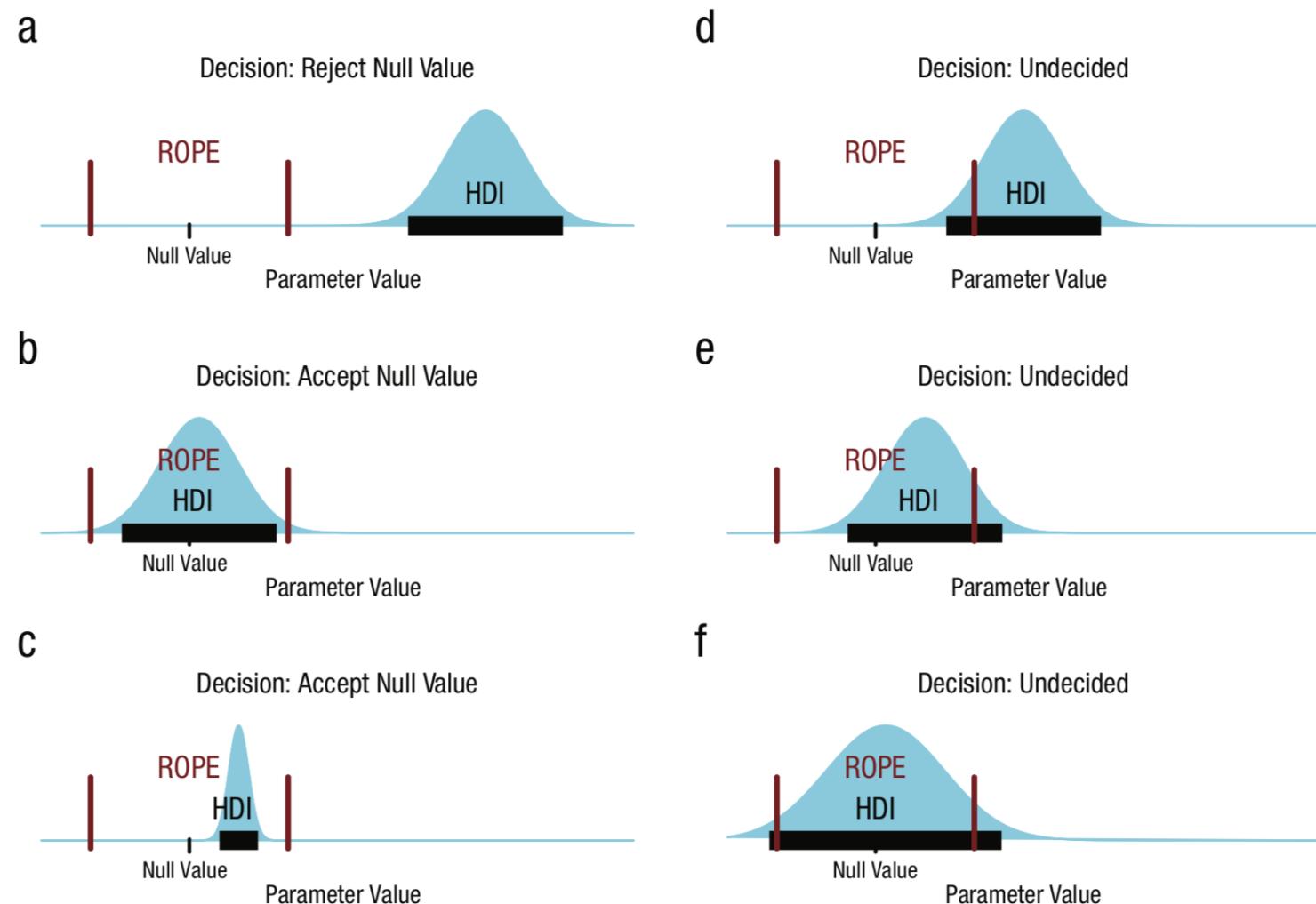
# Highest density interval (HDI)

```
1 library(imbsb)
2
3 set.seed(666)
4 p_grid <- seq(from = 0, to = 1, length.out = 1e3)
5 pTheta <- dbeta(p_grid, 3, 10)
6 massVec <- pTheta / sum(pTheta)
7 samples <- sample(x = p_grid, size = 1e4, replace = TRUE, prob = pTheta)
8
9 posterior_plot(samples = samples, credmass = 0.89)
```



# Region of practical equivalence (ROPE)

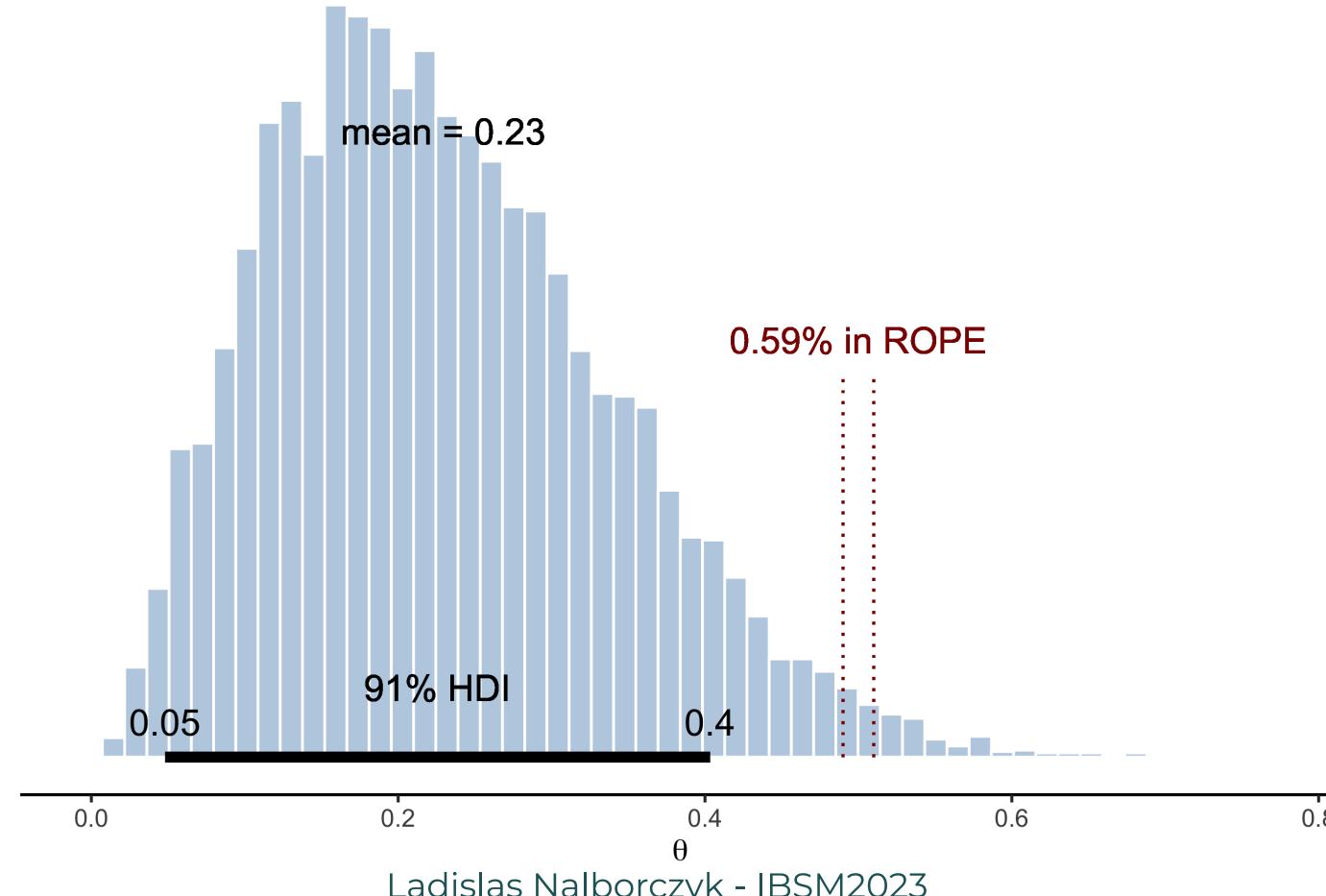
This procedure can be used to accept or reject a null value. The region of practical equivalence or **region of practical equivalence** (ROPE) defines an interval of values that are considered to be “equivalent” to the null value. The figure below summarises the possible decisions resulting from this procedure ([Kruschke, 2018](#)).



# Region of practical equivalence (ROPE)

The value of the parameter (e.g.,  $\theta = 0.5$ ) is rejected if the HDI is entirely outside the ROPE. The parameter value (e.g.,  $\theta = 0.5$ ) is accepted if the HDI is entirely within the ROPE. If the HDI and the ROPE overlap, we cannot conclude...

```
1 posterior_plot(samples = samples, rope = c(0.49, 0.51) ) +  
2     labs(x = expression(theta))
```



# Model comparison

We tossed coin 200 times and got 115 “Faces”. Is the coin coin biased? We build two models and try to find out which one which one best accounts for the data.

$$\begin{cases} \mathcal{M}_0 : Y \sim \text{Binomial}(n, \theta = 0.5) & \text{The coin is not biased} \\ \mathcal{M}_1 : Y \sim \text{Binomial}(n, \theta \neq 0.5) & \text{The coin is biased} \end{cases}$$

The Bayes factor is the ratio of the marginal likelihoods (of the two models).

$$\frac{p(\mathcal{M}_0 \mid \text{data})}{p(\mathcal{M}_1 \mid \text{data})} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} \times \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}$$



# Model comparison

The Bayes factor is the ratio of the marginal likelihoods (of the two models).

$$\frac{p(\mathcal{M}_0 \mid \text{data})}{p(\mathcal{M}_1 \mid \text{data})} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} \times \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}$$

In our example:

$$\text{BF}_{01} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} = \frac{0.005955}{0.004975} \approx 1.1971.$$

This BF indicates that the (prior) odds ratio increased (or should be updated) by 20% in favour of  $\mathcal{M}_0$  after observing the data. The Bayes factor can also be interpreted as follows: The data are approximately 1.2 times more likely under the  $\mathcal{M}_0$  model than under the  $\mathcal{M}_1$  model.



# Model checking

The two roles of the likelihood function:

- It is a function of  $\theta$  for calculating the posterior distribution:  $\mathcal{L}(\theta | y, n)$ .
- When  $\theta$  is known/fixed, it is a probability distribution:  $p(y | \theta, n) \propto \theta^y(1 - \theta)^{(n-y)}$ .

This probability distribution can be used to generate data... !

For example: Generating 10.000 values from a binomial distribution based on 10 coin tosses and a probability of Heads of 0.6:

```
1 samples <- rbinom(n = 1e4, size = 10, prob = 0.6)
```



# Model checking

In a Bayesian models, there are **two sources of uncertainty** when generating predictions:

- Uncertainty related to the sampling process  
-> We draw data from a Binomial pdf
- Uncertainty about the value of  $\theta$   
-> Our knowledge of  $\theta$  is described by a (posterior) pdf

For example: Generating 10.000 values from a binomial distribution based on 10 coin tosses and a probability of Heads described by the posterior distribution of  $\theta$ :

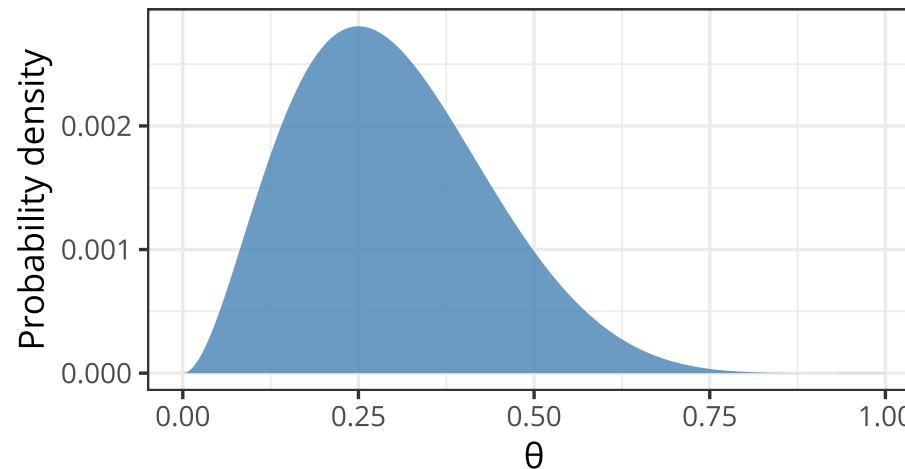
```
1 posterior <- rbeta(n = 1e4, shape1 = 16, shape2 = 10)
2 samples <- rbinom(n = 1e4, size = 10, prob = posterior)
```



# Prior and posterior predictive checking

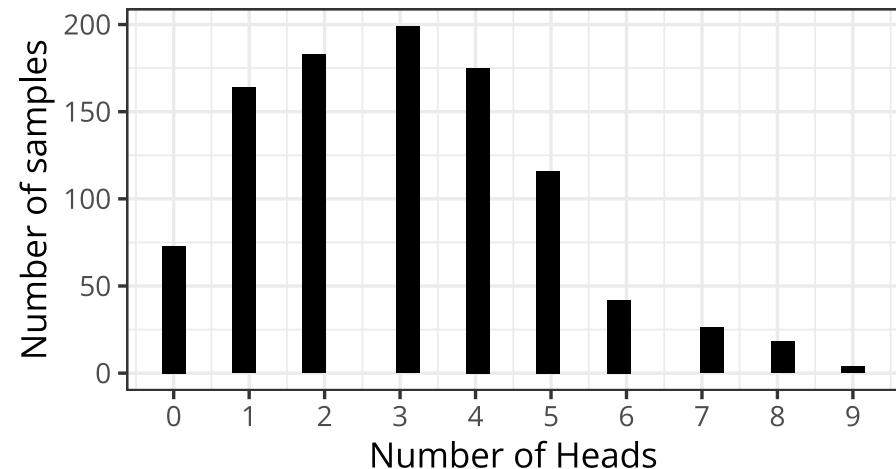
Prior distribution

`rbeta(n = 1e4, shape1 = 3, shape2 = 7)`



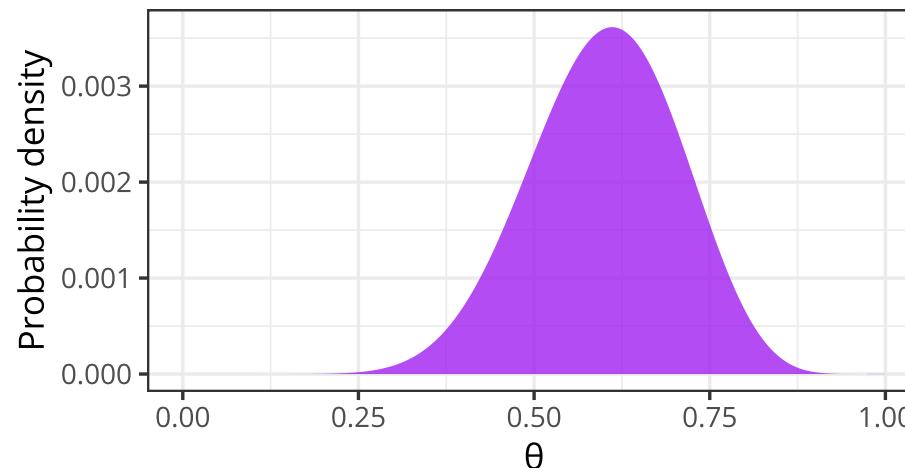
Prior predictive distribution

`rbinom(n = 1e4, size = 10, prob = prior)`



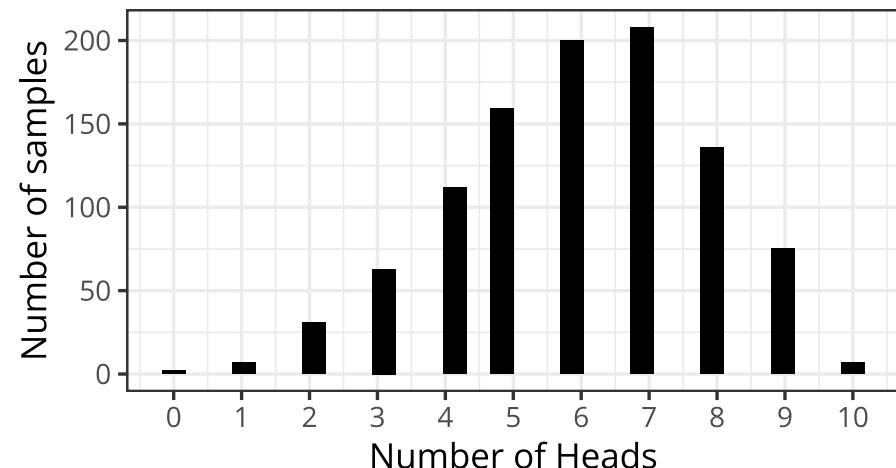
Posterior distribution

`rbeta(n = 1e4, shape1 = 12, shape2 = 8)`

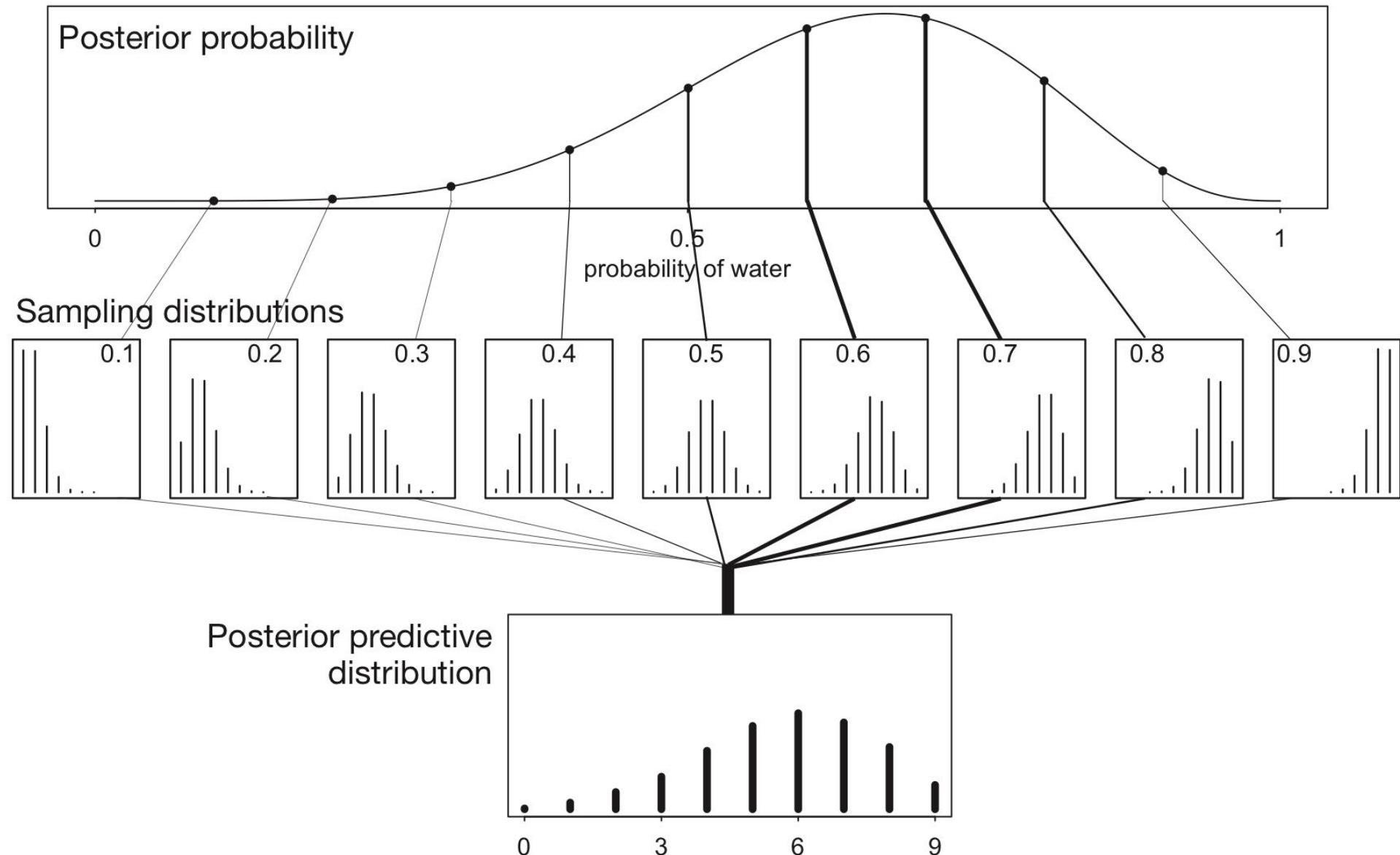


Posterior predictive distribution

`rbinom(n = 1e4, size = 10, prob = posterior)`



# Posterior predictive checking



# Practical work

RICHARD ROBINSON

## WHY THE TOAST ALWAYS LANDS BUTTER SIDE DOWN

THE SCIENCE OF  
MURPHY'S LAW



An analyst who works in a factory that makes famous Swedish bread rolls read a book that raised a thorny question... Why does the toast always land on the butter side? Failing to come up with a plausible answer, she set out to verify this assertion. The first experiment was to drop a slice of buttered bread from the height of a table. The results of this experiment are available in the `tartine1` dataset from the `imsb` package.



# Retrieving the data

First task: Retrieving the data.

```
1 # importing the data
2 data <- open_data(tartine1)
3
4 # summary of the data
5 str(data)
```

'data.frame': 500 obs. of 2 variables:  
\$ trial: int 1 2 3 4 5 6 7 8 9 10 ...  
\$ side : int 1 1 0 1 0 0 1 1 1 0 ...



# Questions

- Since the toast only has two sides, the result is similar to a draw according to a binomial distribution with an unknown parameter  $\theta$ . What is the posterior distribution of the parameter  $\theta$  given these data and given that the analyst had no prior knowledge (you can also use your own prior)?
- Calculate the 95% HDI of the posterior distribution and give a graphical representation of the result (hint: use the function `imsb::posterior_plot()`).
- Can the null hypothesis that  $\theta = 0.5$  be rejected? Answer this question using the HDI+ROPE procedure.
- Import the `tartine2` data from the `imsb` package. Update the model using the mode of the posterior distribution calculated previously.



# Proposed solution - Question 1

Since the toast only has two sides, the result is similar to a draw according to a binomial distribution with an unknown parameter  $\theta$ . What is the posterior distribution of the parameter  $\theta$  given these data?

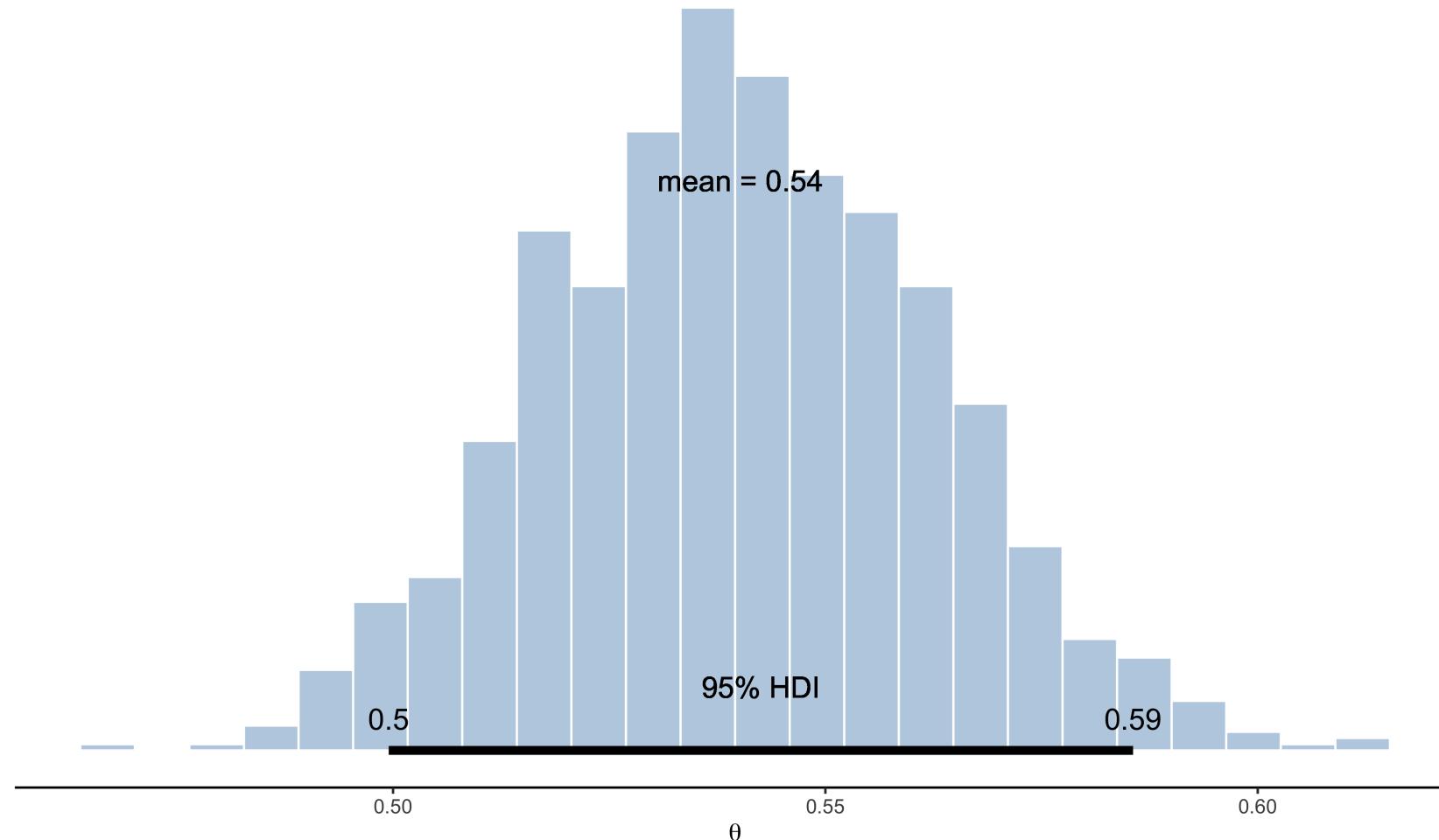
```
1 # number of trials
2 nbTrial <- length(data$trial)
3
4 # number of "successes" (i.e., when the toast lands on the butter side)
5 nbSuccess <- sum(data$side)
6
7 # size of the grid
8 grid_size <- 1e3
9
10 # generating the grid
11 p_grid <- seq(from = 0, to = 1, length.out = grid_size)
12
13 # uniform prior
14 prior <- rep(1, grid_size)
15
16 # computing the likelihood
17 likelihood <- dbinom(x = nbSuccess, size = nbTrial, prob = p_grid)
18
19 # computing the posterior
20 posterior <- likelihood * prior / sum(likelihood * prior)
```



# Proposed solution - Question 2

Calculate the 95% HDI of the posterior distribution and give a graphical representation of the result.

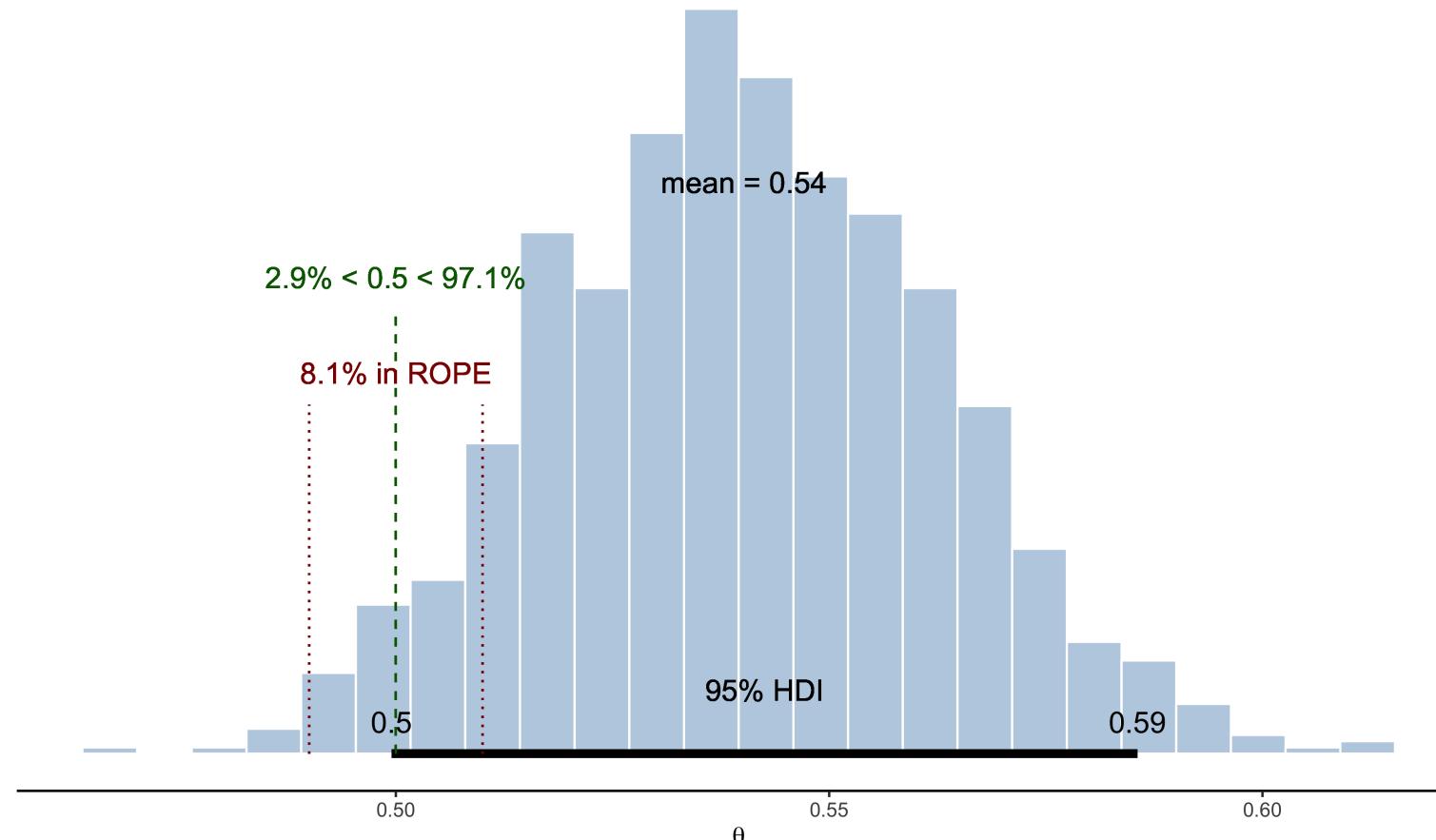
```
1 samples <- sample(x = p_grid, prob = posterior, size = 1e3, replace = TRUE)
2 posterior_plot(samples = samples, credmass = 0.95) + labs(x = expression(theta))
```



# Proposed solution - Question 3

Can the null hypothesis that  $\theta = 0.5$  be rejected? No, because the HDI overlaps with the ROPE...

```
1 posterior_plot(  
2   samples = samples, credmass = 0.95,  
3   compval = 0.5, rope = c(0.49, 0.51)  
4 ) + labs(x = expression(theta))
```



# Proposed solution - Question 4

At this point, no conclusion can be drawn. The analyst decides to repeat a series of observations to refine her results.

```
1 data2 <- open_data(tartine2)
2 str(data2)

'data.frame':   100 obs. of  2 variables:
$ trial: int  1 2 3 4 5 6 7 8 9 10 ...
$ side : int  0 0 1 0 0 1 1 1 0 0 ...
```

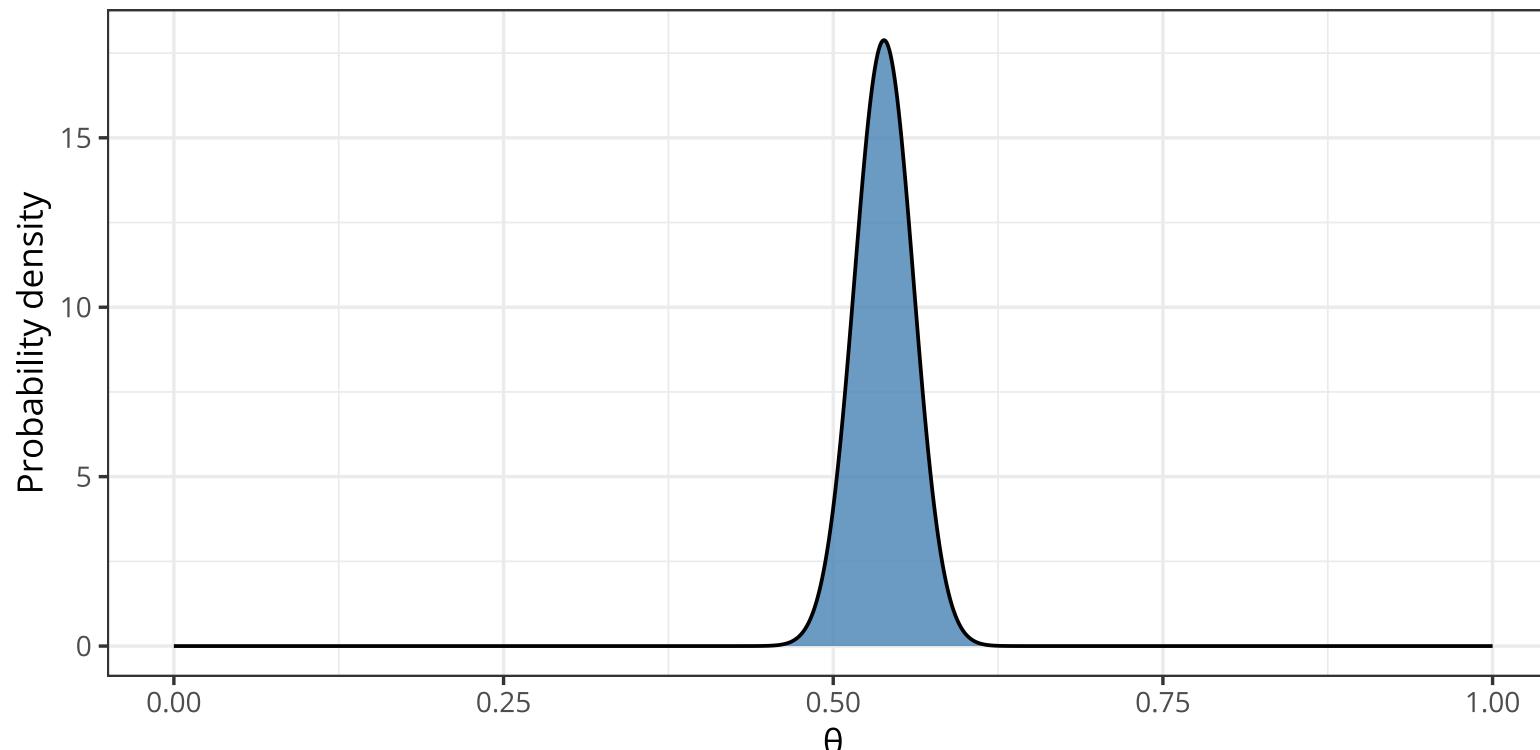
```
1 nbTrial2 <- length(data2$trial) # number of trials
2 nbSuccess2 <- sum(data2$side) # number of "successes"
```



# Proposed solution - Question 4

We use the previous posterior as the prior for this new model.

```
1 model <- find_mode(samples)
2 prior2 <- dbeta(p_grid, model * (nbTrial - 2) + 1, (1 - model) * (nbTrial - 2) + 1)
3
4 data.frame(x = p_grid, y = prior2) %>%
5   ggplot(aes(x = x, y = y)) +
6   geom_area(alpha = 0.8, fill = "steelblue") +
7   geom_line(size = 0.8) +
8   labs(x = expression(theta), y = "Probability density")
```

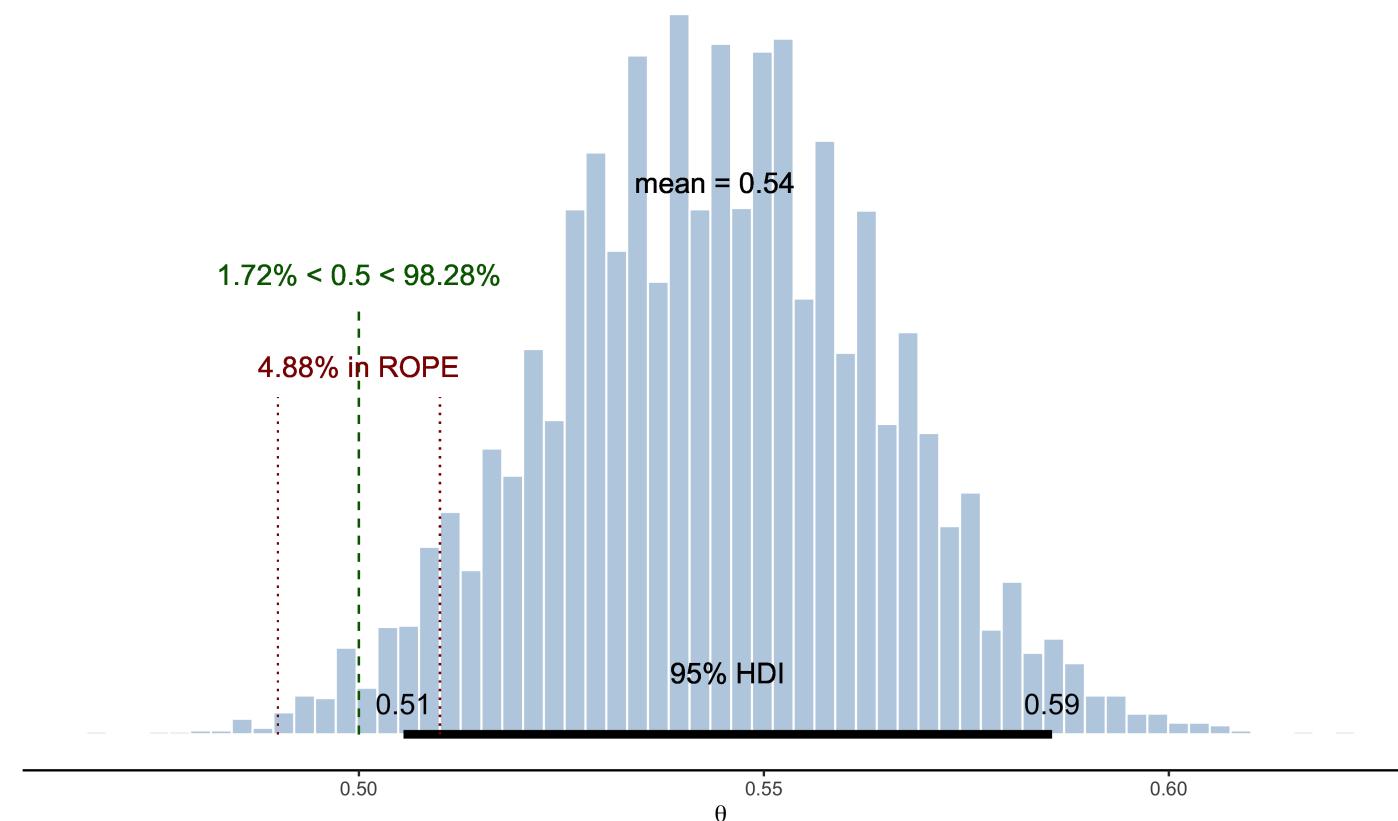


# Proposed solution - Question 4

```

1 likelihood2 <- dbinom(x = nbSuccess2, size = nbTrial2, prob = p_grid)
2 posterior2 <- likelihood2 * prior2 / sum(likelihood2 * prior2)
3 samples2 <- sample(p_grid, prob = posterior2, size = 1e4, replace = TRUE)
4
5 posterior_plot(
6   samples = samples2, credmass = 0.95,
7   compval = 0.5, rope = c(0.49, 0.51)
8 ) + labs(x = expression(theta) )

```



# References

- Carnap, R. (1971). *Logical foundations of probability* (4. impr). Univ. of Chicago Press [u.a.].
- Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition*. CRC Press, Taylor & Francis Group.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8(2), 53–96. <https://doi.org/10.1111/j.1539-6053.2008.00033.x>
- Jaynes, E. T. (1986). *Bayesian methods: General background*.
- Keynes, J. M. (1921). *A Treatise On Probability*. Macmillan And Co.,  
<http://archive.org/details/treatiseonprobab007528mbp>
- Kolmogorov, A. N. (1933). *Foundations of the theory of probability*. New York, USA: Chelsea Publishing Company.
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan* (2nd Edition). Academic Press.
- Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.  
<https://doi.org/10.1177/2515245918771304>
- McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC Press/Taylor & Francis Group.
- McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Taylor; Francis, CRC Press.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4). <https://doi.org/10.1214/aos/1176346785>  
Ladislás Náborczyk - IBSM2023

