Machine Learning Applications: Continuous variables

Alessio Benavoli

CSIS
University of Limerick

Summing up

Given a discrete variable

- a variable, x,
- its domain $\mathrm{dom}(x)$ (only a finite number of possible values),

we determined $p(x = \mathsf{x})$ for all of the possible values of x, and called it the probability mass function (PMF).

Give $p(x = \mathsf{x})$, we can compute everything we are interested in probabilistic analysis.

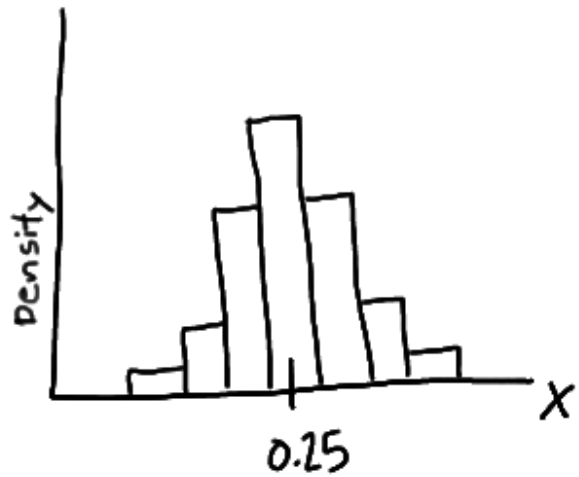
Is it also true for continuous random variables?

Introduction

- A continuous random variable takes on an uncountably infinite number of possible values.
- For continuous random variables, the probability that x takes on any particular value x is 0. That is, finding $p(x = \mathrm{x})$ for a continuous random variable x in general is not going to work.
- Instead, we will need to find the probability that x falls in some interval $[a, b]$, that is, we will need to find $P(a \leq x \leq b)$. We will do that using a probability density function (PDF).

An example

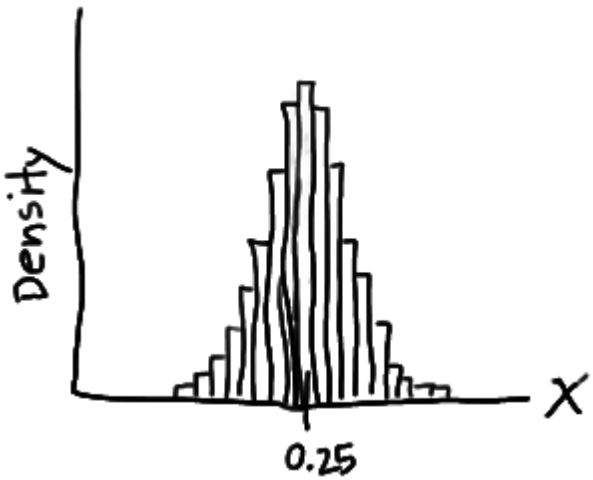
Imagine you could randomly selecting, let's say, 100 pairs of fish cakes advertised to weigh a 0.25 kg. If you weighed them, and created a histogram of the resulting weights it might look something like this:



In this case, the histogram illustrates that most of the sampled fish cakes do indeed weigh close to 0.25 Kg, but some are a bit more and some a bit less.

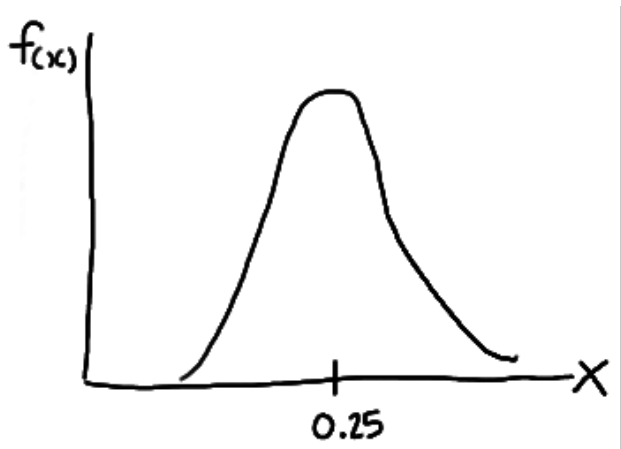
An example

We can decrease the length of the class interval on that density histogram:



An example

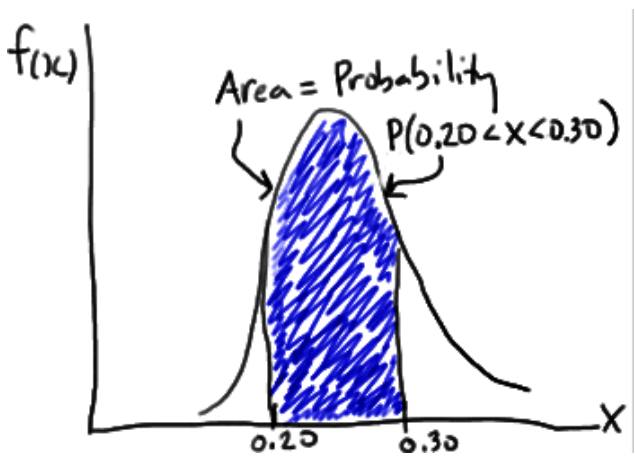
Now, what if we pushed this further and decreased the intervals even more? The intervals get so small that we can represent the probability distribution of x, not as a histogram, but rather as a curve (by connecting the "dots" at the tops of the tiny tiny rectangles) that, in this case, might look like a curve.



Such a curve is denoted $f(x)$ and is called a (continuous) probability density function.

An example

Note that, a histogram is defined so that the area of each rectangle equals the relative frequency of the corresponding class, and the area of the entire histogram equals 1.
That suggests then that finding the probability that a continuous random variable x falls in some interval of values involves finding the area under the curve $f(x)$ sandwiched by the endpoints of the interval. In the case of this example, the probability that a randomly selected fish cakes weighs between 0.20 and 0.30 Kg is then this area:



Definition

The probability density function (PDF) of a continuous random variable x with domain $\mathrm{dom}(x) = \Omega$ is an integrable function $f(x)$ satisfying the following:

1. $f(x)$ is nonnegative everywhere in the domain Ω, that is, $f(x) \geq 0$, for all x in Ω
2. The area under the curve $f(x)$ in the support Ω is 1, that is:

$$\int_{\Omega} f(x)dx = 1$$

3. If $f(x)$ is the PDF of x, then the probability that x belongs to some subset A of Ω is given by the integral of $f(x)$ over that subset, that is:

$$P(x \in A) = \int_{A} f(x)dx$$

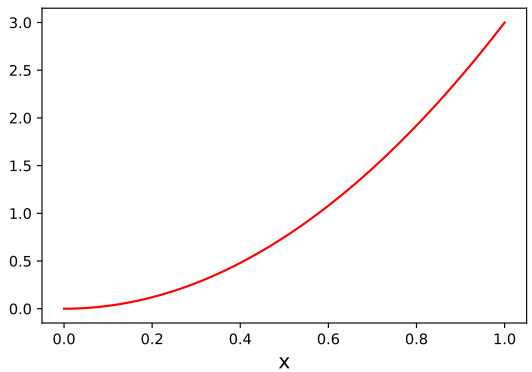
As you can see, the definition for the PDF of a continuous random variable differs from the definition for the PMF of a discrete random variable by simply changing the summations that appeared in the discrete case to integrals in the continuous case.

An example

Let x be a continuous random variable whose probability density function is:

$$f(x) = 3x^2$$

for $0 \leq x \leq 1$ (that means that $\Omega = [0, 1]$).



First, note again that $f(x) \neq p(x = \mathsf{x})$.

For example, $f(0.9) = 3(0.9)^2 = 2.43$, which is clearly not a probability!

In the continuous case, $f(x)$ is instead the height of the curve at $x = \mathsf{x}$, so that the total area under the curve is 1.

In the continuous case, it is areas under the curve that define the probabilities.

An example

Now, let's first start by verifying that $f(x)$ is a valid probability density function.

$$f(x) = 3x^2$$

Is it nonnegative in $\Omega = [0, 1]$?

- Yes

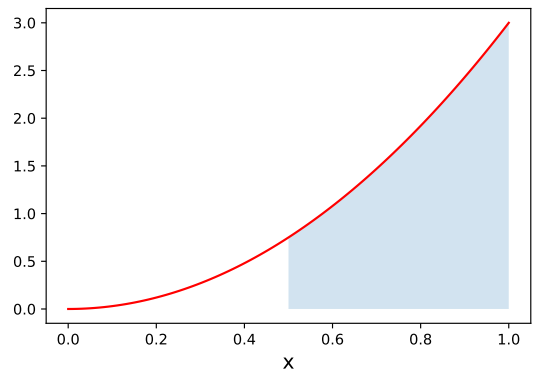
Does it integrate to one in $\Omega = [0, 1]$?

- YES! $\int_0^1 3x^2 = x^3\Big|_0^1 = 1$

An example

What is the probability that x falls between 0.5 and 1? That is, what is $P(0.5 \leq x \leq 1)$?

$$P(0.5 \leq x \leq 1) = \int_{0.5}^{1} 3x^2 = x^3 \Big|_{0.5}^{1} = 1 - 0.5^3 = \frac{7}{8}$$



What is $P(x = 0.5)$?

$$P(0.5 \leq x \leq 0.5) = \int_{0.5}^{0.5} 3x^2 = x^3 \Big|_{0.5}^{0.5} = 0.5^3 - 0.5^3 = 0$$

Note that I am using the notation capital P, because often the probability density function is denoted as $p(x)$ (instead of $f(x)$).

Cumulative Distribution Function

You might recall that the CDF is defined for discrete variables as:

$$F(\mathrm{x}) = P(x \leq \mathrm{x}) = \sum_{t \leq \mathrm{x}} p(x)$$

Again, F(x) accumulates all of the probability less than or equal to x.

The cumulative distribution function for continuous random variables is just a straightforward extension of that of the discrete case. All we need to do is replace the summation with an integral.

$$F(\mathrm{x}) = P(x \leq \mathrm{x}) = \int_{-\infty}^{\mathrm{x}} p(x)dx$$

Similar definitions

Conditional PDF

The conditional PDF of x given y is defined as:

$$p(x|y) \equiv \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \quad \textit{this equation is still called Bayes' Rule}$$

and

$$p(y) = \int_{\mathrm{dom}(x)} p(x,y)dx$$

Independence

Variables x and y are independent if knowing one event gives no extra information about the other event. Mathematically, this is expressed by

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

Independence of x and y is equivalent to

$$p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y)$$

Expectation

Expectation

Given a function f of x, the expectation of f with respect to the PDF $p(x)$ is defined as:

$$E[f(x)] = \int_{\mathrm{dom}(x)} f(x)p(x)dx$$

For instance, when $f(x) = x$, the expectation is

$$E[x] = \int_{\mathrm{dom}(x)} xp(x)dx$$

and it is equal to the mean of x.

Conditional Expectation

$$E[f(x)|y] = \int_{\mathrm{dom}(x)} f(x)p(x|y)dx$$

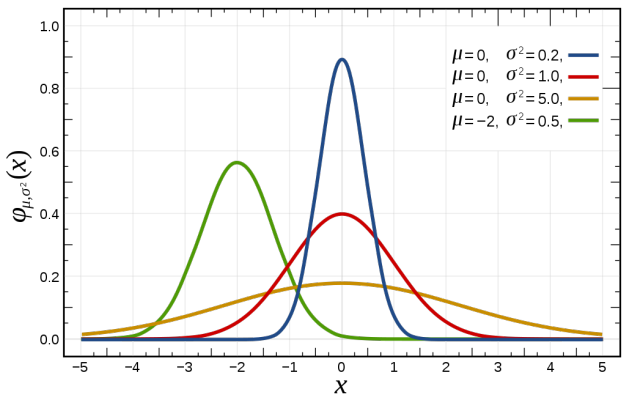
Probability Density functions or distributions?

Gaussian (or Normal) distribution has PDF:

- variable x
- $\mathrm{dom}(x) = \mathbb{R}$ (a real number);
- $p(x) := N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)}{2\sigma^2}\right)$;

It has two parameters

- $\mu \in \mathbb{R}$ is called mean;
- $\sigma \in \mathbb{R}$ with $\sigma \geq 0$ is called standard deviation;



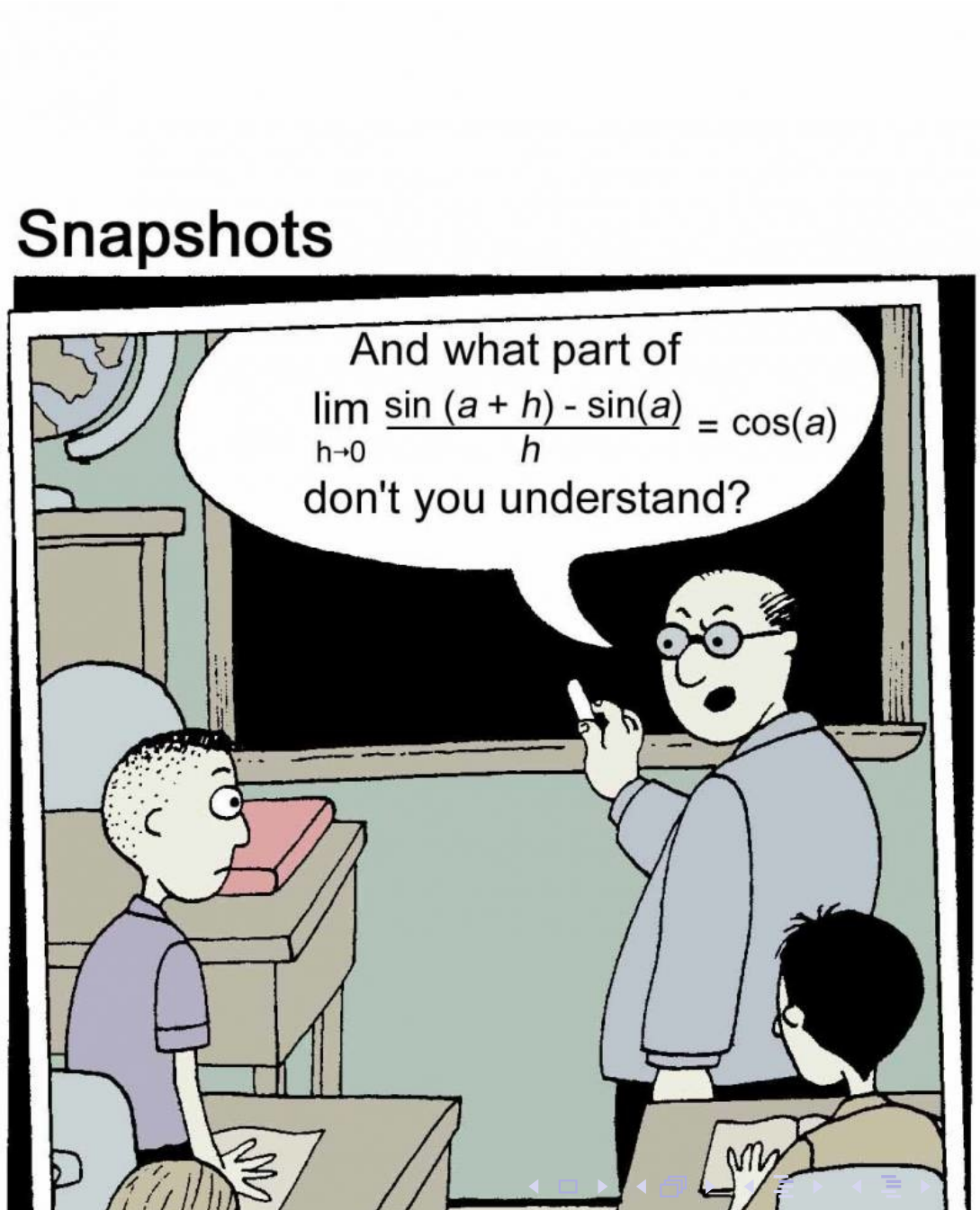
The Gaussian PDF is unimodal (it has a single maximum) and the value of x corresponding to the maximum is equal to the mean μ.
Note that

$$E[x] = \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)}{2\sigma^2}\right) dx = \mu$$

$$E[(x-\mu)^2] = \int (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)}{2\sigma^2}\right) dx = \sigma^2$$

that is why μ, σ^2 are called mean and variance.

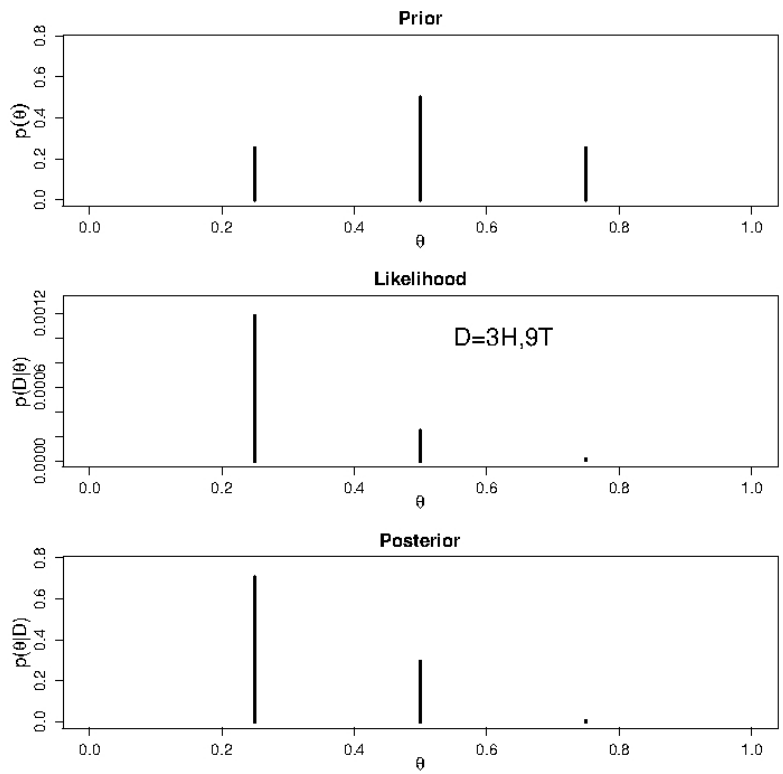
Don't Panic!


Snapshots
And what part of
$\lim_{h \to 0} \frac{\sin(a+h) - \sin(a)}{h} = \cos(a)$
don't you understand?

Is the coin fair?

- Denote with C a certain coin.
- Denote its bias with θ (the probability of getting Heads).
- **Query:** We want to know if C is fair (equivalently if $\theta = 0.5$)?
- Previously we considered the case in which only three values of θ are possible:
  - $\theta = 0.25$, $\theta = 0.5$, $\theta = 0.75$

  but why? The bias can be any number in $[0, 1]$!



Is the coin fair?

- Denote with C a certain coin.
- Denote its bias with θ (the probability of getting Heads).
- **Query:** We want to know if C is fair (equivalently if $\theta = 0.5$)?
- Previously we considered the case in which only three values of θ are possible:
  - $\theta = 0.25$, $\theta = 0.5$, $\theta = 0.75$

  but why? The bias can be any number in $[0, 1]$!

How can we assign probability to the values of the continuous interval $[0, 1]$?

We can for example use a probability density function (PDF) like

$$p(\theta) = 3\theta^2$$

we have seen that

$$3\theta^2 \geq 0 \quad \text{and} \quad \int_0^1 3\theta^2 d\theta = 1$$

so this is a valid PDF for the variable θ.

The Beta PDF

The PDF

$$p(\theta) = 3\theta^2$$

is a "Beta distribution".
That is it belongs to the following family of PDFs:

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where

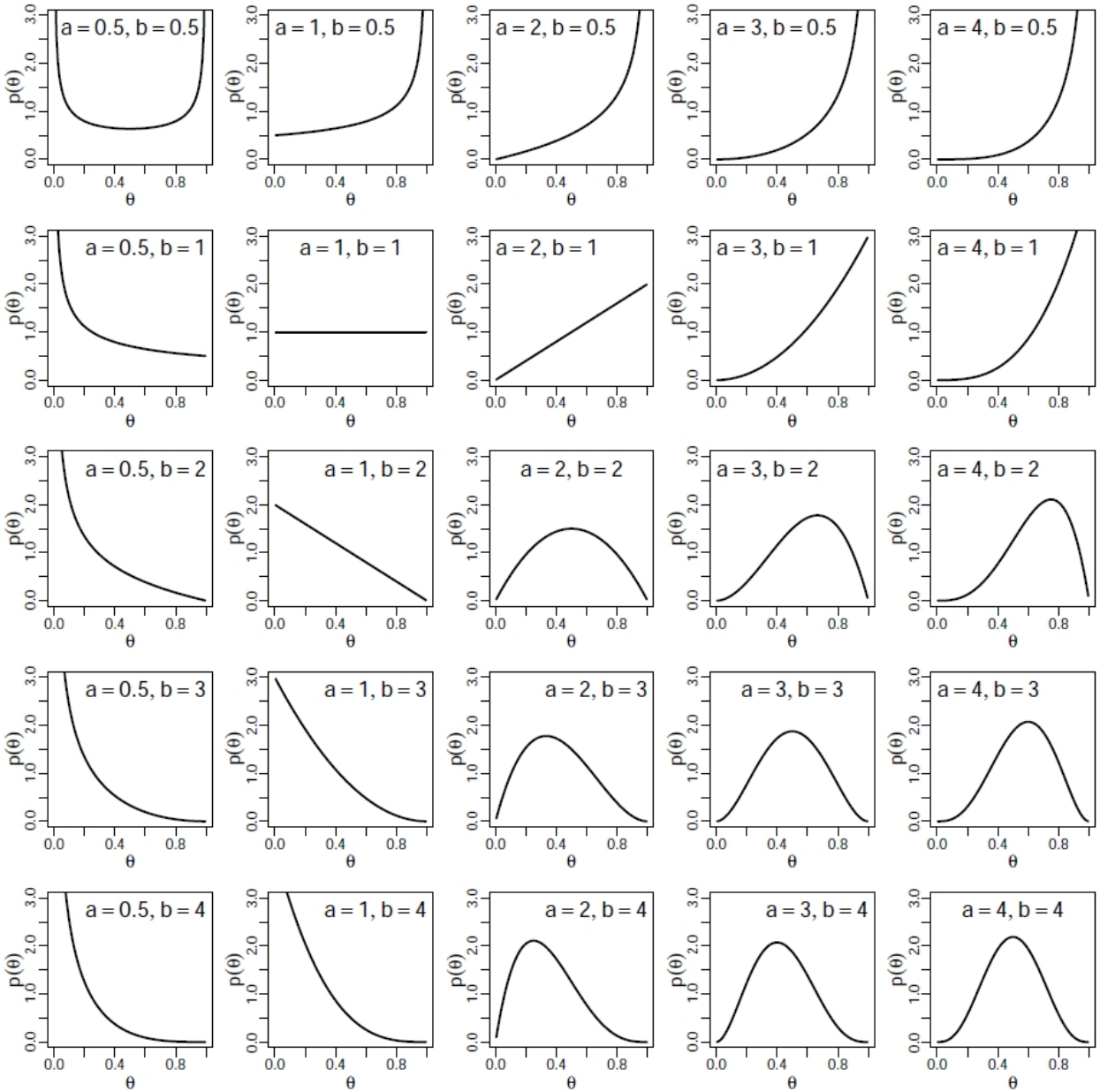
- $\alpha, \beta > 0$ are parameter, we can use α, β to change the shape of the PDF to match our prior beliefs about the bias of the coin.
- $B(\alpha, \beta) \equiv \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$ is the constant of normalisation.

Note that for $\alpha = 3$ and $\beta = 1$, we obtain

$$p(\theta) = \frac{1}{B(3,1)} \theta^2 = 3\theta^2$$

where $B(3,1) = \int_0^1 \theta^2 d\theta = \frac{1}{3}$.

Beta


a = 0.5, b = 0.5
a = 1, b = 0.5
a = 2, b = 0.5
a = 3, b = 0.5
a = 4, b = 0.5
a = 0.5, b = 1
a = 1, b = 1
a = 2, b = 1
a = 3, b = 1
a = 4, b = 1
a = 0.5, b = 2
a = 1, b = 2
a = 2, b = 2
a = 3, b = 2
a = 4, b = 2
a = 0.5, b = 3
a = 1, b = 3
a = 2, b = 3
a = 3, b = 3
a = 4, b = 3
a = 0.5, b = 4
a = 1, b = 4
a = 2, b = 4
a = 3, b = 4
a = 4, b = 4
p(θ)
θ
0.0
1.0
2.0
3.0
0.4
0.8

Posterior probability

Consider the following prior PDF:

$$p(\theta) = beta(\theta, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

In the previous lecture, we have seen that the likelihood of observing z Heads in N coin tosses is:

$$p(D|\theta) = \theta^z(1-\theta)^{N-z}$$

We aim to compute the posterior PDF:

$$p(\theta|D) = ?$$

this is the posterior belief about θ after observing D (i.e., z out of N).

How do we compute it?

Bayes' rule for PDFs

Bayes' rule for PDFs is:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}$$

From the posterior PDF, we can compute the probability that θ is in some set A:

$$P(\theta \in A|D) = \int_A p(\theta|D)d\theta$$

like for instance $A = [0.4, 0.5]$.

Posterior PDF

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

Note that

$$\begin{aligned} p(D|\theta)p(\theta) &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{z}(1-\theta)^{N-z} \\ &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha+z-1}(1-\theta)^{\beta+N-z-1} \end{aligned}$$

and, by definition,

$$P(D) = \int_0^1 p(D|\theta)p(\theta)d\theta$$

Posterior PDF

Note that:

$$\begin{aligned} P(D) &= \int_\Theta p(D|\theta)p(\theta)d\theta \\ &= \frac{1}{B(\alpha,\beta)} \int_\Theta \theta^{\alpha+z-1}(1-\theta)^{\beta+N-z-1}d\theta \end{aligned}$$

but we have seen that the definition of Beta function is

$$B(\alpha+z,\beta+N-z) = \int_\Theta \theta^{\alpha+z-1}(1-\theta)^{\beta+N-z-1}d\theta$$

and, therefore,

$$P(D) = \frac{B(\alpha+z,\beta+N-z)}{B(\alpha,\beta)}$$

Posterior PDF

Summing up:

$$p(\theta|D) = beta(\theta, \alpha + z, \beta + N - z) = \frac{1}{B(\alpha + z, \beta + N - z)} \theta^{\alpha+z-1}(1-\theta)^{\beta+N-z-1}$$

the posterior PDF is still a Beta distribution but with **updated** parameters $\alpha + z$ and $\beta + N - z$.

Let's see some examples.

$\alpha = \beta = 2$, $z = 3$ and $N = 5$

Prior

$$p(\theta) = beta(\theta, \alpha = 2, \beta = 2) = \frac{1}{B(2,2)}\theta(1-\theta)$$

Likelihood (we observe z=3H in N=5 tosses):

$$p(z = 3, N = 5|\theta) = \theta^3(1-\theta)^{5-3} = \theta^3(1-\theta)^2$$

Posterior:

$$\begin{aligned} p(\theta|z = 3, N = 5) = beta(\theta, \alpha = 2 + 3, \beta = 2 + 2) &= beta(\theta, \alpha = 5, \beta = 4) \\ &= \frac{1}{B(5,4)}\theta^{5-1}(1-\theta)^{4-1} \\ &= \frac{1}{B(5,4)}\theta^4(1-\theta)^3 \end{aligned}$$

```
from scipy.special import beta
beta(2,2)=1/6
beta(5,4)=1/280
```

General Recipe ML

What is the **general recipe ML** estimate of θ?

The Maximum Likelihood Estimator:

$$\hat{\theta} = \arg\max_{\theta} p(z = 3, N = 5|\theta)$$

The result is

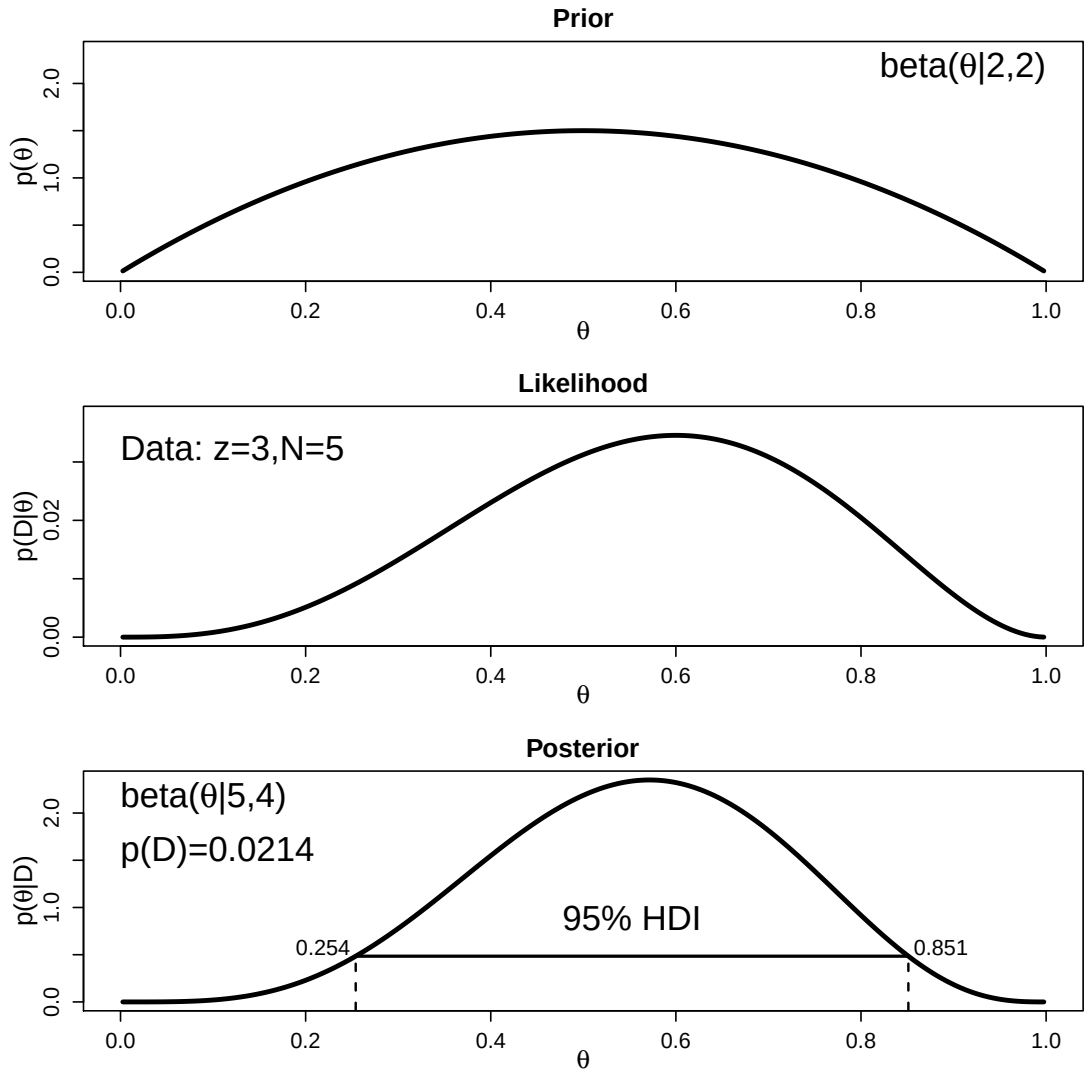
$$\hat{\theta} = \frac{3}{5} = 0.6$$

Is it bad?

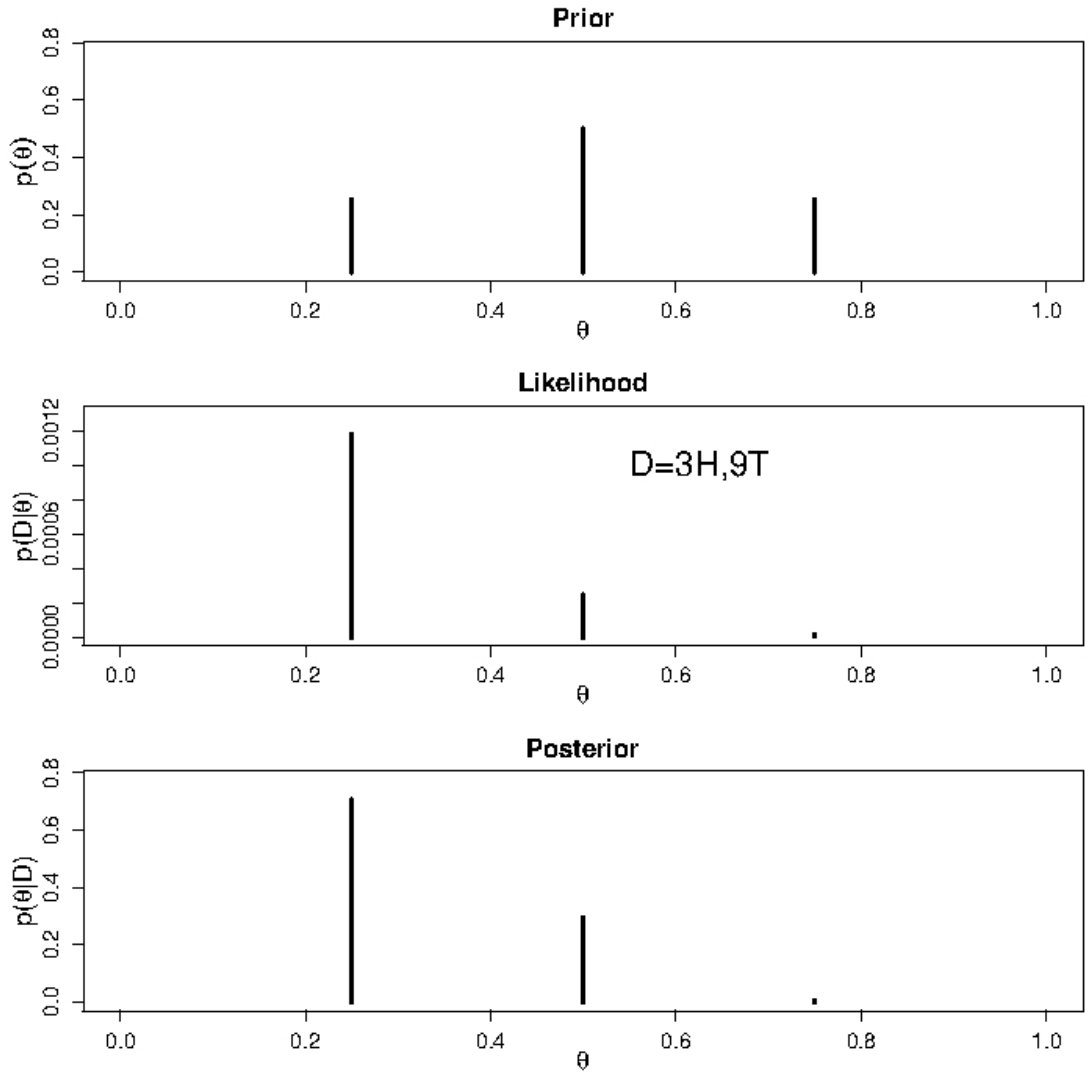
Yes it is very bad because

$$0.6 = \frac{3}{5} = \frac{30}{50} = \frac{300}{500} = \frac{3000}{5000}$$

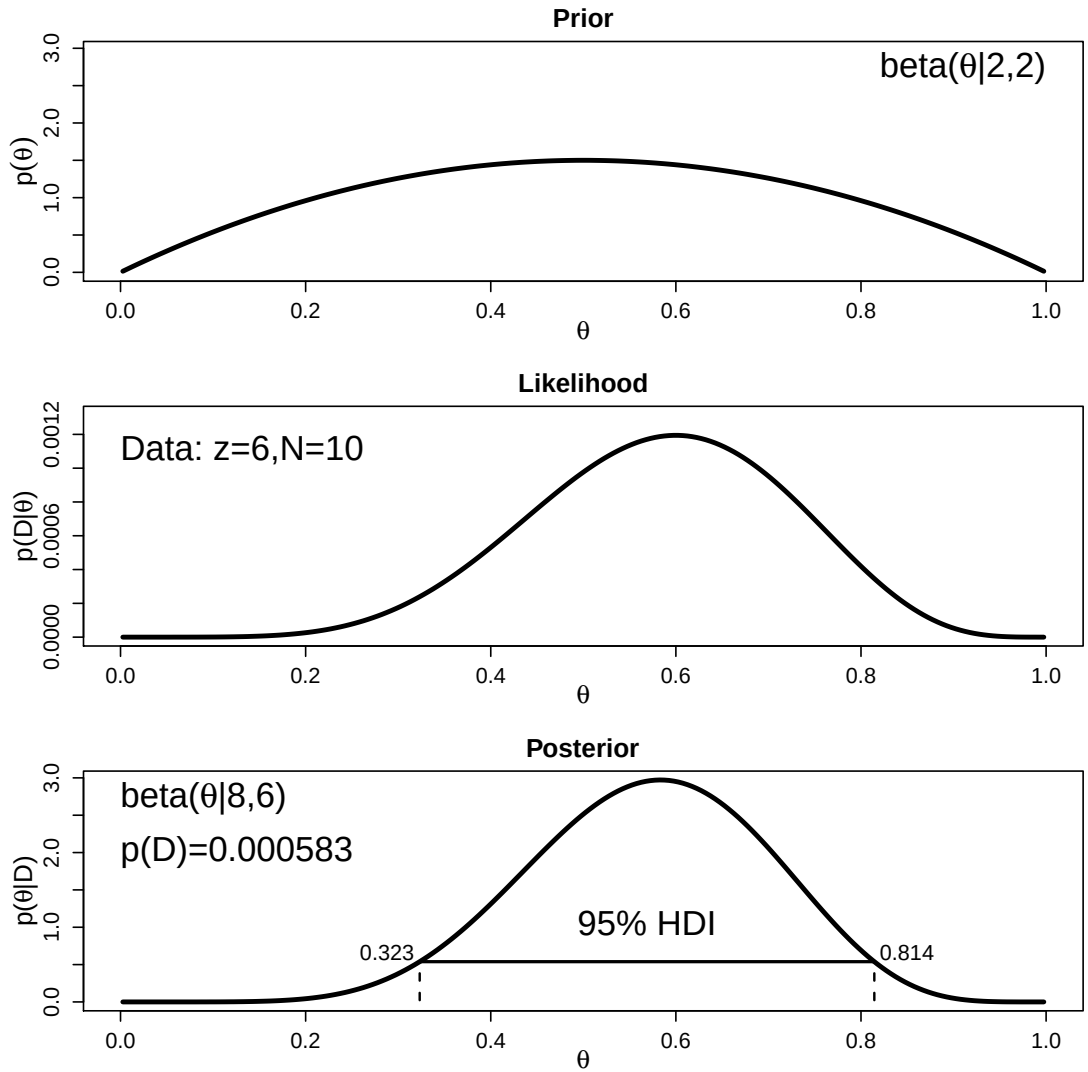
$\alpha = \beta = 2$, $z = 3$ and $N = 5$



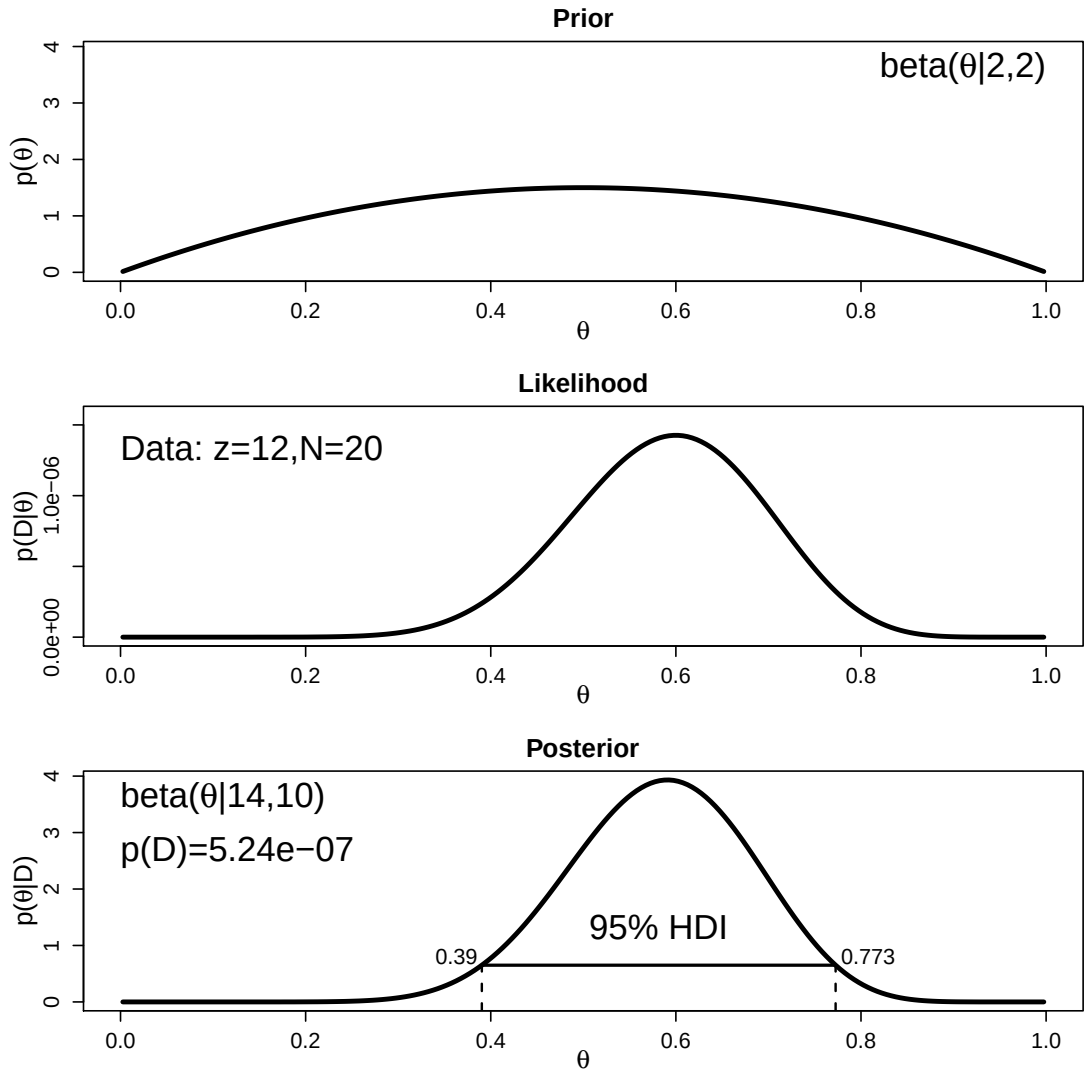
Compare with last week analysis



$\alpha = \beta = 2,\ z = 6$ e $N = 10$



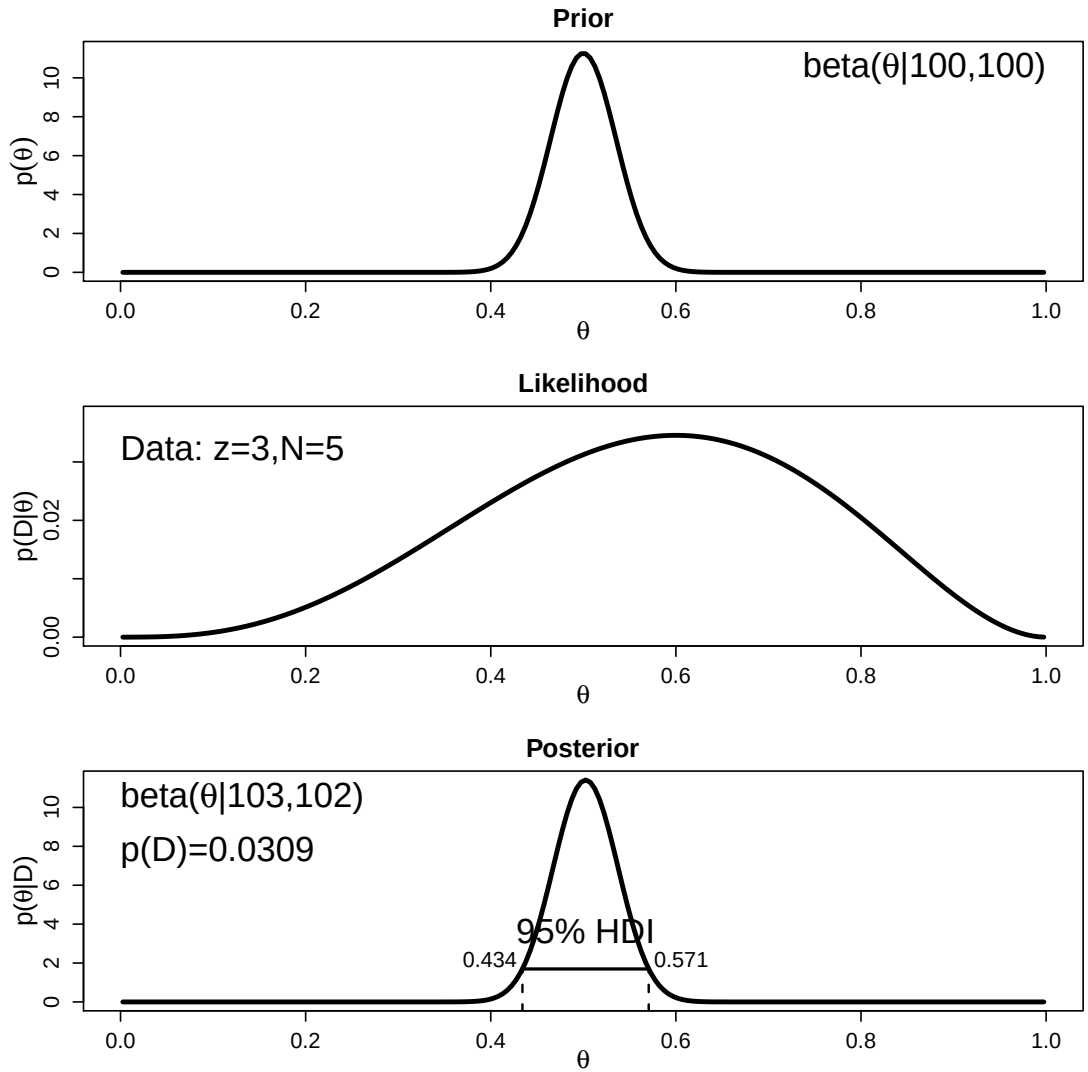
$\alpha = \beta = 2,\ z = 12$ e $N = 20$



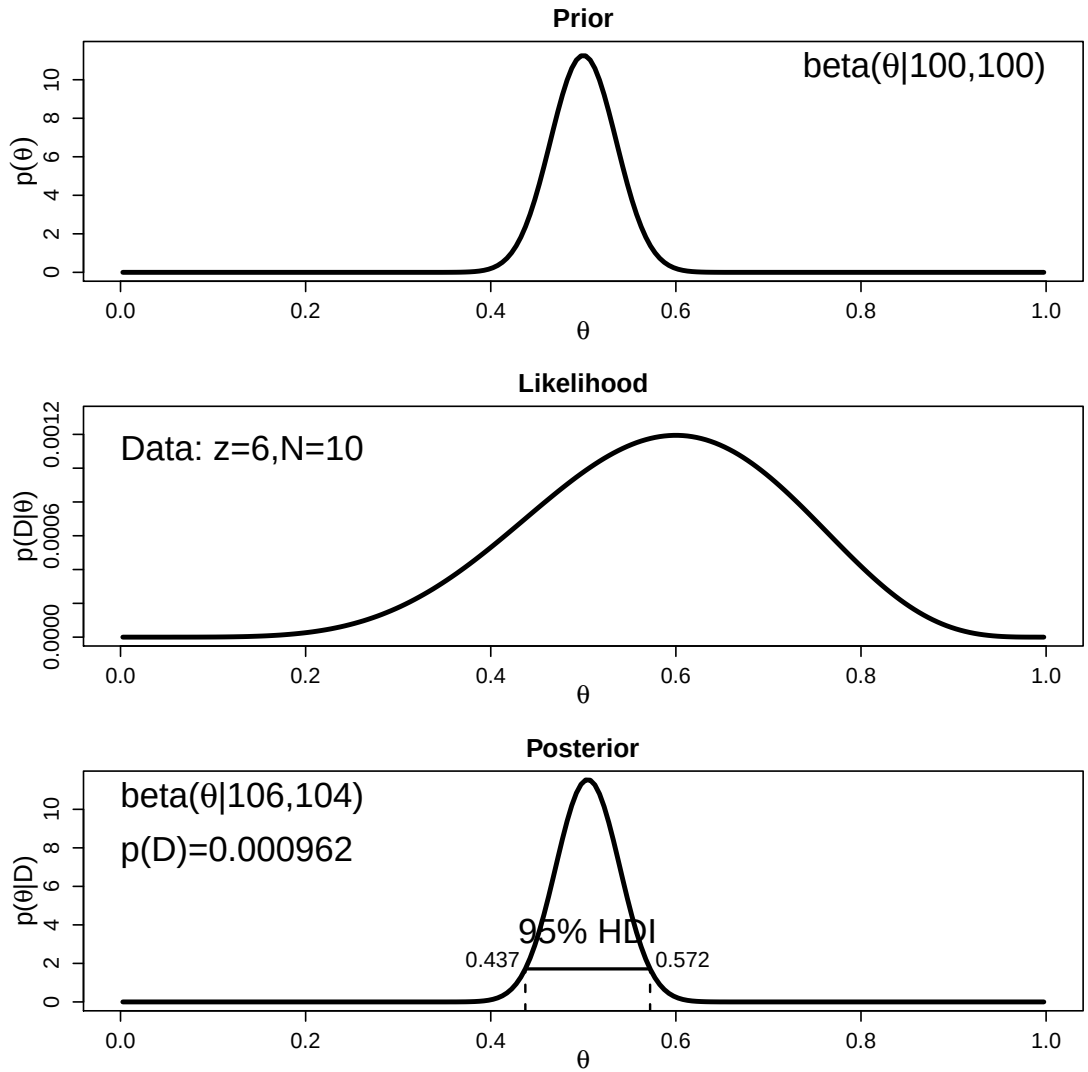
Remarks

- The prior PDF models the fact that, before observing the data (a-priori), we believe that the bias of the coin is 0.5, but we also believe that other values may be possible (although these other values are less and less probable closer to either 0 or 1).
- In all three cases the likelihood says that $\theta = 3/5 = 6/10 = 12/20 = 0.6$ is the most probable value (maximum likelihood estimation). Note that the likelihood becomes more concentrated around $\theta = 0.6$ at the increase of the number of observations N.
- The posterior probability, when the number of observations is low ($N = 5$), it is between the likelihood and the prior (note that the maximum is different from both 0.5 and 0.6).
- At the increase of the number of observations $N = 10, 20$, the posterior probability becomes concentrated around 0.6.
- Assume we now decide to toss a new official coin (from the state mint), the a-priori we should be very certain that $\theta = 0.5$. How do we model this case?

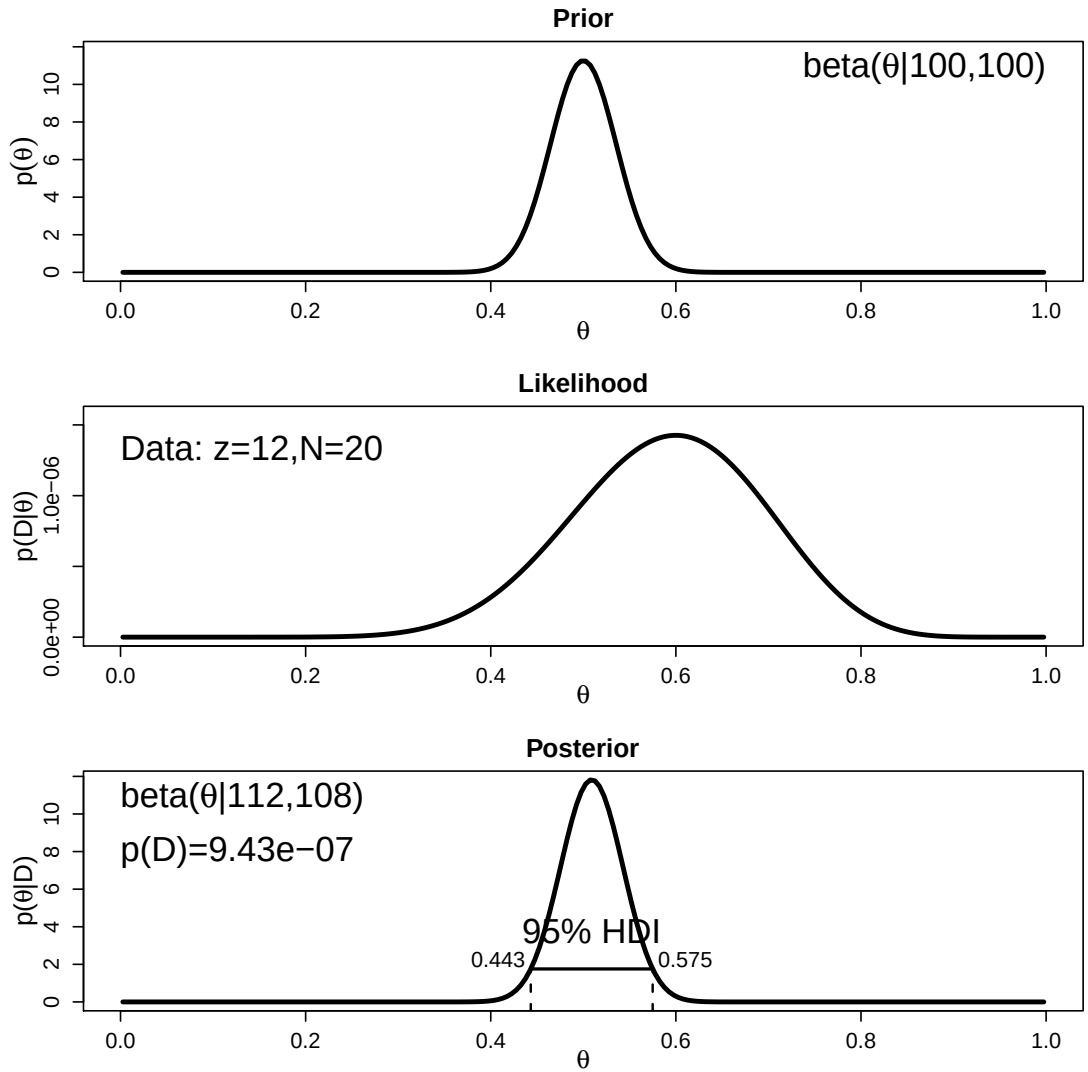
$\alpha = \beta = 100,\ z = 3$ e $N = 5$



$\alpha = \beta = 100,\ z = 6$ e $N = 10$



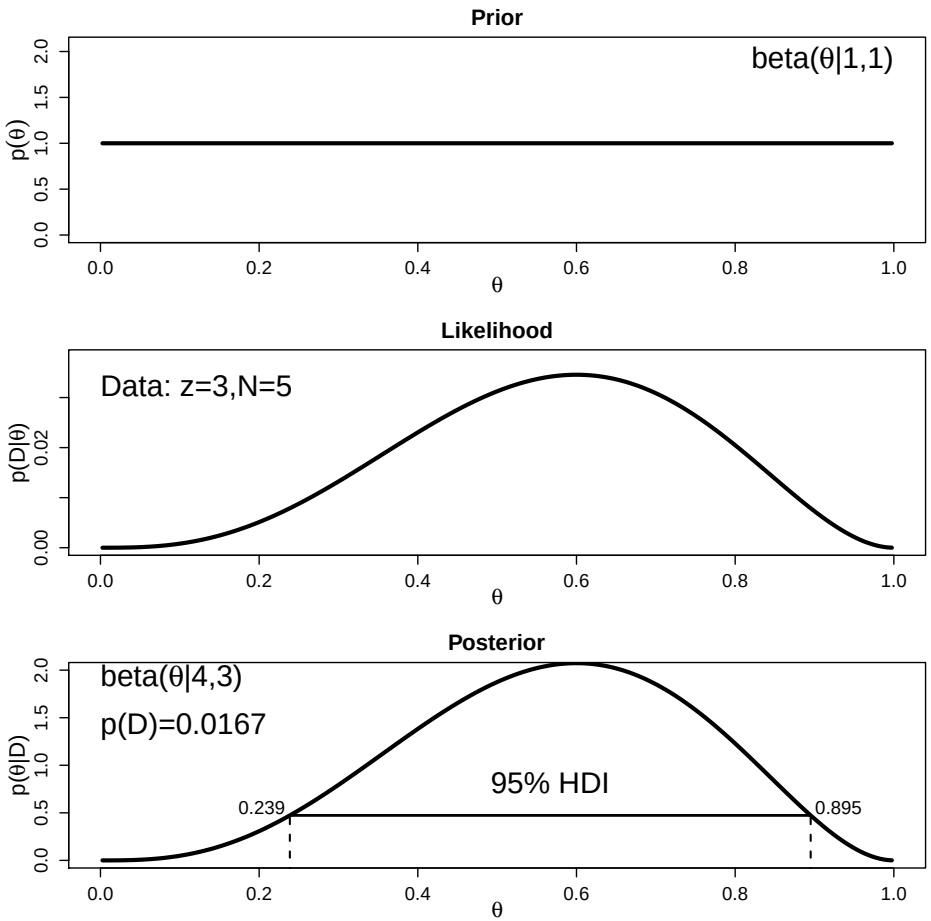
$\alpha = \beta = 100,\ z = 12$ e $N = 20$



Remarks

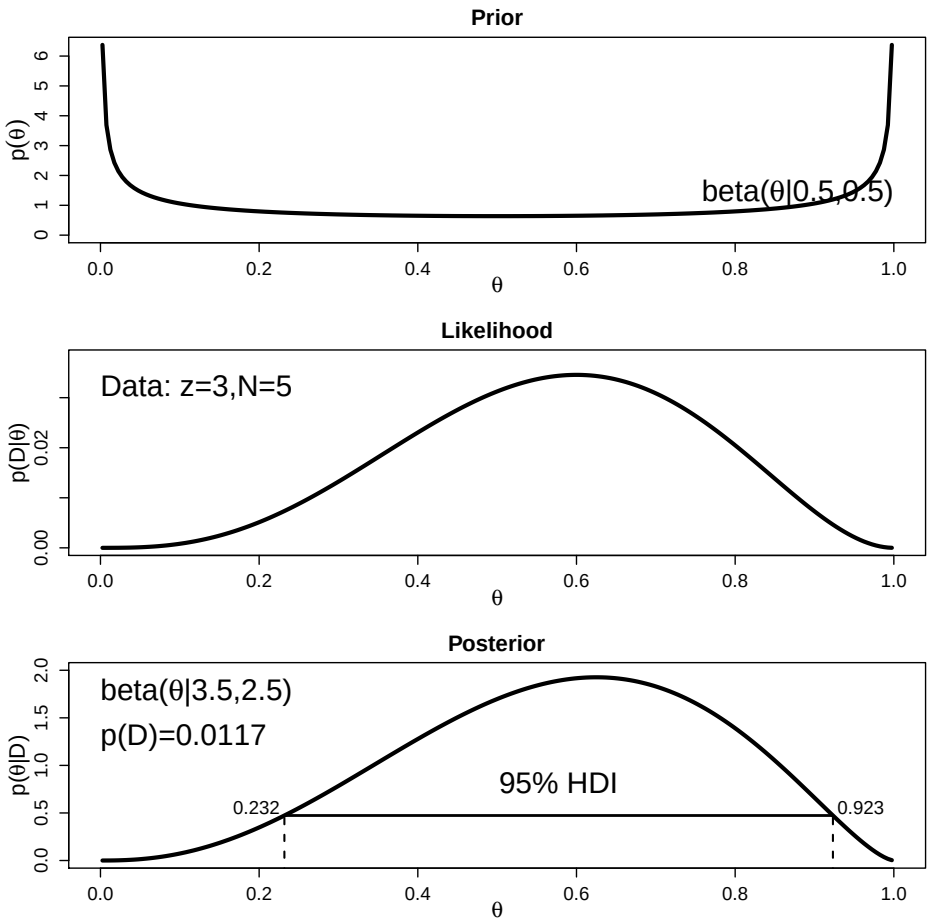
- Our prior belief is now very concentrated around 0.5. It means we strongly believe that the coin is unbiased ($\theta = 0.5$).
- The posterior probability is always basically equal to the prior. This means that even after $N = 20$ tosses we don't change much our belief about the coin.
- More observations are necessary to change our opinion.
- We have seen that at the increase of α, β our prior beliefs become stronger.
- What does it happen when α, β decrease?

$\alpha = \beta = 1$, $z = 3$ e $N = 5$



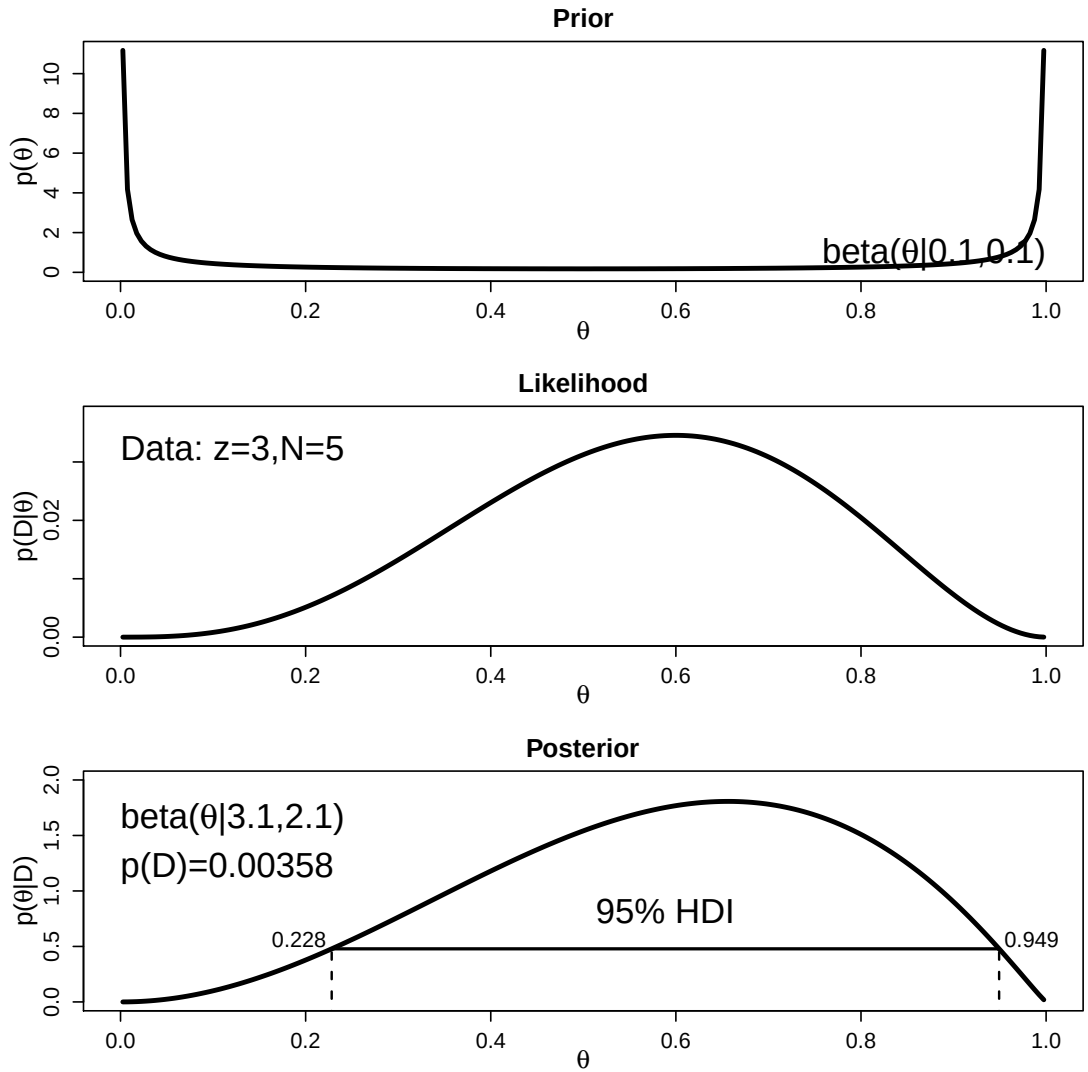
- This prior PDF is called **uniform**, because it is constant for all values of θ.
- It means that all values of θ are equally probable (a-priori).
- Note that even for $N = 5$, the posterior probability has the same shape of the likelihood.

$\alpha = \beta = 0.5$, $z = 3$ e $N = 5$



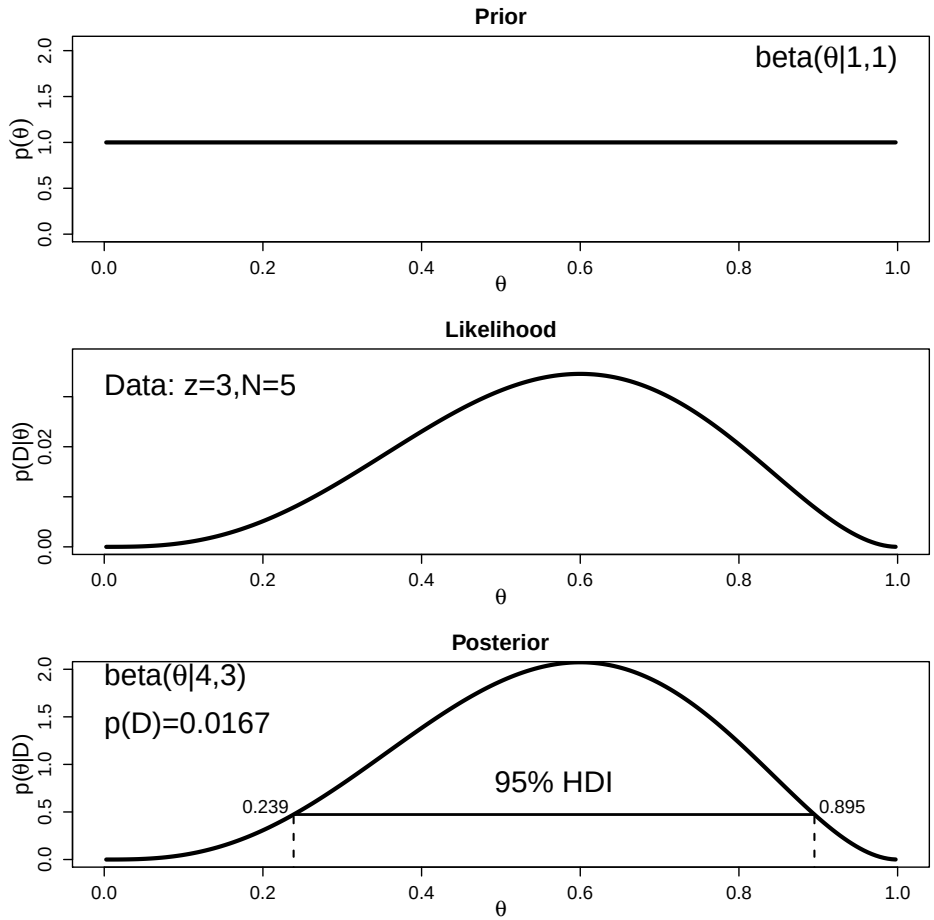
- Note that, this PDF is also constant for almost all values of θ but then it goes to infinity for 0 and 1.
- It models the fact we think the coin can be heavily tricked (two heads or two tails).
- In this case, as soon as we observe at least one heads and one tails we know that the coin is not heavily tricked. After that the prior becomes as the uniform distribution.
- Note in fact that, even with $N = 5$, the posterior probability has the same shape of the likelihood.

$\alpha = \beta = 0.1$, $z = 3$ e $N = 5$



Descriptors

Consider again the case $\alpha = \beta = 1$, $z = 3$ e $N = 5$. What does the PDF say to us?



Descriptors

- The posterior PDF has a maximum at 0.6, this value of θ is the posterior most probable value (maximum a-posteriori estimation).
- The posterior PDF is not symmetrical around the maximum, it means that we believe that the probability that θ is less than 0.6 is larger than the probability that θ is greater than 0.6.
- When the function is not symmetrical, its center of mass gives us more meaningfull information about θ. The center of mass is

$$\mu = E[\theta] = \int_0^1 \theta \cdot \frac{1}{B(4,3)} \theta^{4-1}(1-\theta)^{3-1} d\theta = \frac{4}{7} = 0.57$$

The center of mass is (by definition) the expected value (**mean**) of θ.

Descriptors

- Another interesting quantity is the variance

$$\sigma^2 = Var(\theta) = E[(\theta - E[\theta])^2] = \int_0^1 (\theta - E[\theta])^2 \frac{1}{B(4,3)} \theta^{4-1}(1-\theta)^{3-1} d\theta = 0.031$$

and its square root called **standard deviation**

$$\sigma = Std(\theta) = \sqrt{Var(\theta)} = 0.174$$

These quantities tell us how much the distribution is concentrated around the mean. Small variance means that the PDF is *narrow and peaked*, large variance means that the PDF is *large and flat*.

- If we integrate the posterior PDF in the interval $[0.239, 0.895]$ results that

$$P(\theta \in [0.239, 0.895]) = \int_{0.239}^{0.895} \frac{1}{B(4,3)} \theta^{4-1}(1-\theta)^{3-1} d\theta = 0.95$$

This means that the posterior probability that θ is in $[0.239, 0.895]$ is equal to 0.95. This interval is called 95% "**High Posterior Density Interval**" (HPDI).

Descriptor: mean and variance

For the Beta PDF $B(\theta, \alpha, \beta)$ the mean is equal to:

$$\mu = E[\theta] = \frac{\alpha}{\alpha + \beta}$$

and the variance is

$$E[(\theta - E[\theta])^2] = \frac{\mu(1-\mu)}{\alpha + \beta + 1}$$

that is they have a closed form.

Example ($\alpha = \beta = 2$, $z = 3$ and $N = 5$): The posterior mean is

$$\mu = E[\theta|\mathcal{D}] = \frac{\alpha + z}{\alpha + \beta + N} = \frac{2+3}{2+2+5} = \frac{5}{9} \approx 0.55$$

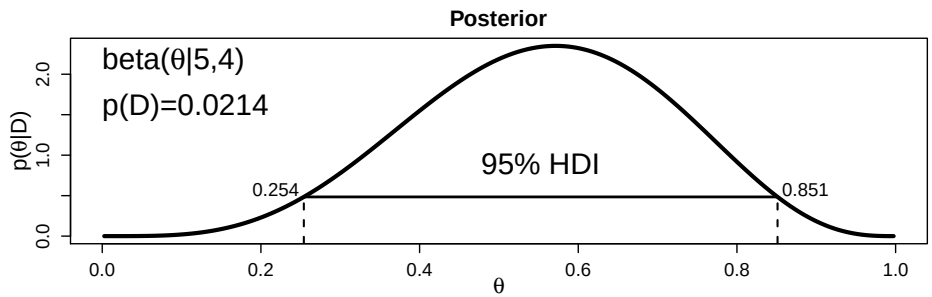
Meaning of mean and variance

$p(\theta)$ represents our degree of belief about the value of θ and:

- the mean of $p(\theta)$ can be interpreted as the value of θ that represents our typical (central) belief .
- the variance, that measures the dispersion of $p(\theta)$ around the mean, can be interpreted as our uncertainty about the possible values of θ. If the variance is small, then we strongly believe on the values of θ close to the mean. If the variance is large, then we are uncertainty about the mean.

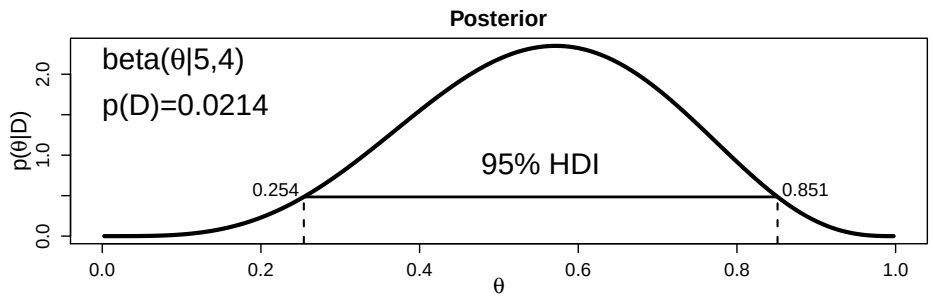
Highest Density Interval (HDI)

- Another descriptor of $p(\theta)$ is HDI.
- The 95% (or any other percentage) HDI is the smallest interval of θ that includes the 95% of the probability mass.



Highest Density Interval (HDI)

- Obviously we can compute any other interval 50% o 99% o 99.9% HDI.
- 95% HDI is also a measure of uncertainty. If the interval is large then we are quite uncertain about the value of θ, if it is small we are instead quite certain.



Why can the Bayesian inference be difficult?

- When we have a (many) continuous variable, we need to compute an (multiple) integral at the denominator

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}$$

  and this can be very very difficult.
- In the case of the Beta prior and Binomial likelihood, we can easily compute the denominator and the posterior PDF is still a Beta distribution.
- This is the reason we have selected the Beta family of distributions as prior PDF.
- Moreover, in this case, also the mean and variance can be computed in closed form.
- In more complex problems, we do not know how to solve the integral in closed form and we need to use numerical approximation (Monte Carlo methods, Variational approximation etc.).

Why can the Bayesian inference be difficult?

- Another difficulty is how to choose the prior probability ?
- The prior PDFs must be "reasonable" and must model our beliefs (information) about the parameter.
- The prior must not be selected using the data, otherwise we will count the same information twice and our posterior will be wrong.
- This is another reason to choose the Beta prior because its shape is very intuitive and also the meaning of its parameters α, β.
- It gives us expressiveness and flexibility. We will use this (and other standard priors) even when we cannot compute the integrals analytically.