Week 5 Module

Bayes Rule with Parameters and Data

Bayesian data analysis refers to a fairly specific application of this Bayes rule where:

- there is a statistical model of observed data, conditional on parameter values, $p(\mathcal{D}|\theta)$. We will use $\mathcal{D}$ to represent data and θ to represent the parameters. (Note the statistical model of observed data exists in a frequentist paradigm as well, as this function is sometimes called a *statistical likelihood* which is used for Maximum Likelihood Estimation (MLE).)

- there exists a **prior** belief on the parameter values, $p(\theta)$, in the form a probability distribution and

- Bayes rule is used to convert the prior belief on the parameters *and* the statistical model into a **posterior** belief $p(\theta|\mathcal{Y})$.

  $$p(\theta|\mathcal{Y}) = \frac{p(\mathcal{Y}|\theta)p(\theta)}{p(\mathcal{Y})}$$

  The denominator $p(\mathcal{Y})$ is referred to as the marginal likelihood of the data and is computed as: $p(\mathcal{Y}) = \sum_{\theta'} p(\mathcal{Y}|\theta^{'})p(\theta^{'})$ in the case where the parameters are discrete and $p(\mathcal{Y}) = \int p(\mathcal{Y}|\theta^{'})p(\theta^{'})d\theta^{'}$ in the case where the parameters are continuous. The $\theta^{'}$ is used as we enumerate or integrate across all possible values of the parameter values.

Example of Bayesian Analysis on a binary outcome

Consider estimating the probability that a die will roll a six based on the results from a large number of rolls

Define a descriptive statistical model for the relevant data.

A descriptive model denoted as $p(\mathcal{Y}|\theta)$ is needed for the die rolling experiment.

- what is $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$?

- what is θ?

- what is a descriptive model for $p(\mathcal{Y}|\theta)$, For a single roll of the die, $y_i = 1$ if a 6 is rolled and $y_i = 0$ otherwise, use the Bernoulli distribution:

$$p(y_i = 1|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

- By assuming that each die roll is independent, they can be combined as:

$$\begin{aligned} p(\mathcal{Y}|\theta) &= \prod_i p(y_i|\theta) \\ &= \prod_i \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_i y_i}(1-\theta)^{\sum_i (1-y_i)} \\ &= \theta^{\sum_i y_i}(1-\theta)^{N-\sum_i y_i} \\ &= \theta^{\#\text{ of heads}}(1-\theta)^{\#\text{ of tails}} \end{aligned}$$

With the probability distribution function, or as it is sometimes called the sampling function, the function is about the data y conditioned upon the the parameter(s) θ.

Now consider the data value y to be fixed and the function to be a function of the parameter θ. This is known as the *Likelihood Function*, as it provides a measure of how likely a parameter value θ is conditioned on the data. Points about a likelihood function:

- in this case, the function is about about a

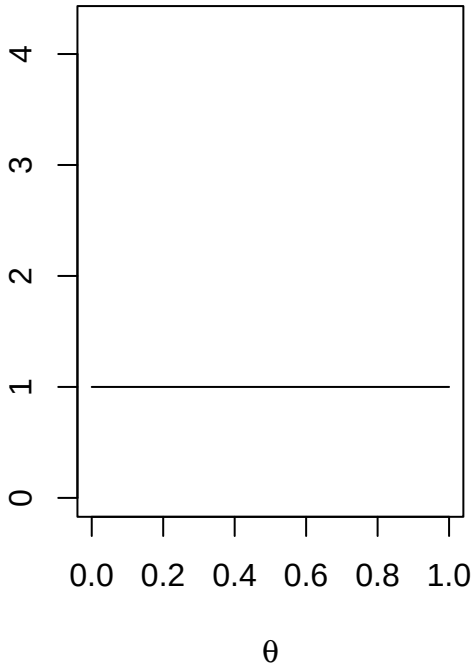
- the likelihood is not a probability distribution function - as it does not integrate to one.

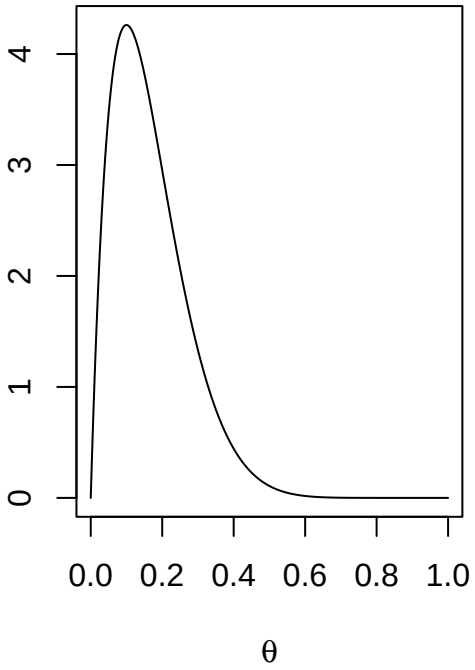
- in classical statistics, this function is maximized

For notational purposes, I'll usually denote the likelihood function as $\mathcal{L}(\theta|y)$, but unfortunately it can also be denoted as $p(y|\theta)$.

Specify a prior distribution on the parameters

Here are a couple of reasonable prior distributions on θ, the probability of rolling a 6.





Each figure can be formulated in terms of a Beta distribution:

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

for $\theta \in [0,1]$ and 0 otherwise.

- The first figure has: $\alpha = 1$ and $\beta = 1$, which results in a uniform distribution.
- The second figure has: $\alpha = 2$ and $\beta = 10$.

We will see that:

- a corresponds to
- b corresponds to
- $a + b$ corresponds to
- as $a + b$ gets large, the distribution get tighter around
- the larger a relative to b the distribution has more density closer
- the larger b relative to a the distribution has more density closer

- What is $\int \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\theta^{(a-1)}(1-\theta)^{(b-1)}d\theta =$

- What is $\int \theta^{(a-1)}(1-\theta)^{(b-1)}d\theta =$

- What is

$$\int \theta^{(a-1)}(1-\theta)^{(b-1)}\theta^{z}(1-\theta)^{(N-z)}d\theta \quad = \quad \int$$

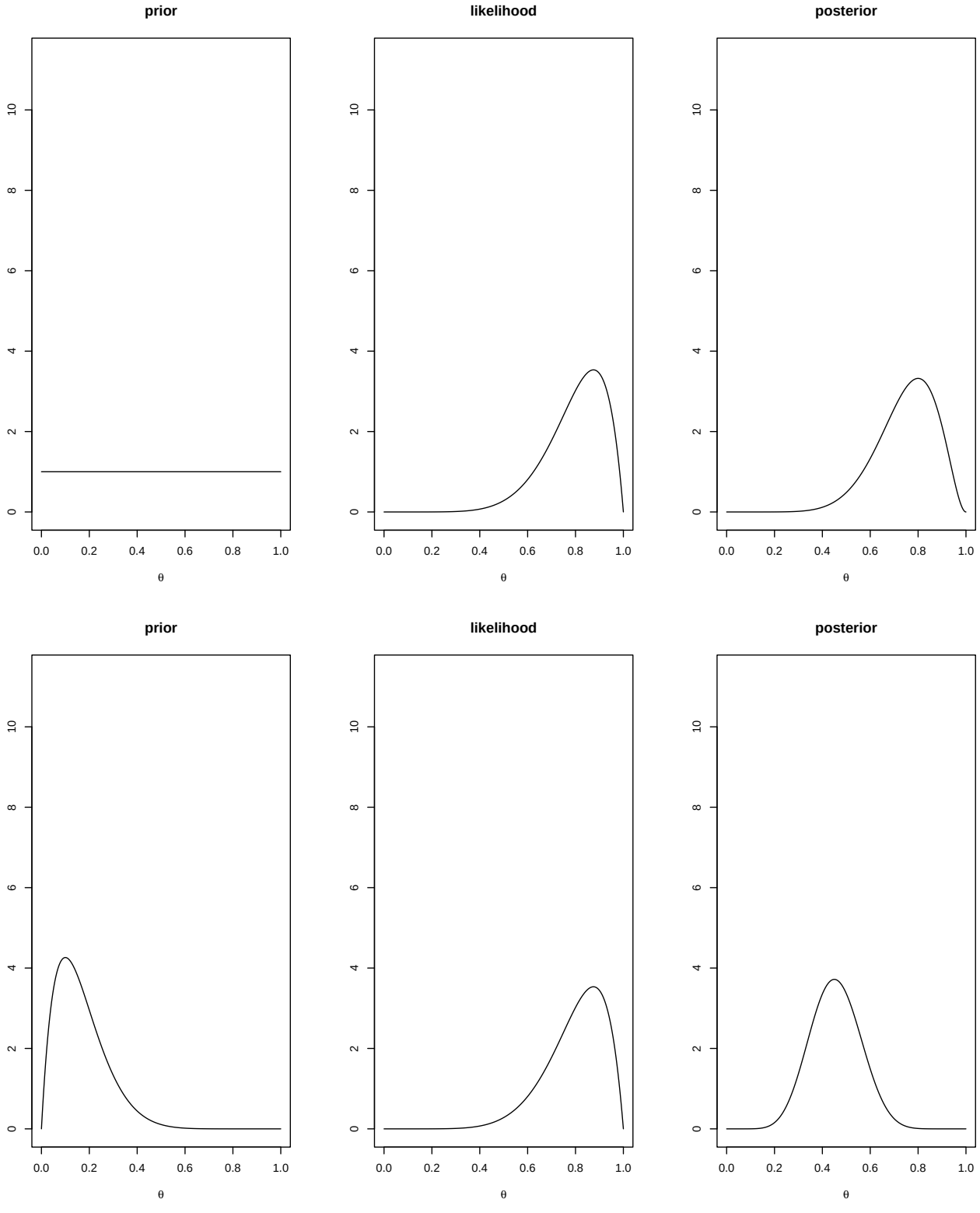
Use Bayesian inference to re-allocate credibility across parameter values.

Recall the goal of this analysis was to learn about θ the probability of rolling a six. Specifically, we are interested in the posterior distribution $p(\theta|\mathcal{Y})$.

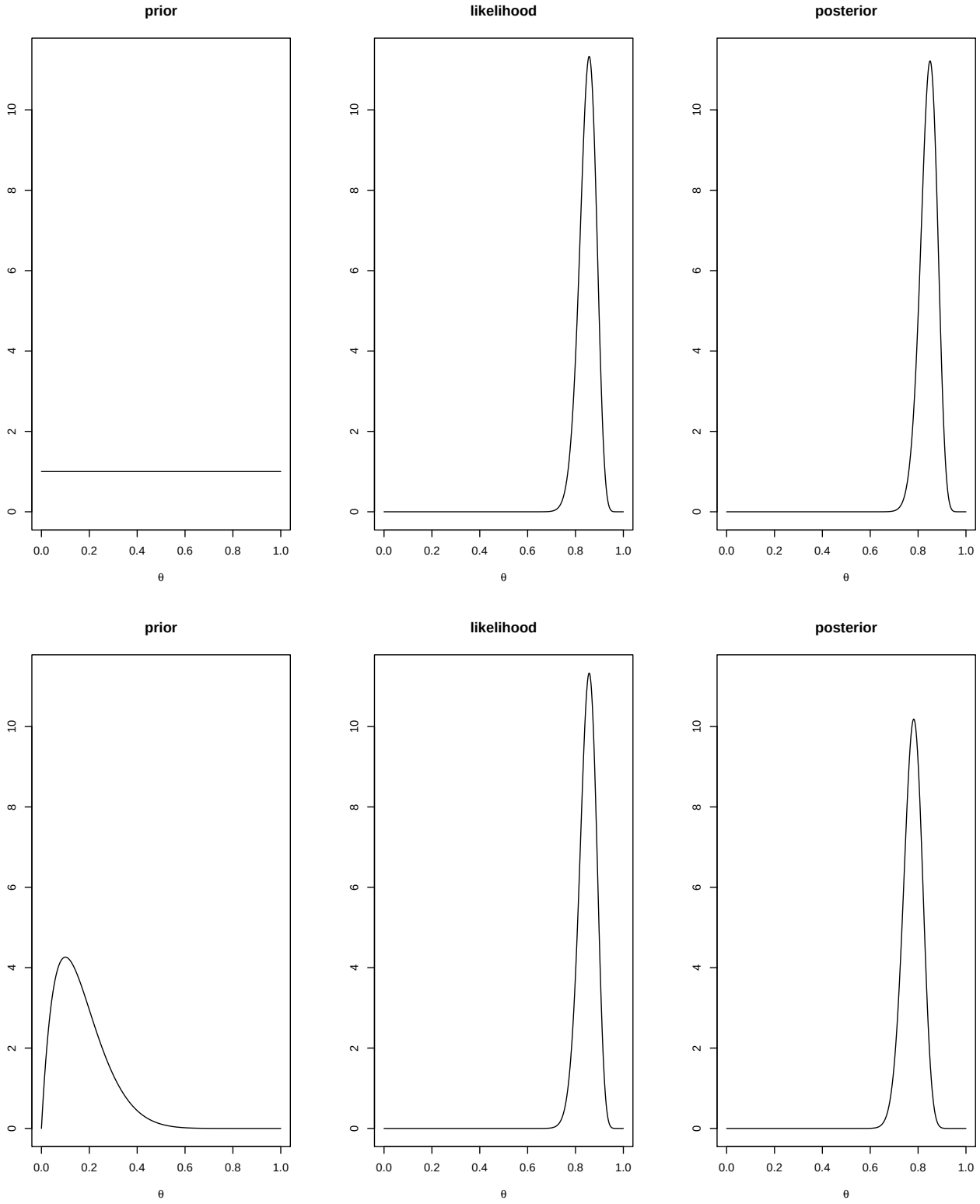
Lets assume a few data collection procedures:

1. 10 rolls of the die, with 8 6's
2. 100 rolls of the die, with 85 6's

With 10 rolls



Then with 100 rolls



Posterior Distribution

The goal of the analysis is to infer $p(\theta|z, N)$, in other words we want to learn about the parameter θ from a series of N trials with z successes.

Assume we use a Beta(a,b) as a prior distribution on θ.

$$\begin{aligned} p(\theta|z,N) &= \frac{p(z,N|\theta)p(\theta)}{p(z,N)} \\ &= \frac{p(z,N|\theta)p(\theta)}{\int p(z,N|\theta)p(\theta)d\theta} \\ &= \frac{\theta^z(1-\theta)^{(N-z)} \times (\Gamma(a)\Gamma(b)/\Gamma(a+b))\theta^{(a-1)}(1-\theta)^{(b-1)}}{\int \theta^z(1-\theta)^{(N-z)} \times (\Gamma(a)\Gamma(b)/\Gamma(a+b))\theta^{(a-1)}(1-\theta)^{(b-1)}d\theta} \end{aligned}$$

Posterior is a compromise of likelihood and prior For more intuition, consider the posterior mean:

$$\frac{z+a}{N+a+b} \tag{1}$$

The posterior mean is a weighted average of the data mean ($\frac{z}{N}$) and the prior mean ($\frac{a}{a+b}$), where the weights are ($\frac{N}{N+a+b}$) for the data piece and ($\frac{a+b}{N+a+b}$) for the prior.