DATA 582: Bayesian Inference

Lecture 2: Bayesian Updating

Dr. Irene Vrbik

UBCO MDS

Put generally, the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior probability distribution for the parameter that contains less uncertainty.[1]

[1] SL pg 50

The four steps can be summarized as:

1. Define the prior
2. Gather data, summarize in a likelihood
3. Update your prior distribution via Bayes' theorem to obtain a posterior distribution
4. Analyze the posterior distribution and summarize it, eg. mean, sd

Recap:

Recall Bayes theorem[2] as it is used in Bayesian inference:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \qquad \text{in Bayesian language}$$

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \qquad \text{in symbols}$$

where

- $p(\theta)$ is our *prior distribution* for the parameter θ
- $p(y \mid \theta)$ is our *likelihood function* for the data x
- $p(\theta \mid y)$ is our *posterior distribution* for the parameter θ
- $p(y)$ is a *marginal probability* of the data

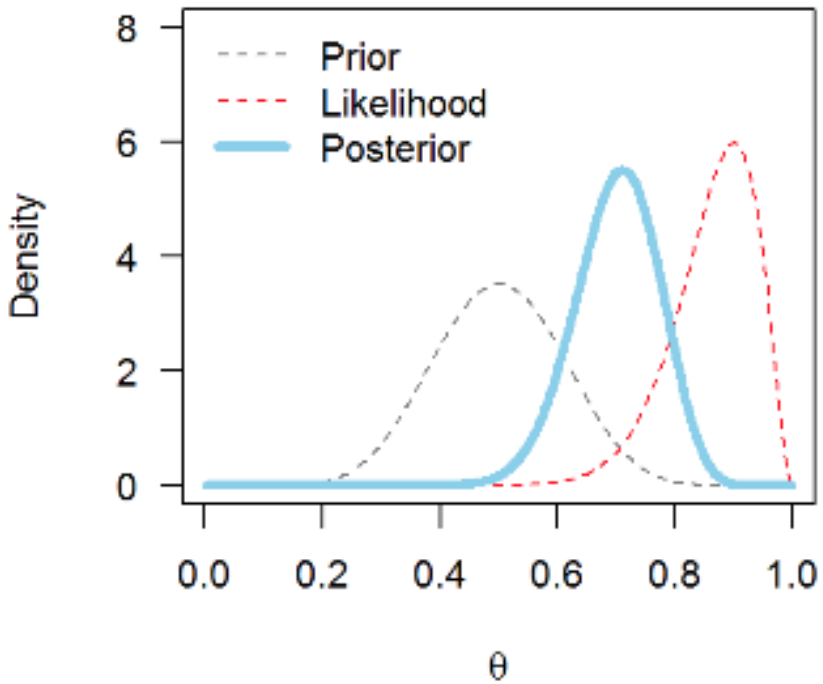
[2] or "Bayes rule"

- For a continuous sample space, $p(y)$, AKA Bayes denominator or marginal likelihood, produces a scaling factor that normalizes posterior so that $\int p(\theta \mid y) d\theta = 1$.
- In practice, this is often hard to calculate.
- Luckily, we can avoid it entirely by noting:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \times p(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned} \tag{1}$$

- When the posterior is constructed in this way we call it the *unnormalized posterior* distribution since it will not general integrate/sum to 1.

- We often see the tug-of-war between these three important functions displayed on a single "triplot".[3]



- The general trend is that posterior distribution lies somewhere between the prior and the likelihood.

[3] Image sourced from: *(Pesky?) Priors*, a blog post written by Jim Grange

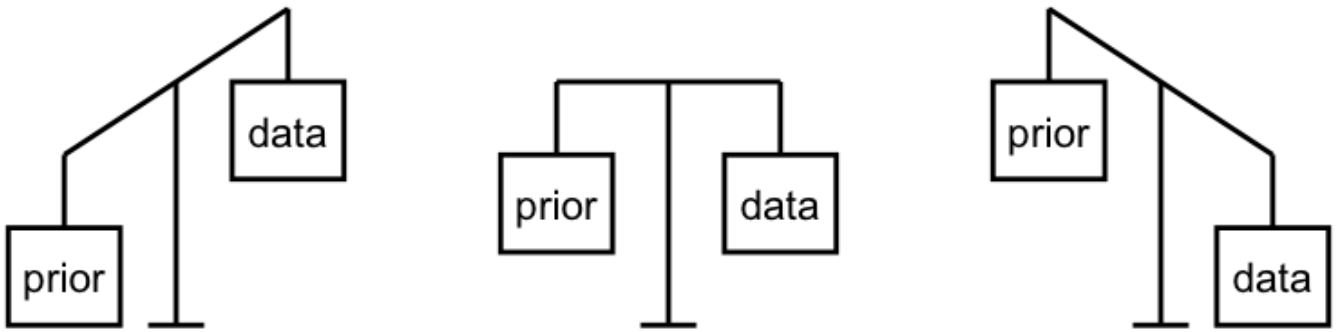


Figure: Bayesian analyses balance our prior experiences with new data. Depending upon the setting, the prior is given more weight than the data (left), the prior and data are given equal weight (middle), or the prior is given less weight than the data (right) [Source: Ch 1 of Bayes Rules!]

- Remember, in the Bayesian framework θ is being treated as a random variable which is why we can assign a probability distribution to it.

- Today we will discuss the process of *Bayesian updating* – that is, how exactly do we revise our prior distribution with the information provided in our likelihood function to obtain a posterior distribution for our population parameter θ.

- First let's discuss in greater detail what each of these terms represent individually . . .

Likelihood

- The *likelihood function* (or simply *likelihood*) plays an important role in both Bayesian and Frequentist frameworks.
- The likelihood, $p(y \mid \theta)$, is written as conditional pdf $p(\cdot \mid \cdot)$, however, it is generally *not* a proper[4] probability distribution.
- Notice in (1) how our construction of the posterior relies on y solely through the $p(y \mid \theta)$, hence the likelihood contains **all the evidence in a sample relevant to parameter** θ.
- As such, any two samples that have the same likelihood will yield the same inference for θ - this is known as the *likelihood principle*.

[4] A proper pdf/pmf is one that integrates/sums to the unity.

Likelihood

- Suppose Y_1, Y_2, ..., Y_n are random variables sampled from the same population assumed to conditionally independent given θ.
- Based on the multiplication rule for independent events we can construct the likelihood in the following way:

$$p(y_1, y_2, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) \tag{2}$$

- Writing the data as a vector $y = (y_1, \ldots, y_n)$ we have:

$$Y_1, \ldots, Y_n \overset{i.i.d}{\sim} p(y \mid \theta)$$

Likelihood

- You have probably seen the likelihood being used in the context of *maximum likelihood estimation* (a Frequentist technique)
- In words, the MLE corresponds to the value of the parameter θ that makes the occurrence of the data most likely to have occurred.
- Mathematically, the MLE is given by

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i \mid \theta)$$

- We often like to think of the likelihood as a function of θ (where θ is our generic parameter of interest) and the data as begin fixed and write: $\mathcal{L}(\theta \mid y)$ or even $\mathcal{L}(\theta)$.

Coin Example

A coin is tossed ten times to obtain the following data.

Roll number:	1	2	3	4	5	6	7	8	9	10
Result	H	H	T	H	H	H	T	H	T	H

In total this accounts for 7 "heads" and 3 "tails". What is the likelihood function for these data?

Coin toss example

- Lets define a random variable Y_i for the outcome of the ith toss. More specifically let

$$Y_i = \begin{cases} 0 & \text{if coin lands T} \\ 1 & \text{if coin lands H} \end{cases}$$

What distribution does Y_i follow?

Coin toss example

- Lets define a random variable Y_i for the outcome of the ith toss. More specifically let

$$Y_i = \begin{cases} 0 & \text{if coin lands T} \\ 1 & \text{if coin lands H} \end{cases}$$

What distribution does Y_i follow?

$$Y_i \sim Bernoulli(\theta)$$

where $\theta \in [0, 1]$ is the probability of Heads.

- Recall the pmf of a Bernoulli(θ) random variable:

$$p(y_i) = \theta^{y_i}(1-\theta)^{1-y_i} \quad \text{for } y_i \in \{0,1\}$$

where θ is probability of the coin landing heads.

- Since coin flips are independent, we simply plug in the above PMF to the likelihood function given in (2) :

$$\mathcal{L}(\theta) = p(y \mid \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Coin Toss Example

Converting our result:

Result $= H, H, T, H, H, H, T, H, T, H$ to

$y = (y_1, y_2, \ldots y_n) = (1, 1, 0, 1, 1, 1, 0, 1)$,

we get ...

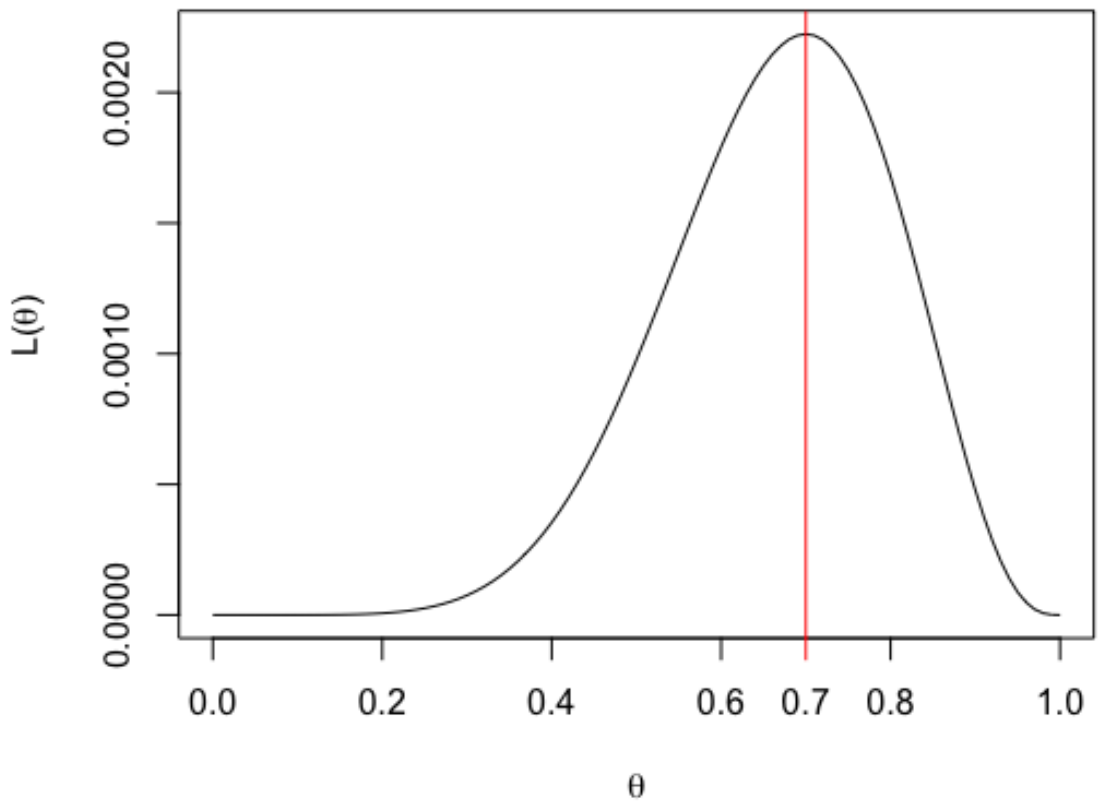
$$
\begin{aligned}
p(y \mid \theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\
&= \theta^{y_1}(1-\theta)^{1-y_1} \times \cdots \times \theta^{y_{10}}(1-\theta)^{1-y_{10}} \\
&= \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \times \theta^{0}(1-\theta)^{1} \times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \\
&\times \theta^{1}(1-\theta)^{0} \times \theta^{0}(1-\theta)^{1} \times \theta^{1}(1-\theta)^{0} \times \theta^{0}(1-\theta)^{1} \times \theta^{1}(1-\theta)^{0} \\
&= \theta^{7}(1-\theta)^{3}
\end{aligned}
$$

Likelihood Function: Coin example

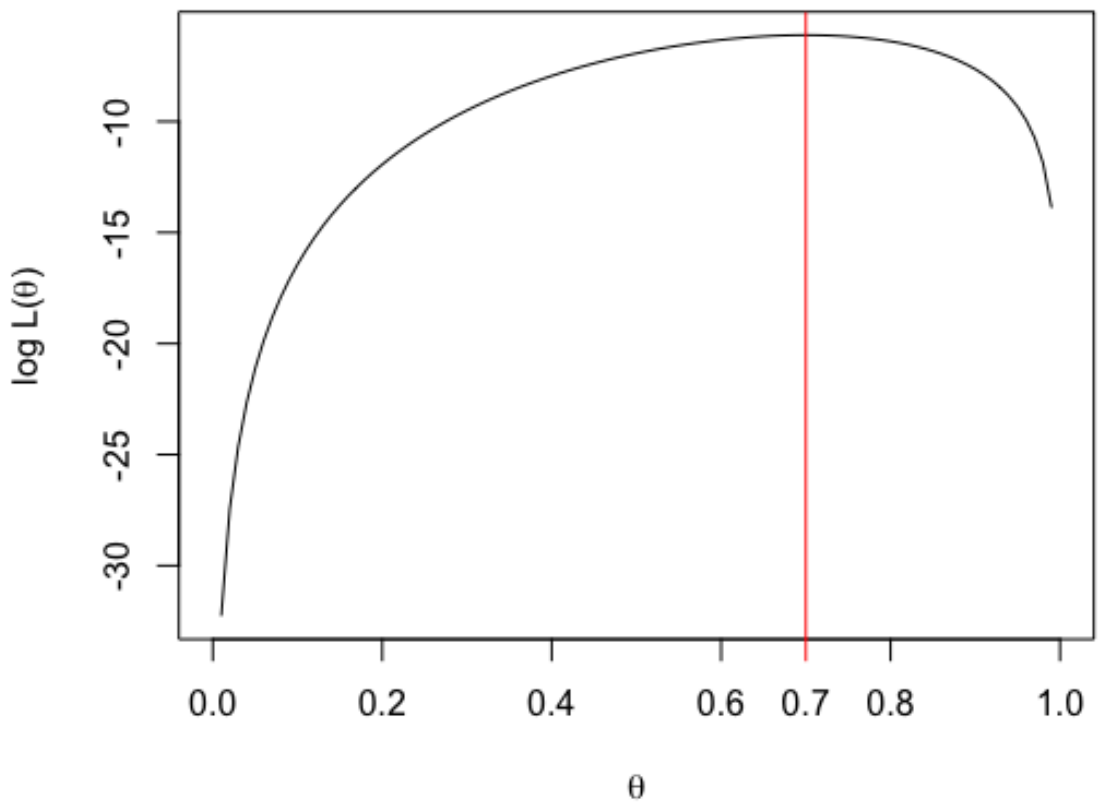
Adopting the MLE approach, we aim to find the value of θ which maximizes $\mathcal{L}(\theta) = \theta^7(1-\theta)^3$. We can arrive at this answer using derivatives:

Likelihood Function: Coin example

Graphically, the MLE is located at the peak of this curve:



This maximum occurs at the same value on the log-likelihood curve $\ell(\theta) := \log \mathcal{L}(\theta) = 7 \log(\theta) + 3 \log(1 - \theta)$



Notes on the likelihood

- Notice that if we had observed 7 heads and 3 tails in any other order, eg:
  - Result 1 = TTTHHHHHHH
  - Result 2= TTHTHHHHHH,
  - Result 3 = HHHHHHHTTT, ...

  each would produce the exact same likelihood.

- This relates back the likelihood principle.

- All three of the results above provide the same inference for θ...

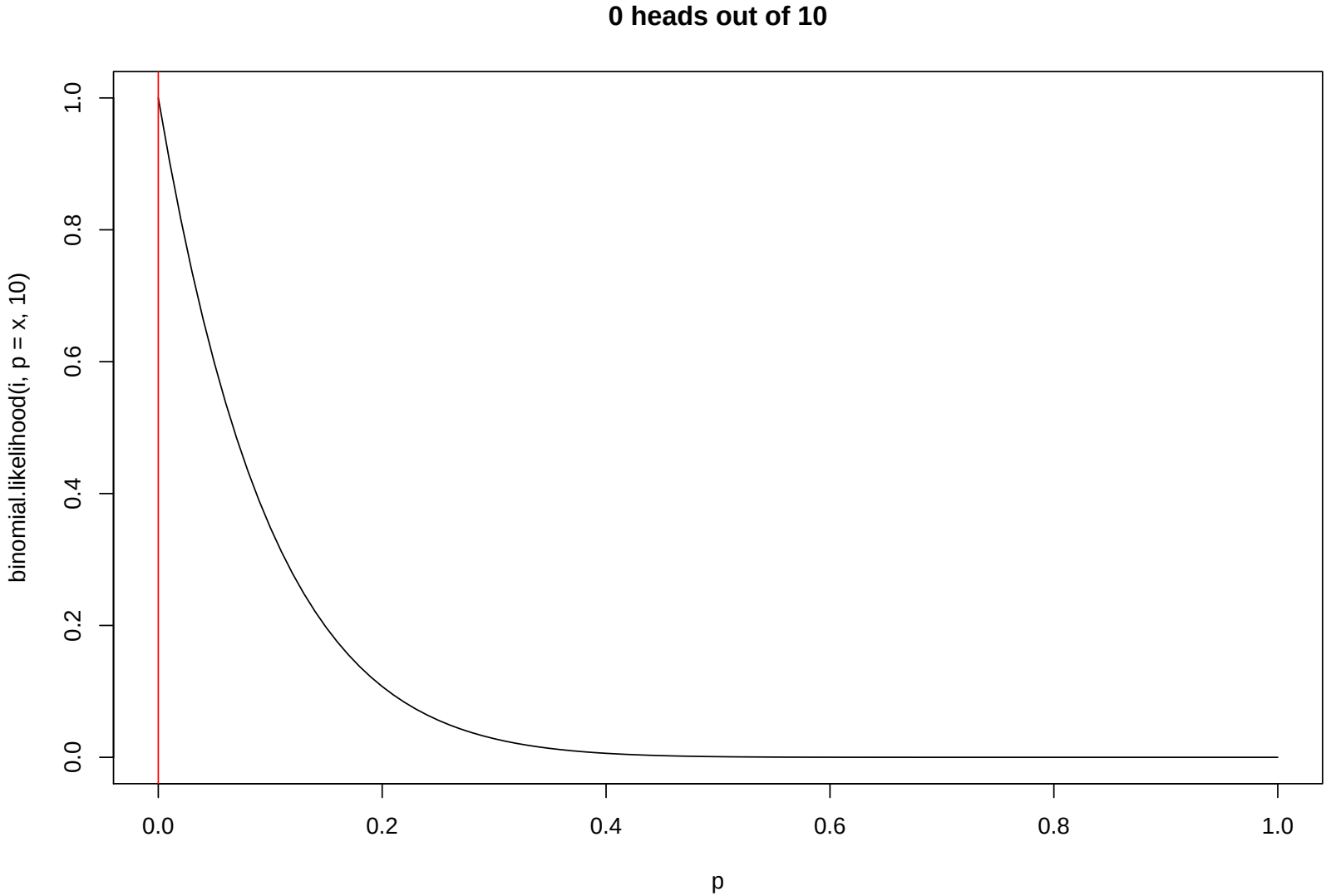
Notes on the likelihood

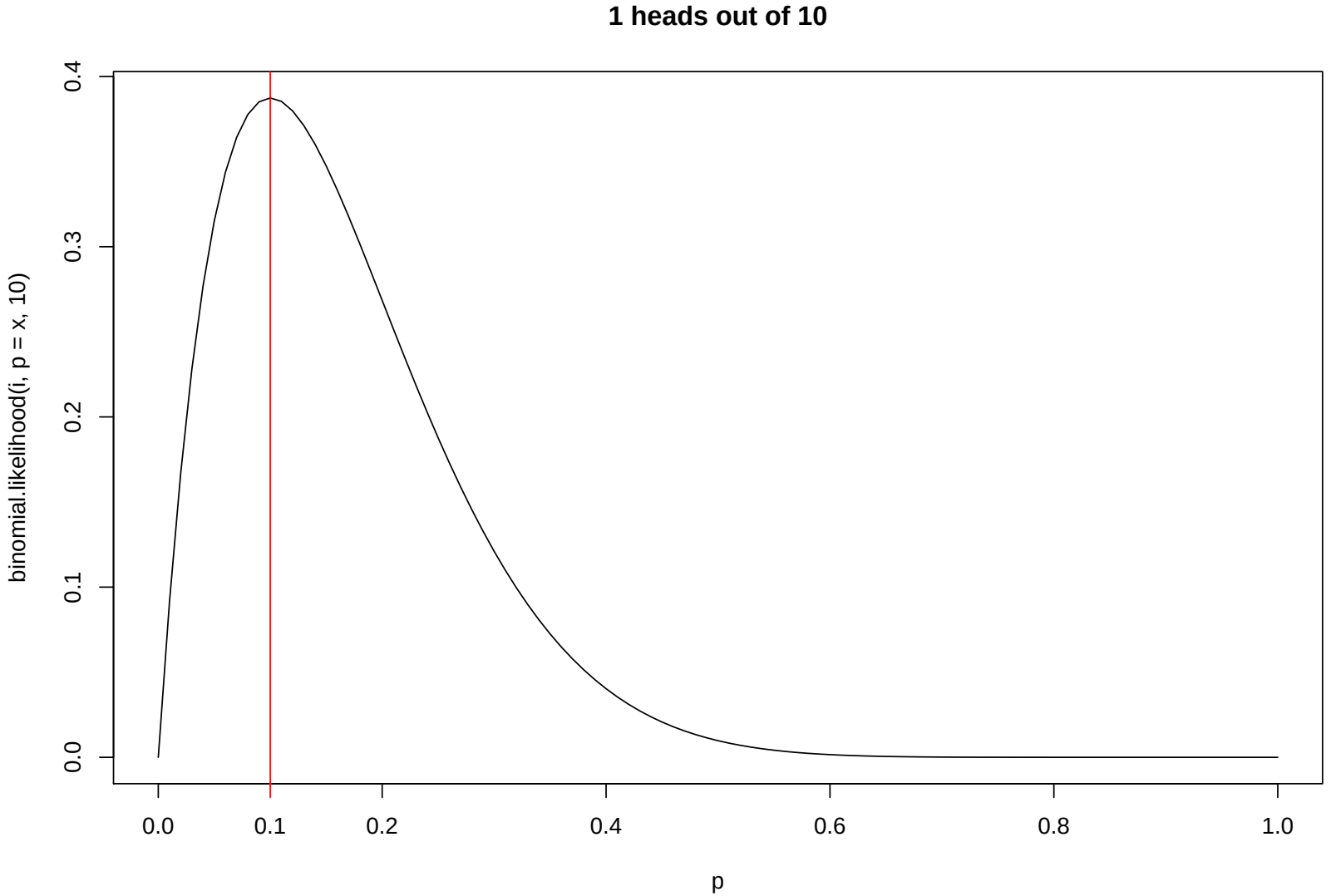
Taking the first example from the previous slide...
Result $= T, T, T, H, H, H, H, H, H, H$

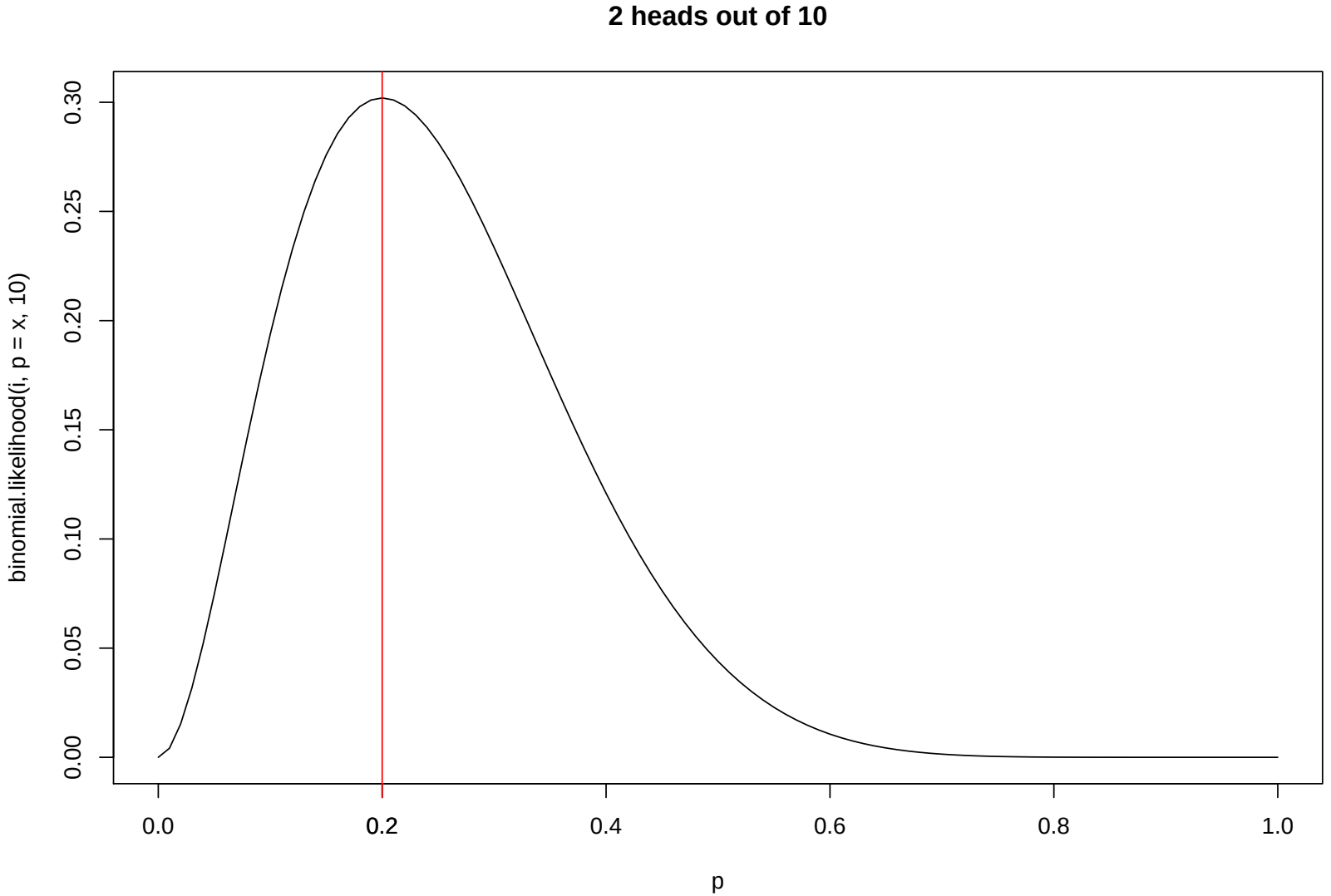
$$
\begin{aligned}
p(y \mid \theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \mathcal{L}(\theta) \\
&= \theta^{y_1}(1-\theta)^{1-y_1} \times \cdots \times \theta^{y_{10}}(1-\theta)^{1-y_{10}} \\
&= \theta^{0}(1-\theta)^{1} \times \theta^{0}(1-\theta)^{1} \times \theta^{0}(1-\theta)^{1} \times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \\
&\times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \times \theta^{1}(1-\theta)^{0} \\
&= \theta^{7}(1-\theta)^{3}
\end{aligned}
$$

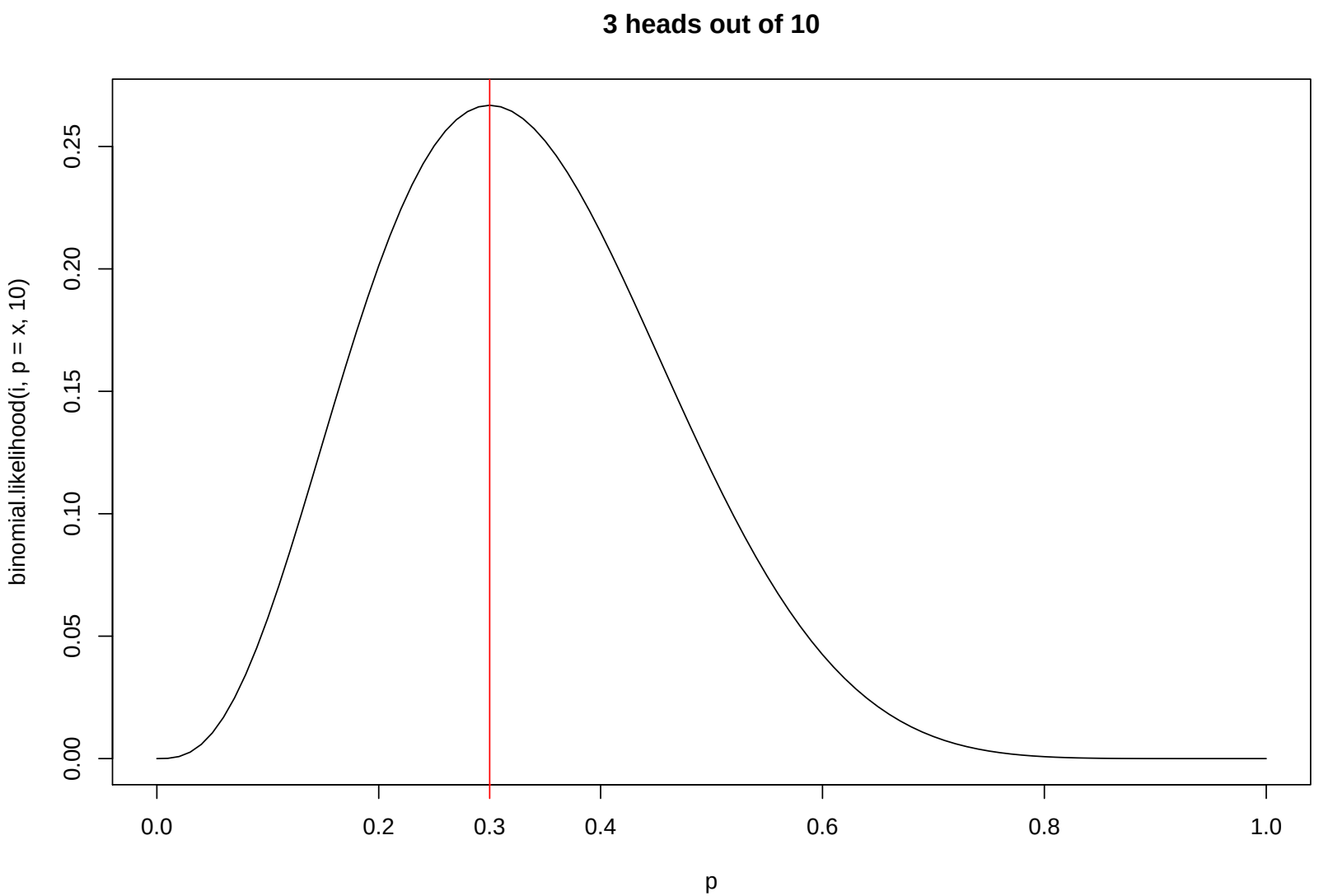
Notes on the likelihood

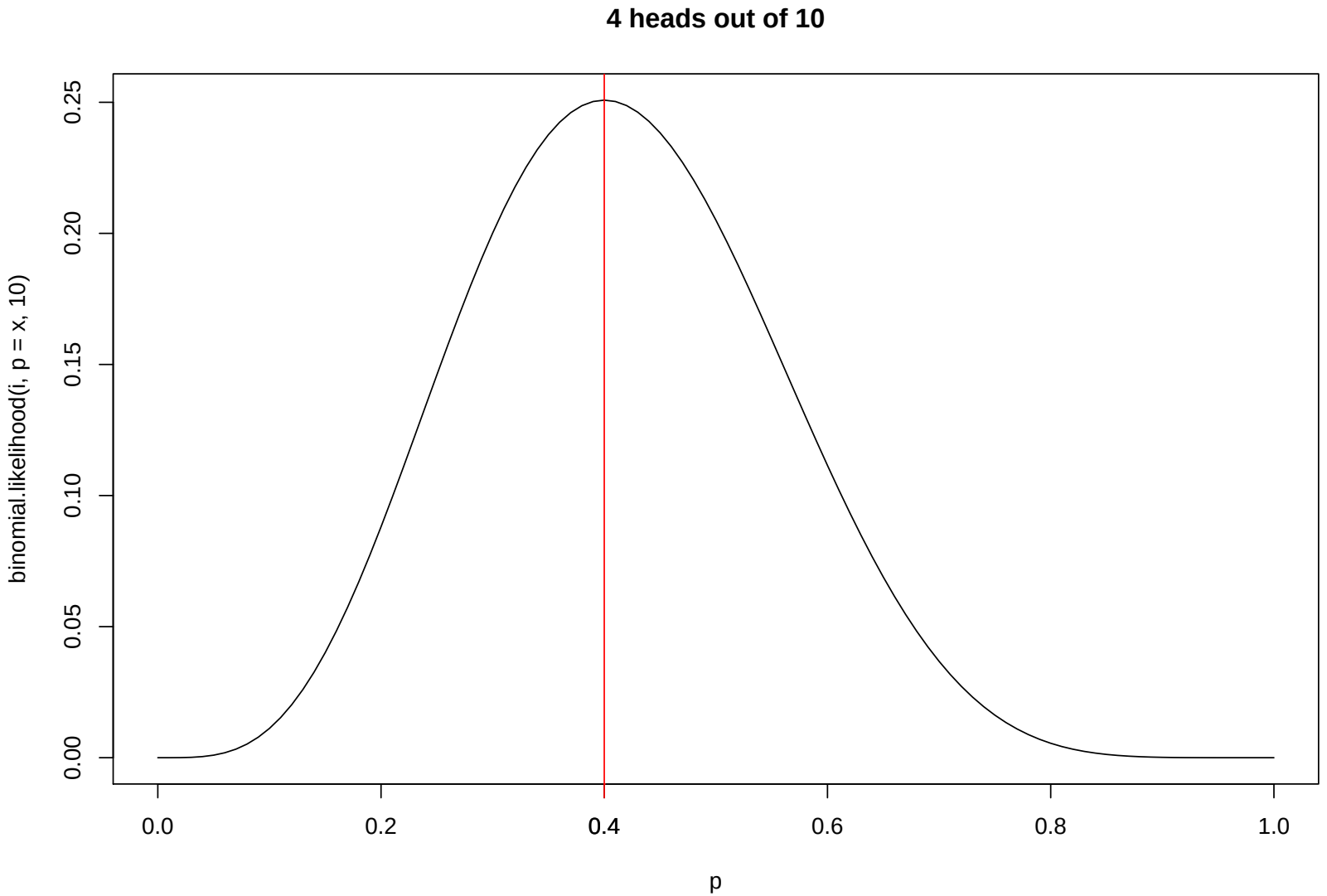
- The MLE produced from the data: $TTTHHHHHHH$ is exactly the same as the one produced by: $HHTHHHTHTH$ (that is 3 tails and 7 heads but in a different order).
- Hence it is sufficient to summarize our data in terms of the **total number of heads and tails** since the order in which they occur provide no additional information on θ.
- While the ordering of these flips are inconsequential, the total number of heads/tails **do** have an affect on the likelihood.
- The next 11 slides demonstrates how the likelihood changes as the number of observed heads (out of 10 flips) changes the likelihood function.

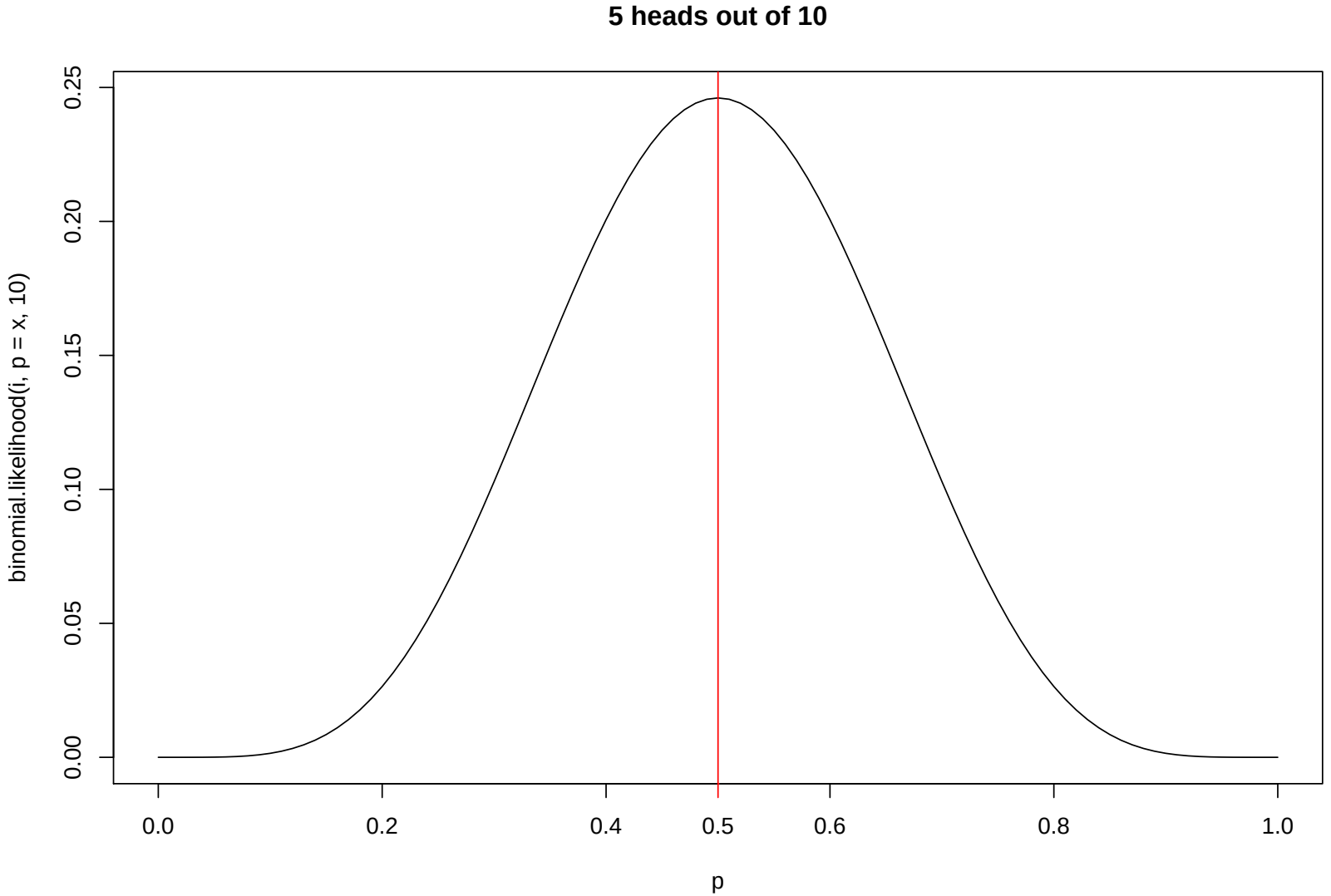

0 heads out of 10
binomial.likelihood(i, p = x, 10)
0.0
0.2
0.4
0.6
0.8
1.0
p
0.0
0.2
0.4
0.6
0.8
1.0

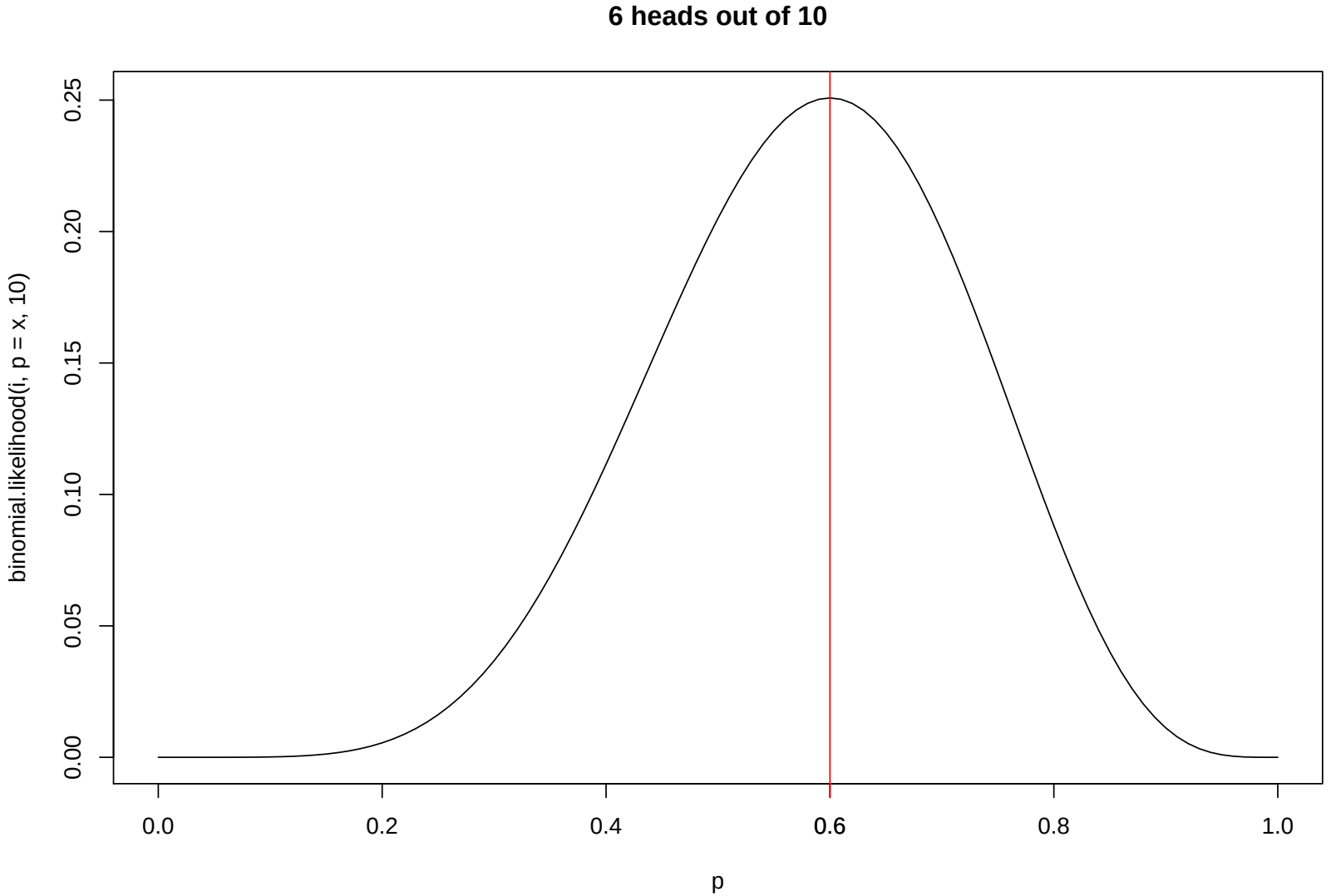

1 heads out of 10
binomial.likelihood(i, p = x, 10)
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.4
0.6
0.8
1.0
p

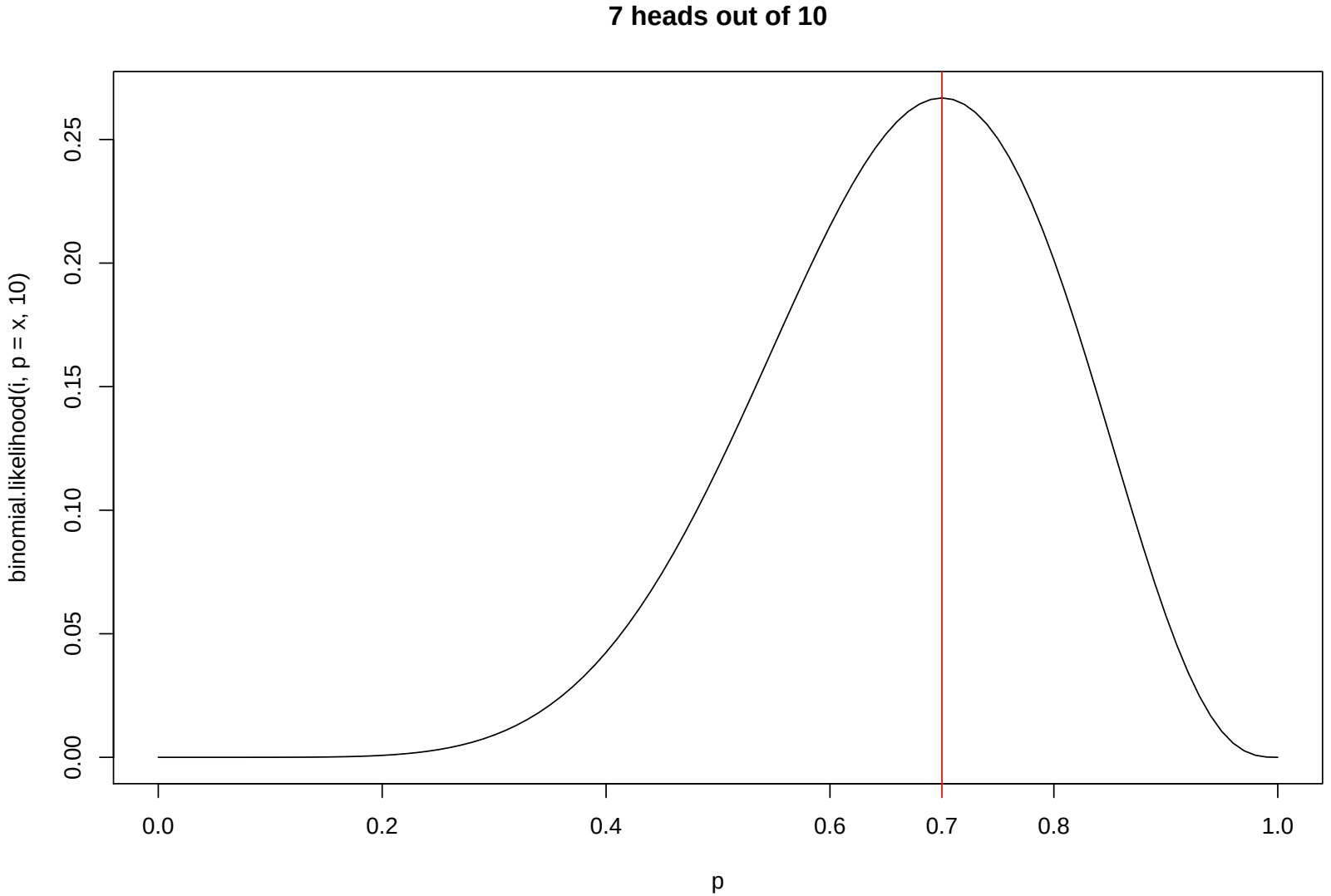

2 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0
p

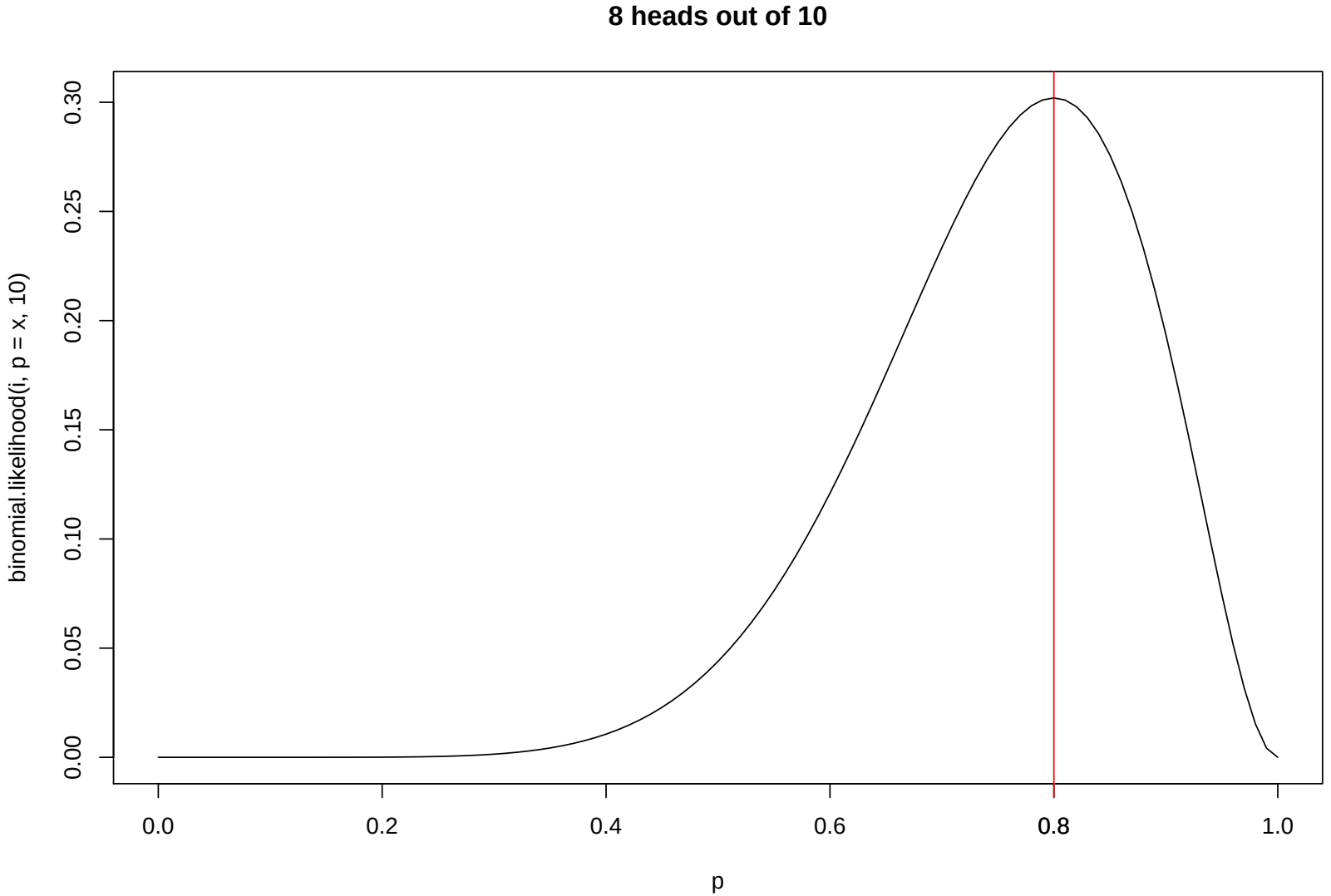

3 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.3
0.4
0.6
0.8
1.0
p

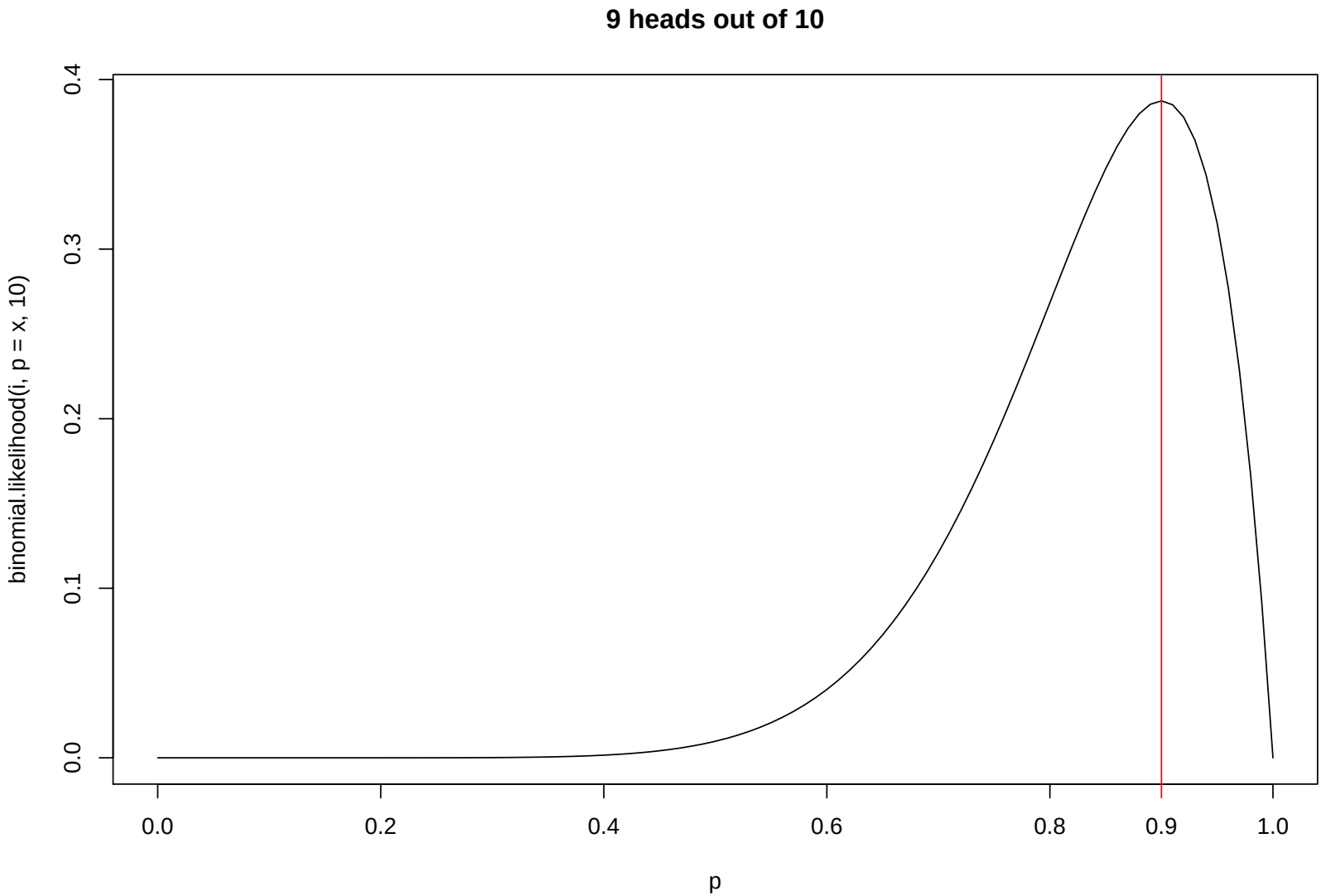

4 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
p

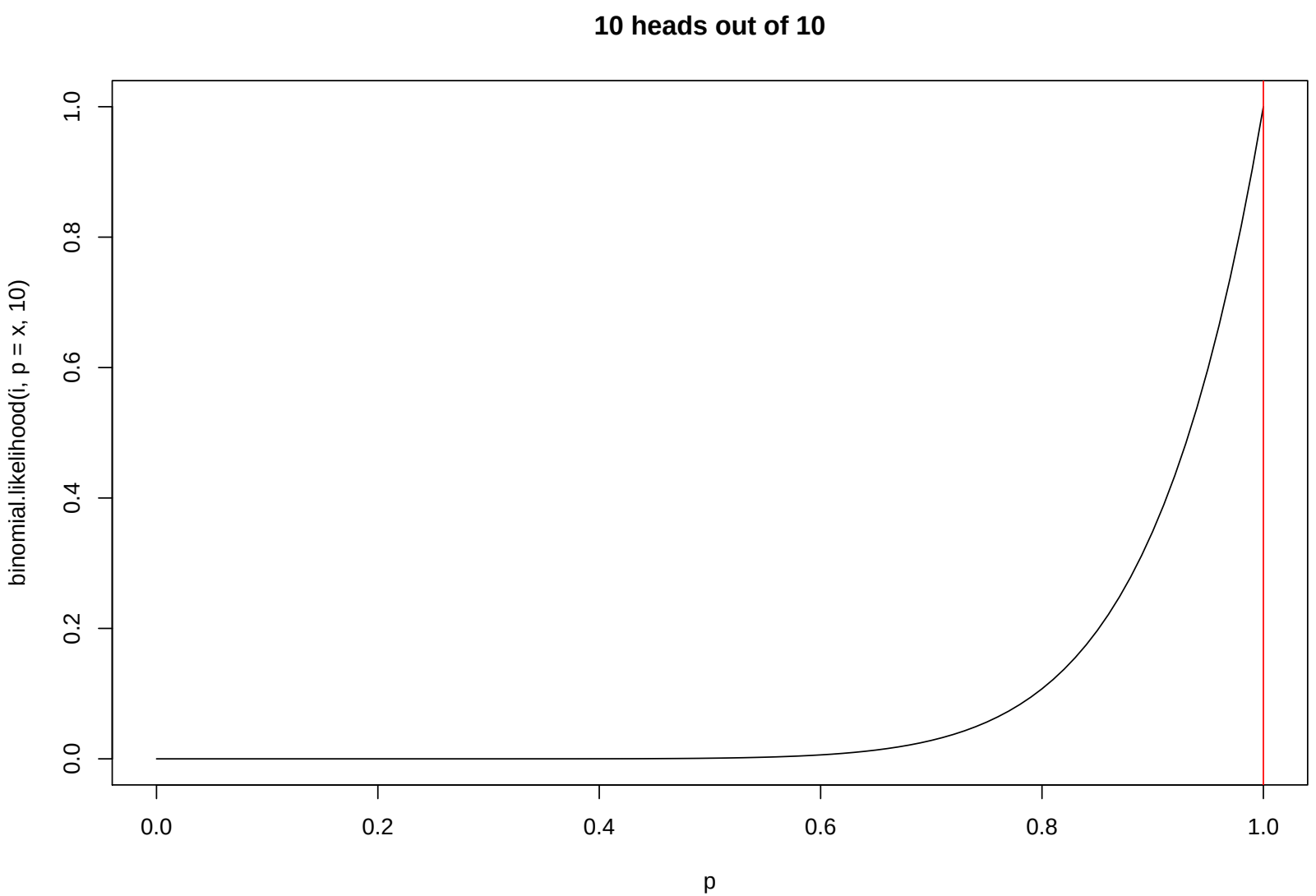

5 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.5
0.6
0.8
1.0
p


6 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
p


7 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.7
0.8
1.0
p


8 heads out of 10
binomial.likelihood(i, p = x, 10)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0
p


9 heads out of 10
binomial.likelihood(i, p = x, 10)
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
0.9
1.0
p


10 heads out of 10
binomial.likelihood(i, p = x, 10)
0.0
0.2
0.4
0.6
0.8
1.0
p
0.0
0.2
0.4
0.6
0.8
1.0

- While the above assumed $n = 10$ independent Bernoulli observations, we could have just as easily modeled this using $X \sim \text{Binomial}(n, \theta)$ having pmf given by:

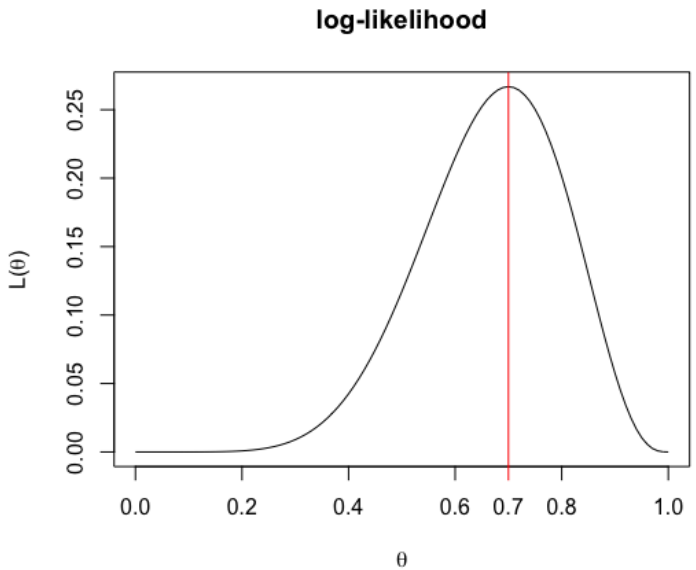
$$p(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^n \tag{3}$$

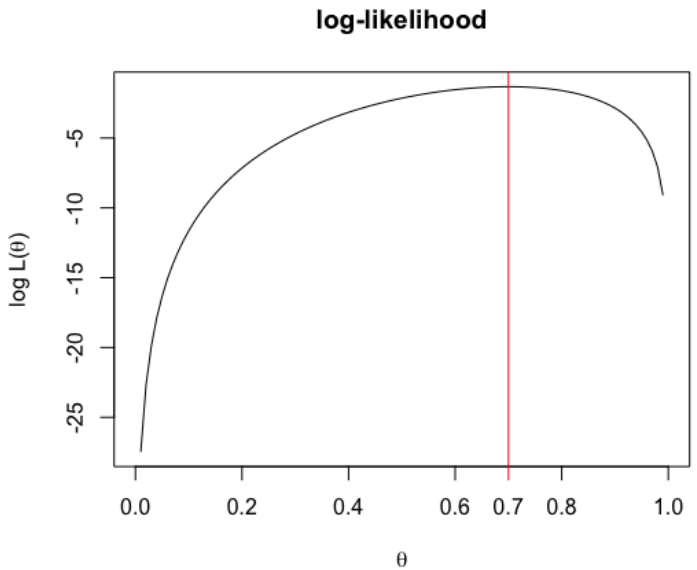
  for $x = 0, 1, \ldots, n$ and $\theta \in (0, 1)$

- Since we observed 7 heads in 10 trials the likelihood is:

$$\mathcal{L}(\theta) = \binom{10}{7} \theta^7 (1-\theta)^3 \tag{4}$$

The maximum of the likelihood function is still located at $\theta = 0.7$. The only thing that has changed is the scale on the y-axis.





- This exercise emphasizes an important point. The likelihood function need only be defined **up to a constant of proportionality**.
- That is we could scale the y-axis by any factor and $\mathcal{L}(\theta)$ would still yield the same inference for θ.
- From a practical standpoint this means we can remove any constants with respect to θ and still obtain the same MLE.

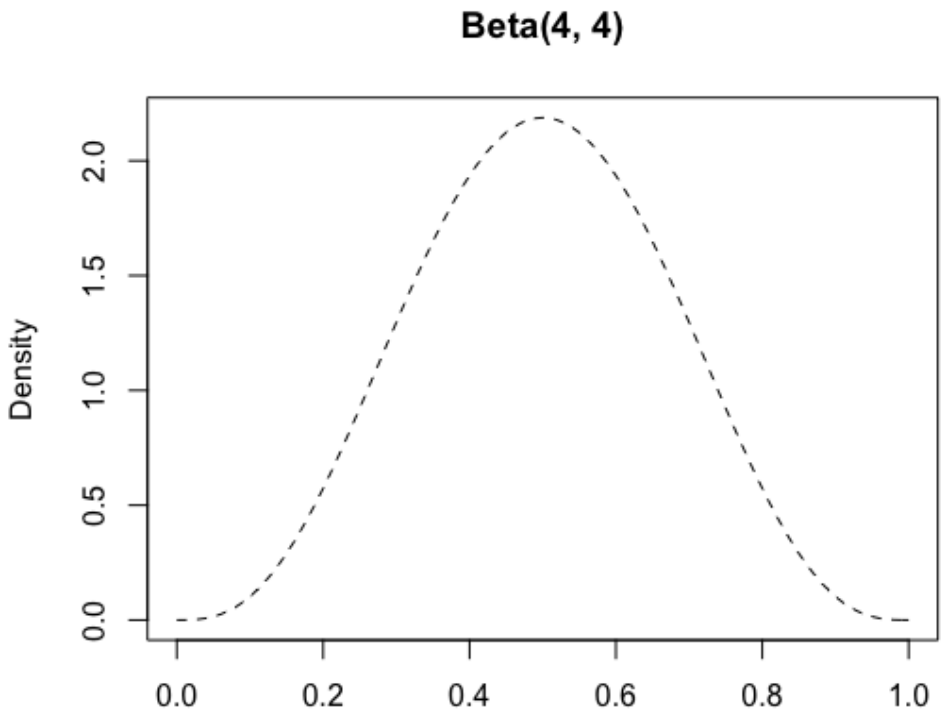
Definition (*proportionality* as defined in SL pg 52)

If a is proportional to b, then a and b only differ by a multiplicative constant.

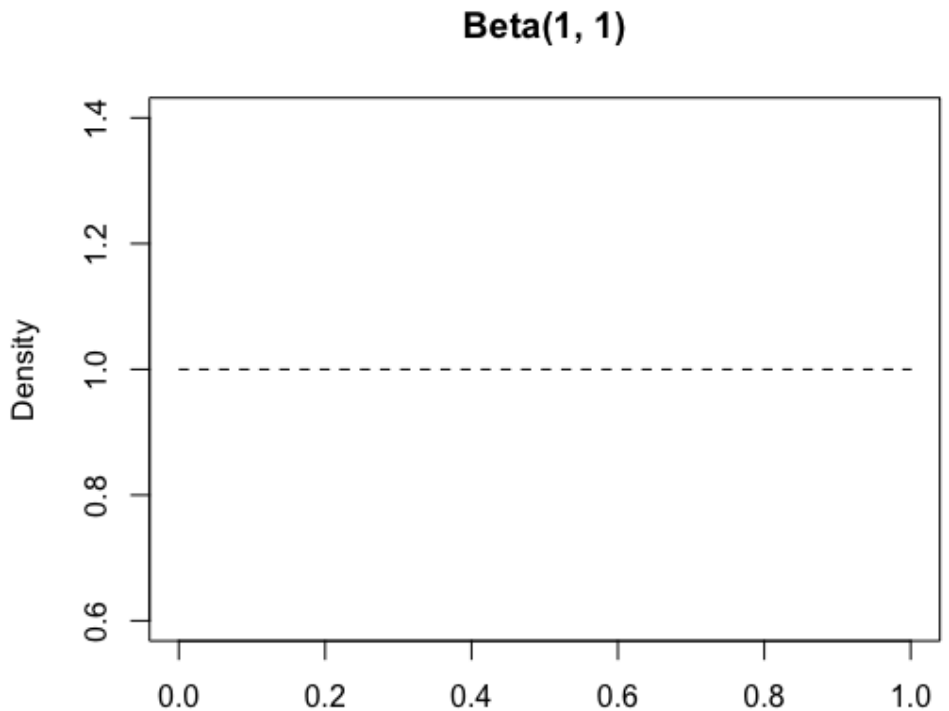
The Prior

- Before collecting any data, a Bayesian would specify what their belief about θ in the form or a *prior distribution*.
- There is great flexibility in doing this and specifying a distribution deemed "reasonable" is subjective.
- The subjective nature of this stage of Bayesian inference is a major point of contention.
- While there is no "right" way to define a prior, we will investigate how our choice at this stage can influence our analysis.

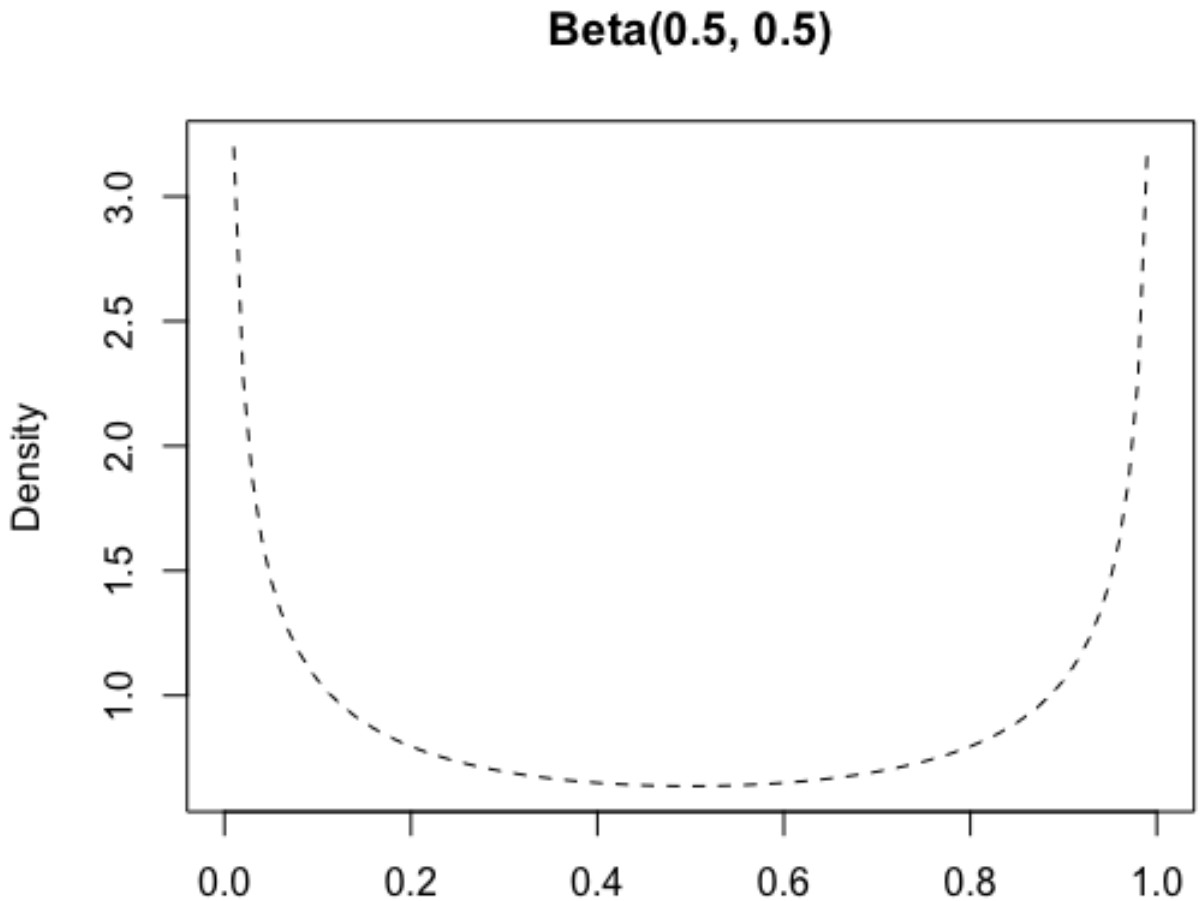
For instance, we might suppose the coin is most-likely fair but have some uncertainty about it. This could be summarized using the Beta distribution with parameters α and β set equal to 4.



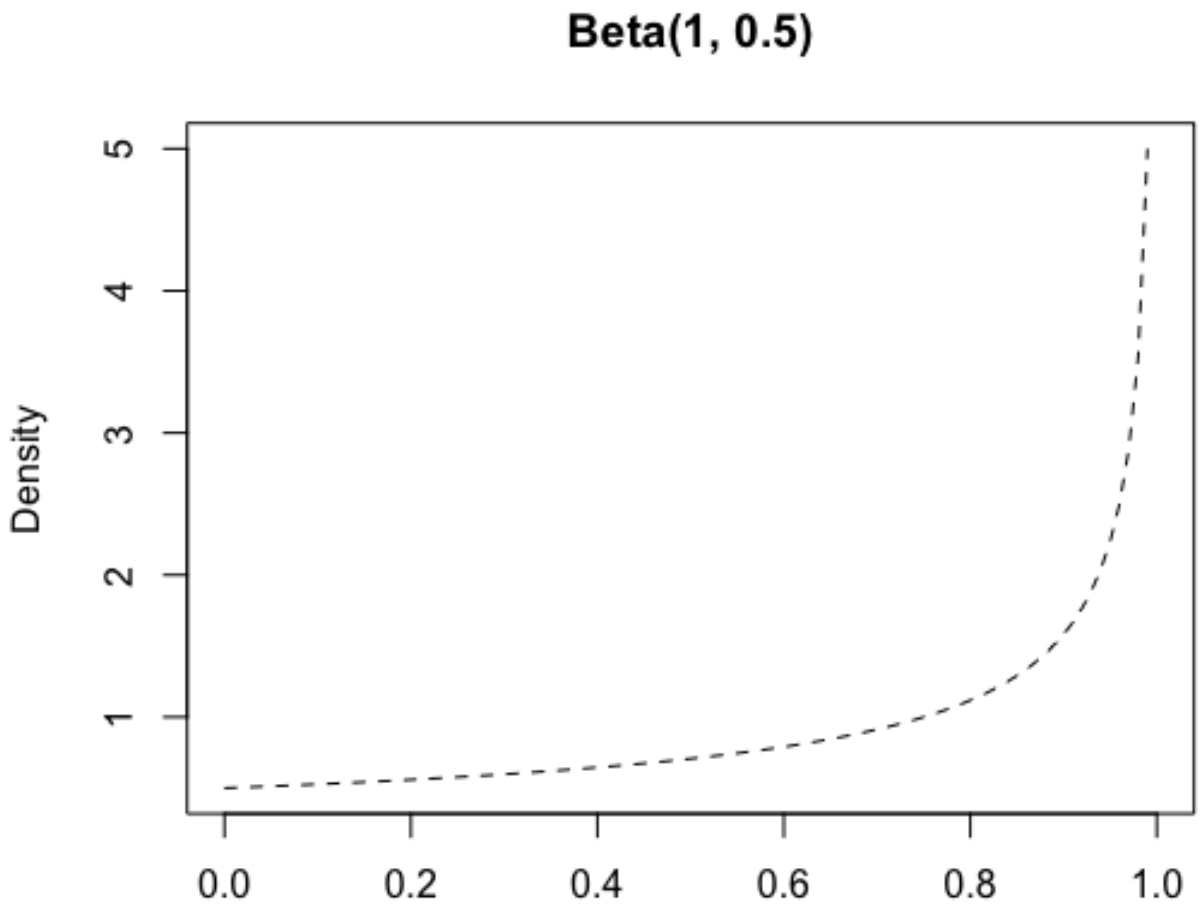
Alternatively, if we have little or no prior information, we might give all possible values of θ an equal probability. This could be achieved using a uniform distribution which is actually just a special case of the beta distribution.



Maybe we're highly suspicious that the coin is biased towards either tails or heads. In that case we might use $\theta \sim \text{Beta}(\alpha = 0.5, \beta = 0.5)$



Here we plot $\theta \sim \text{Beta}(\alpha = 1, \beta = 0.5)$. What might this prior distribution represent in words?



The effect of a Prior

- As mentioned previously, the prior and the likelihood have a tug-of-war match before they arrive at a posterior.
- To continue with this analogy, we can define prior to be weak (uniformative) or strong (informative).
- Specifying a strong prior is akin to assigning very strong weight to our prior belief about θ.
- As in life, if we are extremely confident in a certain belief, it takes a significant amount of contrary information to change our mind.
- Similarly, to compete with a strong prior, a significant amount of evidence would need to be contained in likelihood to influence our posterior away from that belief.

Both of these priors reflect the belief that a coin is fair but the left prior is much weaker than the right.

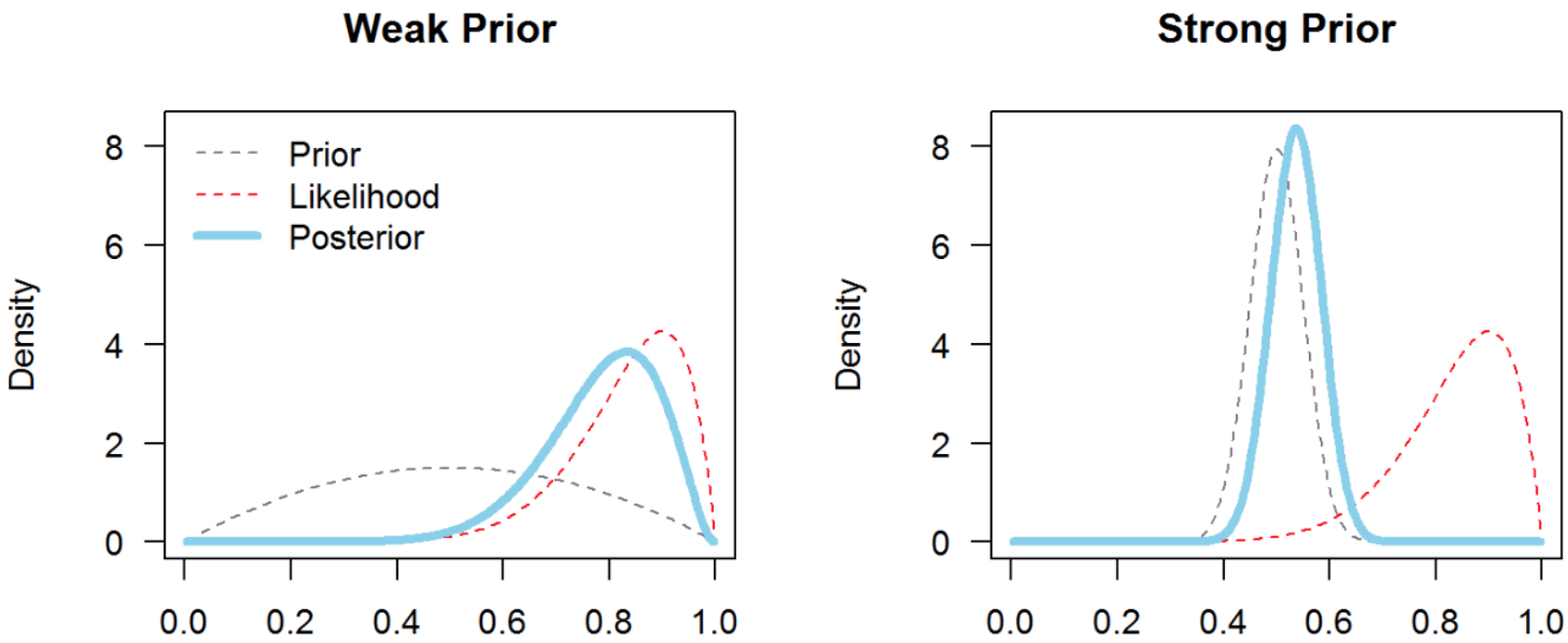


Image source: here

The likelihood is the same in both examples (the result of observing 9/10 heads)

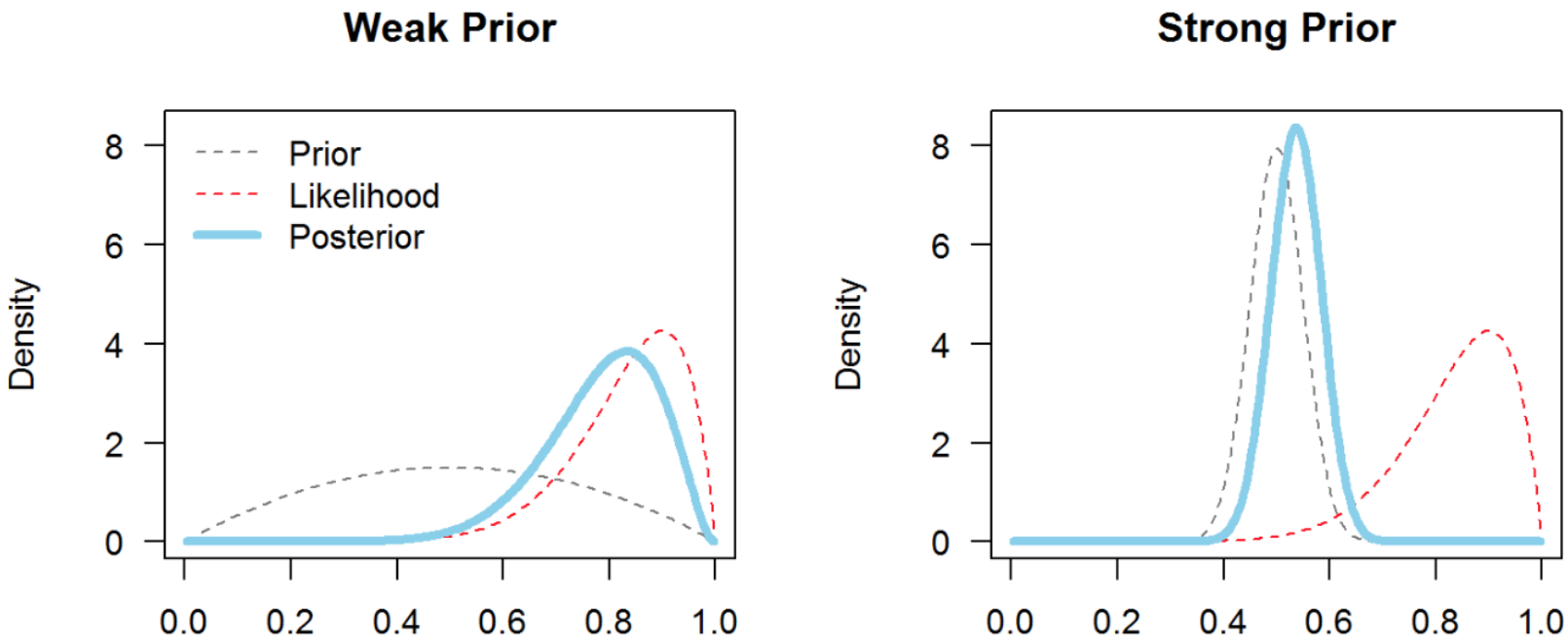


Image source: here

Despite the fact that there is less than 1% chance of seeing 9/10 heads if the coin was fair, our posterior distribution still sits very close to our strong prior, that is our mind wasn't changed much in light of the data.

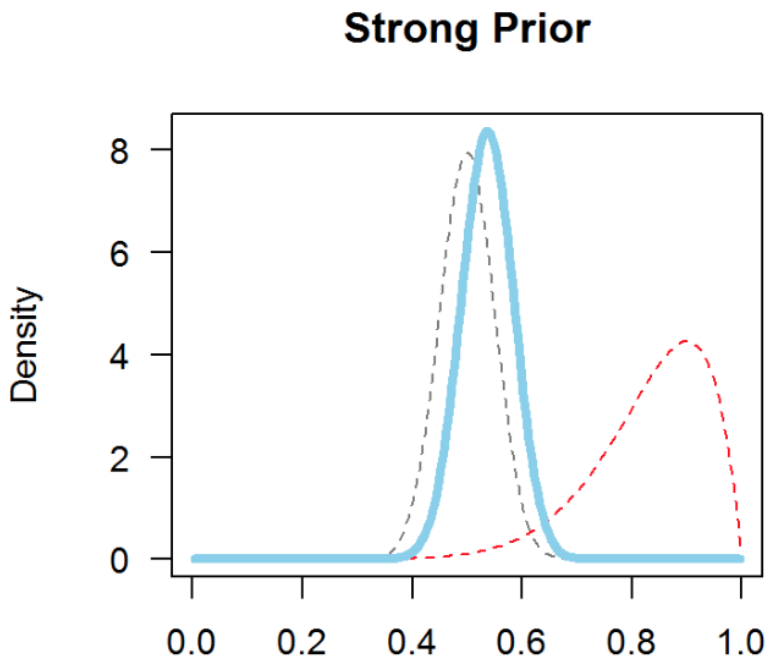


Image source: here

By specifying a weak prior like the one given below, the data have greater influence[5] on the resulting posterior and are not dominated by our prior beliefs.

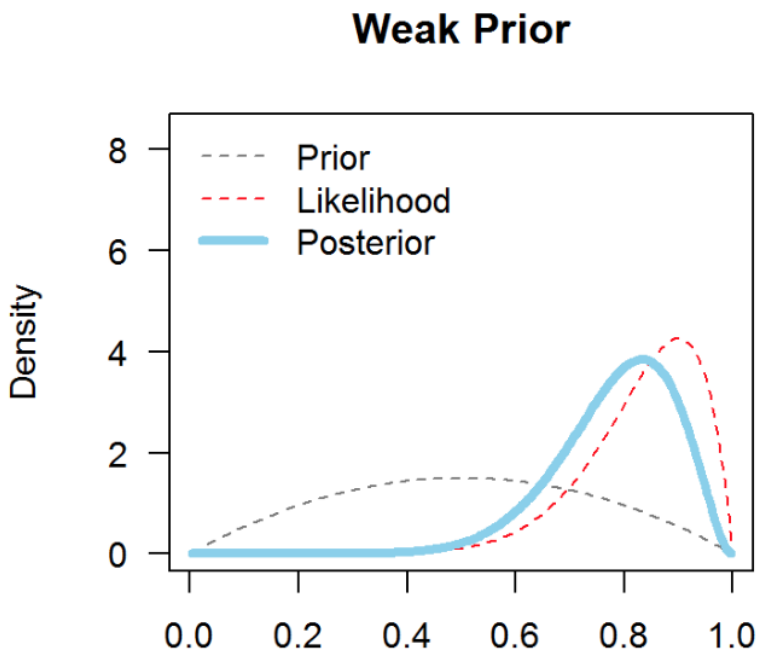


Image source: here

[5] i.e. more data-driven

- The takeaway is that if you are concerned your prior is wrong and is influencing your inferences, we'll need to collect more data.
- The effect of the prior can (and should) be investigated using robustness checks, where one plots the posterior distribution based on a range of (plausible) prior values.
- This process can be iterative, namely, once we have obtained a posterior, we can use it as our prior for the next study.
- We will be expanding on these keys ideas throughout this module and practical solutions for specifying beta parameters will be provided next lecture.

Posterior

- Rather than a single MLE point estimate, the Bayesian framework produces a summary a summary of all the possible values θ can take on along with their relative plausibility in form of a *posterior distribution*.

- The posterior distribution will represent all of "up-to-date" information and uncertainty we have about a population parameter θ in a very comprehensive way.

- Once we have this distribution we can obtain estimates for parameters, probability intervals, and more . . . (more to say about this in future lectures).

Discrete

Considers three types of coins: Type A, B, and C. Each has a different probabilities of landing heads when tossed.

A coins are fair, with probability 0.5 of heads
B coins are bent and have probability 0.6 of heads
C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin, I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Source: Jeremy Orloff, and Jonathan Bloom. 18.05 Introduction to Probability and Statistics. Spring 2014. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu.

Construct the Tree

- While we have done this type of Bayes theorem question before using trees, let's look at it through a slightly different (Bayesian) lense.

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} = \frac{\text{Bayes Numerator}}{\text{Bayes denominator}}$$

$$\text{posterior} \propto \text{Bayes Numerator}$$

Note: In the previous coin toss example $\theta \in (0, 1)$ was treated as a continuous random variable. Here $\theta = \{0.5, 0.6, 0.9\}$ is discrete, so we summarize the likelihood (and prior) in a table

Step 1: Prior

- Before seeing the result of the coin, we know the marginal distribution the coin type.
- Converting this to information related to θ, we can summarize this as a prior distribution $p(\theta)$ with the following PMF:

	θ		
	0.5	0.6	0.9
$p(\theta)$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

- Notice how the probabilities sum to 1 and accounts for all possible events.

Step 2: Likelihood

If we define a random variable $X \sim \text{Bernoulli}(\theta)$ then we can summarizing our data $x = 1$ in the following likelihood.

$$\mathcal{L}(\theta) = \prod_{i=1}^{1} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \begin{cases} p(x \mid \theta = 0.5) = 0.5^1(1-0.5)^{1-1} = 0.5 \\ p(x \mid \theta = 0.6) = 0.6^1(1-0.6)^{1-1} = 0.6 \\ p(x \mid \theta = 0.9) = 0.9^1(1-0.9)^{1-1} = 0.9 \end{cases}$$

Notice how these probabilities do NOT sum to 1.

Step 2: Likelihood

To put another way, our likelihood is the probability of the data given a certain hypothesis H_i.

$$\mathcal{L}(\theta) = \begin{cases} P(\text{Heads} \mid \text{Coin A}) = 0.5 & \text{if } H_1 : \text{Coin A} \\ P(\text{Heads} \mid \text{Coin B}) = 0.6 & \text{if } H_2 : \text{Coin B} \\ P(\text{Heads} \mid \text{Coin C}) = 0.9 & \text{if } H_3 : \text{Coin C} \end{cases}$$

In the frequentist paradigm, coin C in the most likely since it has the largest likelihood.

H_i	θ	prior	likelihood	Bayes num.	posterior*
A	0.5	$\frac{2}{5}$	0.5	$\frac{2}{5} \times 0.5 = 0.2$	0.2
B	0.6	$\frac{2}{5}$	0.6	$\frac{2}{5} \times 0.6 = 0.24$	0.24
C	0.9	$\frac{1}{5}$	0.9	$\frac{1}{5} \times 0.9 = 0.18$	0.18
Total:		1	2	0.62	0.62

*this is an *unnormalized posterior* because it is not yet a proper pmf that sums to 1.

H_i	θ	prior	likelihood	Bayes num.	posterior*
A	0.5	$\frac{2}{5}$	0.5	$\frac{2}{5} \times 0.5 = 0.2$	$c \times 0.2$
B	0.6	$\frac{2}{5}$	0.6	$\frac{2}{5} \times 0.6 = 0.24$	$c \times 0.24$
C	0.9	$\frac{1}{5}$	0.9	$\frac{1}{5} \times 0.9 = 0.18$	$c \times 0.18$
Total:		1	2	0.62	

We can multiply the unnormalized posterior by any constant (w.r.t θ) and get the same "answer" up to a constant of proportionality.

Posterior

- To summarize the posterior in a meaningful way (eg. finding mean values, sd, etc), we would like to scale the last column such that it can can be a *proper*[6] pmf.

- We need to find the constant c such that the last column sums to 1. This special constant is known as the *normalizing constant*

$$1 = c \times 0.2 + c \times 0.24 + c \times 0.18$$

$$\implies c = \frac{1}{0.2 + 0.24 + 0.18} = \frac{1}{0.62}$$

[6] A proper pdf/pmf is one that integrates/sums to the unity.

H_i	θ	prior	likelihood	Bayes num.	posterior
A	0.5	$\frac{2}{5}$	0.5	$\frac{2}{5} \times 0.5 = 0.2$	$\frac{1}{0.62} \times 0.2 = 0.3226$
B	0.6	$\frac{2}{5}$	0.6	$\frac{2}{5} \times 0.6 = 0.24$	$\frac{1}{0.62} \times 0.24 = 0.3871$
C	0.9	$\frac{1}{5}$	0.9	$\frac{1}{5} \times 0.9 = 0.18$	$\frac{1}{0.62} \times 0.18 = 0.2903$
Total:		1			1

Now the posterior is a proper pmf!

- Returning to our original question.[7]
  - $P(\text{type A} \mid \textit{Head}) = 0.3226$
  - $P(\text{type B} \mid \textit{Head}) = 0.3871$
  - $P(\text{type C} \mid \textit{Head}) = 0.2903$

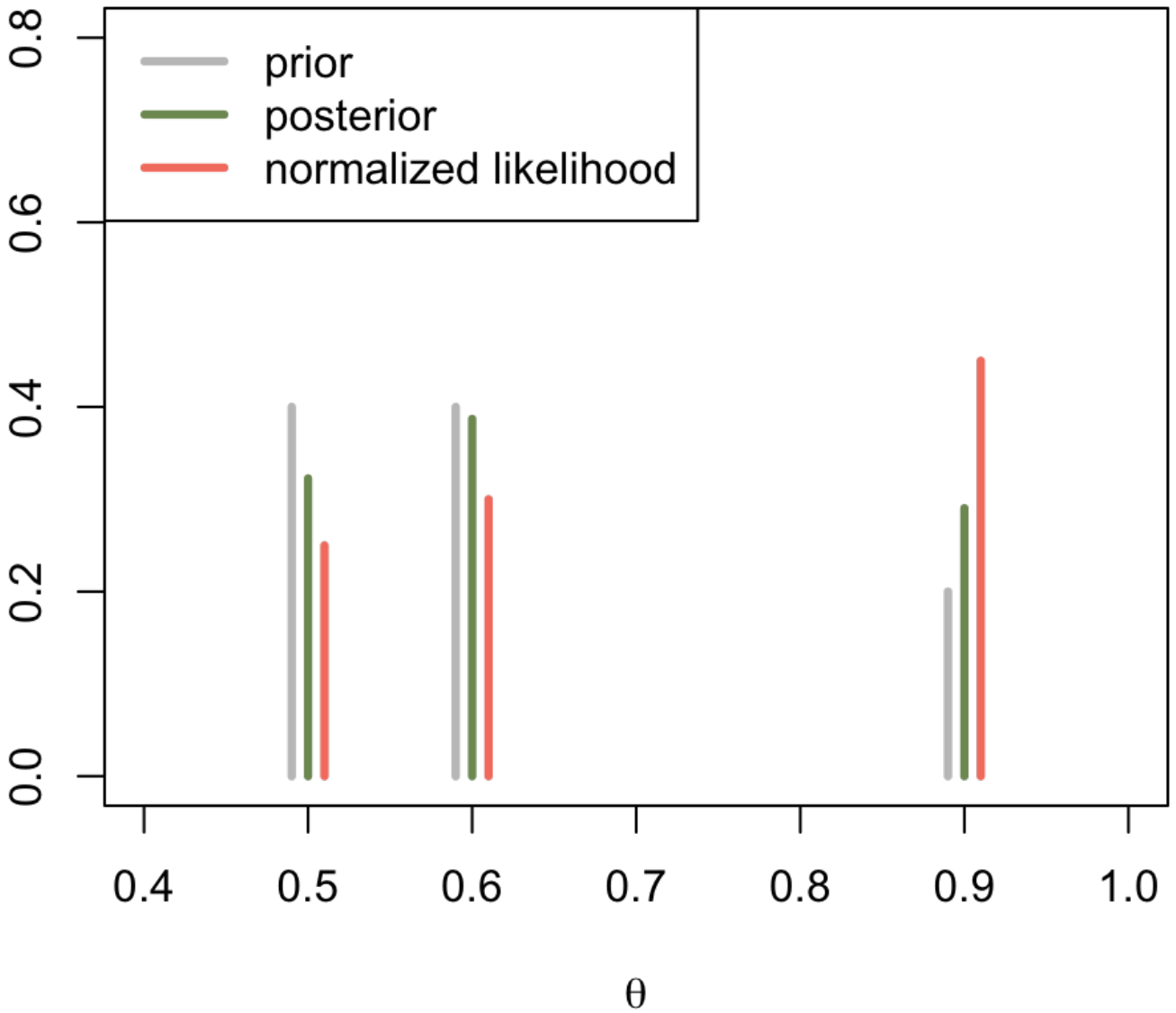
- To put it more into the Bayesian framework, the final column give us the **posterior probabilities** that θ is equal to 0.5, 0.6, or 0.9 provided we have observed one head. That is, our posterior distribution $p(\theta \mid y)$ is given by pmf:

	θ		
	0.5	0.6	0.9
$p(\theta \mid y)$	0.3226	0.3871	0.2903

[7] *N.B.* if we hadn't multiplied the unnormalized posterior by the normalizing constant, the last column of our table could **not** be interpreted as a probability

Expressing the PMF

- In this example we have expressed our pmf in table form.
- To visualize the problem, we can plot the pmf with height of the vertical lines indicating the probability for θ to take on the value corresponding on a x-axis.
- To put the likelihood on the same footing as the prior and the posterior I have scaled the values such that the total height adds to 1
- While we will eventually move to smooth curves in the continuous case, notice that the tug-of-war match between the prior and the likelihood can still be viewed in this visualization.


prior
posterior
normalized likelihood
0.0
0.2
0.4
0.6
0.8
0.4
0.5
0.6
0.7
0.8
0.9
1.0
θ

Comments on Bayes Denominator

- In this discrete example, finding the normalizing constant was easy through some basic algebra.
- Equivalent to finding the normalizing constant, we could have found $p(x)$ directly (ie. Bayes denominator)

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

- This of course requires the calculation of $p(x)$, the marginal probability of the coin landing heads....

This calculation is similar to the $P(+)$ calculation we did in the example in our last lecture:

$$
\begin{aligned}
P(Heads) &= \sum_{i=1}^{3} P(Heads \mid H_i) P(H_i) \\
&= P(Heads \mid H_1) P(H_1) + P(Heads \mid H_2) P(H_2) + P(Heads \mid H_3) P(H_3) \\
&= 0.5 * \frac{2}{5} + 0.6 * \frac{2}{5} + 0.9 * \frac{1}{5} \\
&= 0.62
\end{aligned}
$$

Even with this simple example, it's a bit of a cumbersome task.

Data: Head ($x = 1$)

H_i		$p(\theta)$	$p(x \mid \theta)$	$p(x \mid \theta)p(\theta)$	(unnormalized) posterior
A	0.5	$\frac{2}{5}$	0.5	$\frac{2}{5} \times 0.5 = 0.2$	0.2
B	0.6	$\frac{2}{5}$	0.6	$\frac{2}{5} \times 0.6 = 0.24$	0.24
C	0.9	$\frac{1}{5}$	0.9	$\frac{1}{5} \times 0.9 = 0.18$	0.18
Total:		1	2	$p(x)$ =0.62	0.62

Note that the calculation on the previous slide can be obtained by summing up the Bayes numerator column.

Data: Head ($x = 1$)

H_i		$p(\theta)$	$p(x \mid \theta)$	$p(x \mid \theta)p(\theta)$	$p(x \mid \theta)p(\theta)/p(x)$
A	0.5	$\frac{2}{5}$	0.5	$\frac{2}{5} \times 0.5 = 0.2$	$\frac{1}{0.62} \times 0.2 = 0.3226$
B	0.6	$\frac{2}{5}$	0.6	$\frac{2}{5} \times 0.6 = 0.24$	$\frac{1}{0.62} \times 0.24 = 0.3871$
C	0.9	$\frac{1}{5}$	0.9	$\frac{1}{5} \times 0.9 = 0.18$	$\frac{1}{0.62} \times 0.18 = 0.2903$
Total:		1	2	$p(x) = 0.62$	1

Once we divided the unnormalized posterior by the normalizing constant $= p(x)$ the posterior becomes proper!

- More generally, the marginal distribution $p(x)$ is given by:

$$= \sum_y p(x, y) = \sum_y p(x \mid y)p(y) \qquad \text{for discrete RVs}$$

$$= \int p(x, y)dy = \int p(x \mid y)p(y)dy \qquad \text{for continuous RVs}$$

- Often this marginal is hard if not impossible to compute so we'd like to avoid calculating it if we can.

where RV stands for random variable

Follow up questions: Coin example

- Suppose we flip the same coin once more and it lands heads.
- As discussed earlier, we will use the posterior of the first study as the prior in our second study.
- We will denote the first flip by x_1, and the second flip by x_2.
- θ, the probability of landing heads, is our parameter of interest.
- As you will see, it is not necessary to to find the normalizing constant for the posterior at Study 1 to be used as a prior in Study 2.[8]

[8] this goes back to the fact that the posterior need on be specified up to a constant of proportionality

Results from first study:

$\mathcal{D}$ata $x_1 = 1$ and using prior specified on slide 51

(H_i)	prior $p(\theta)$	likelihood $p(x_1 = 1 \mid \theta)$	Bayes numerator $p(x_1 = 1 \mid \theta)p(\theta)$	*posterior* $p(\theta \mid x_1)$
Type A	0.4	0.5	0.5*0.4 = 0.20	*0.3226*
Type B	0.4	0.6	0.6*0.4 = 0.24	*0.3871*
Type C	0.2	0.9	0.9*0.2 = 0.18	*0.2903*
Total	1	–	*0.62*	*1*

Second study:

$\mathcal{D}$ata $x_2 = 1$ and using a prior equal to the the (proper) posterior obtained from the first experiment (i.e. 4th column from slide 67)

	prior* $p(\theta)$	likelihood $p(x_2 = 1 \mid \theta)$	Bayes numerator $p(x_2 = 1 \mid \theta)p(\theta)$	posterior $p(\theta \mid x_2)$
	0.3226	0.5	0.1613	$\frac{0.1613}{0.6549} = 0.2463$
	0.3871	0.6	0.2323	$\frac{0.2323}{0.6549} = 0.3547$
	0.2903	0.9	0.2613	$\frac{0.2613}{0.6549} = 0.3990$
$\sum$	1	2	0.6549	1

Second study:

$\mathcal{D}$ata $x_2 = 1$ and using a prior equal to the the unnormalized posterior obtained from the first experiment (i.e. 3rd column from slide 67)

	prior $p(\theta)$	likelihood $p(x_2 = 1 \mid \theta)$	Bayes numerator $p(x_2 = 1 \mid \theta)p(\theta)$	posterior $p(\theta \mid x_2)$
	0.20	0.5	0.5*0.20 = 0.100	$\frac{0.100}{0.406} = 0.2463$
	0.24	0.6	0.6*0.24 = 0.144	$\frac{0.144}{0.406} = 0.3547$
	0.18	0.9	0.9*0.18 = 0.162	$\frac{0.162}{0.406} = 0.3990$
$\sum$	1	–	0.406	1

- In the last example we did two subsequent studies.
- It is important to know that if we had inputted the data all at once, we again would have arrived at the same answer.
- Letting $y =$ total number of Heads, and adopting the Binomial likelihood (pmf given in Equation (3)) we get ...

$$\mathcal{L}(\theta) = \binom{2}{2}\theta^2(1-\theta)^{2-2}$$

$$= \begin{cases} p(y \mid \theta = 0.5) = \binom{2}{2}0.5^2 = 0.25 \\ p(y \mid \theta = 0.6) = \binom{2}{2}0.6^2 = 0.36 \\ p(y \mid \theta = 0.9) = \binom{2}{2}0.9^2 = 0.81 \end{cases}$$

Data: HH, $y = 2$

	prior $p(\theta)$	likelihood $p(y \mid \theta)$	Bayes numerator $p(y \mid \theta)p(\theta)$	posterior $p(\theta \mid y)$
	0.40	0.25	0.25*0.40 = 0.100	$\frac{0.100}{0.406} = 0.2463$
	0.40	0.36	0.36*0.40 = 0.144	$\frac{0.144}{0.406} = 0.3547$
	0.20	0.81	0.81*0.2 = 0.162	$\frac{0.162}{0.406} = 0.3990$
$\sum$	1	–	0.406	1

Note that this is the same result as slide 69

Summary

What did this example demonstrate?

proportionality multiplying the prior or likelihood by any constant (with respect to θ) will have no affect on the (proper) posterior.

exchangeability the likelihood is the same no matter which order the variables are observed.

Bayesian updating after updating the prior to the posterior, we can take more data and update again!

$p(x)$ we'd like to bypass finding this (it's usually hard!)

From Discrete to Continuous

- This example was built off the assumption that a coin was being drawn from a drawer containing 5 coins: 2 A-type coins, 2 B-type coins, and 1 C-type coin.

- **Scenario 1:** Now what happens if we have no idea how many coins of each coin I have?

- **Scenario 2:** Worse yet, what happens if I don't know how many coin types I have?

The answer is to simply consider a more flexible distribution for θ.

Scenario 1 we could assume, for example, a discrete uniform:

	θ		
	0.5	0.6	0.9
$p(\theta)$	1/3	1/3	1/3

Scenario 2 we could assume, for example, a continuous uniform on interval $[0, 1]$, i.e. $\theta \sim \mathcal{U}(0, 1)$. Another way of saying this is that we can assume that all possible values for θ, the probability of the coin landing heads, is equally likely.

Looking ahead

- In this discrete case, we used a simple algebra "trick" to avoid explicitly calculating $p(x)$
- In the upcoming lecture we'll be looking at more complicated continuous examples.
- Therein, we will use a different "trick" for avoid the calculation of $p(x)$ which rely on us recognizing the *functional form* of the unnormalized posterior.
- We will eventually run out of "tricks" and we will need to rely on numeric methods to *approximating* the posterior.