Statistics 2

Introduction to Bayesian Statistics

Casper Albers & Jorge Tendeiro

Lecture 13, 2019 – 2020

university of groningen

Overview

Recall from Statistics 1

Bayes' rule

- Prior
- Likelihood
- Marginal likelihood
- Posterior

About the posterior

- Effect of the prior
- Effect of the data I
- Effect of the data II
- Cromwell's rule
- Reallocation of credibility
- Summary of results
- Computation

Literature for this lecture

Read:

Kruschke, J. K. & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155-177. doi:10.3758/s13423-017-1272-1

Recall from Statistics 1

In Statistics 1A and 1B you learned important basics:

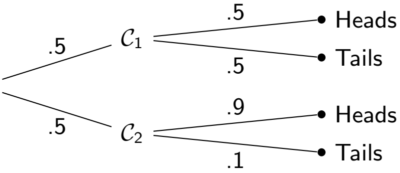
- Stats 1A: Conditional probabilities.
- Stats 1B: Caveats of NHST.
- Stats 1B: Introduction to Bayesian statistics.

Do review that material!

Here we *revisit* the main ideas and expand on them.

Conditional probability

Consider two coins, $\mathcal{C}_1$ and $\mathcal{C}_2$, indistinguishable from each other. $\mathcal{C}_1$ is fair but $\mathcal{C}_2$ is not (heads come 9 out of 10 times on average). You randomly choose one coin and throw it; it turns up heads. What is the probability that it is $\mathcal{C}_1$?



	Heads	Tails	
$\mathcal{C}_1$	.25 $(.5 \times .5)$	.25 $(.5 \times .5)$	.50
$\mathcal{C}_2$	.45 $(.5 \times .9)$	.05 $(.5 \times .1)$	.50
	.70	.30	1

$$
\begin{aligned}
p(\mathcal{C}_1|\text{Heads}) &= \frac{p(\mathcal{C}_1 \text{ and Heads})}{p(\text{Heads})} \\
&= \frac{p(\mathcal{C}_1)p(\text{Heads}|\mathcal{C}_1)}{p(\mathcal{C}_1)p(\text{Heads}|\mathcal{C}_1) + p(\mathcal{C}_2)p(\text{Heads}|\mathcal{C}_2)} \\
&= \frac{.5 \times .5}{.5 \times .5 + .5 \times .9} = \frac{.25}{.70} = .36.
\end{aligned}
$$

Inversed probability

'Direct' probabilities are easy to get, e.g.:

$$p(\text{Heads}|\mathcal{C}_1) = p(\text{effect}|\text{cause}).$$

But of more interest are the 'inversed' probabilities, e.g.:

$$p(\mathcal{C}_1|\text{Heads}) = p(\text{cause}|\text{effect}),$$

because this is the useful direction in statistical inference:

> *"Given the observed data (effect), what can be said about the underlying mechanism that generated them (cause)?"*

If we already knew the underlying mechanism, why would we be collecting data?. . .

So, $p(\text{cause}|\text{effect})$ is of great interest.
And it is Bayes'rule that allows us to compute it!

Bayes' rule

$$\boxed{p(A|B) = \frac{p(A)p(B|A)}{p(B)}}$$

Let's write Bayes' rule in the usual statistics form.

- Assume we want to estimate a parameter, say θ (e.g., the population mean).
- We collect data to update our beliefs about θ.

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

Or, in words,

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}.$$

Bayes' rule

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

About the constituent components of the Bayes' rule:

- Prior: A probability distribution representing our prior beliefs about θ. I.e., what do we think about the value of θ before we observe the data?
- Likelihood: The statistical model. Allows expressing the probability distribution for the data, at any fixed θ value.
- Marginal likelihood: Also referred to as the *weighted* likelihood. Represents the probability of the data across all possible values of θ.
- Posterior: A probability distribution representing our beliefs about θ after we observe the data.

We next visit each of the constituent parts of the Bayes' rule in turn.

Running example

We will use the ubiquitous example of the coin bias:

- We have a coin, we want to learn about its bias.
- Denote $\theta = p(\text{heads})$.
- Assume independence between coin throws.
- Further assume that θ is fixed between throws.
- $N =$ total number of throws;
  $X =$ number of heads in N throws.

Under the above conditions, the number of heads in a total of N throws can be modeled by means of the binomial model:

$$p(X = x|N, \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x},$$

where $\binom{N}{x} = \frac{N!}{(N-x)!x!}$, with $k! = k(k-1)\cdots 2 \cdot 1$ (note: $0! = 1$).

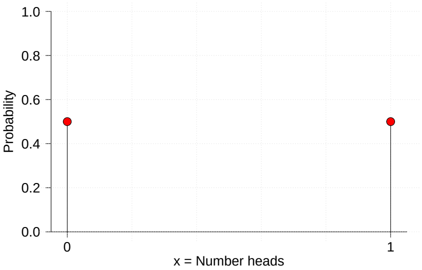
Say what?...

Running example

$$p(X = x|N, \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

One throw, fair coin:

θ	N	x	$p(X = x\|N, \theta)$
.5	1	0	$\binom{1}{0} .5^0 (1-.5)^{1-0} = .5$
		1	$\binom{1}{1} .5^1 (1-.5)^{1-1} = .5$
			Sum $= 1$

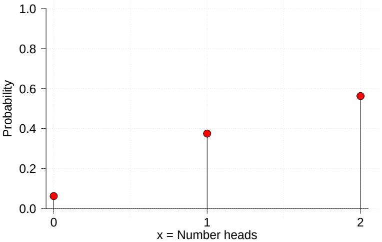


Running example

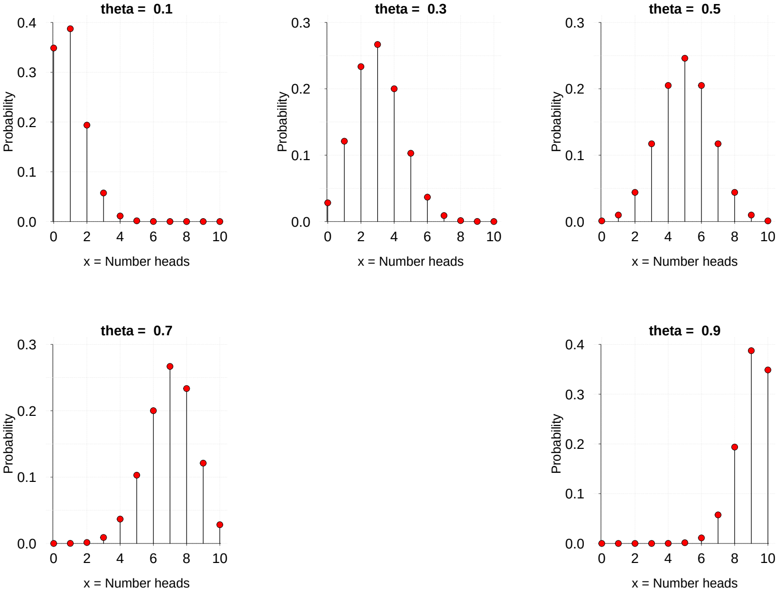
$$p(X = x|N, \theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}$$

Two throws, biased coin:

θ	N	x	$p(X = x\|N, \theta)$
		0	$\binom{2}{0}.75^0(1-.75)^{2-0} = .0625$
.75	2	1	$\binom{2}{1}.75^1(1-.75)^{2-1} = .375$
		2	$\binom{2}{2}.75^2(1-.75)^{2-2} = .5625$
			Sum $= 1$



Running example ($N = 10$)



Prior

$$\boxed{\text{Bayes' rule: } p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})}}$$

$p(\theta)$: A probability distribution representing our prior beliefs about θ. I.e., what do we think about the value of θ before we observe the data?

- The prior expresses our uncertainty about θ, before the data.
- We can look at θ in different ways:
  - As fixed (but unknown).
  - As variable between sample draws.

  In either case, the prior reflects our imprecise knowledge about it.

Bayesian statistics = rational updating of beliefs.
We need to have initial beliefs to begin with!
Otherwise, we can't update or learn anything.

Priors are the initial beliefs that we wish to rationally update, in light of data.

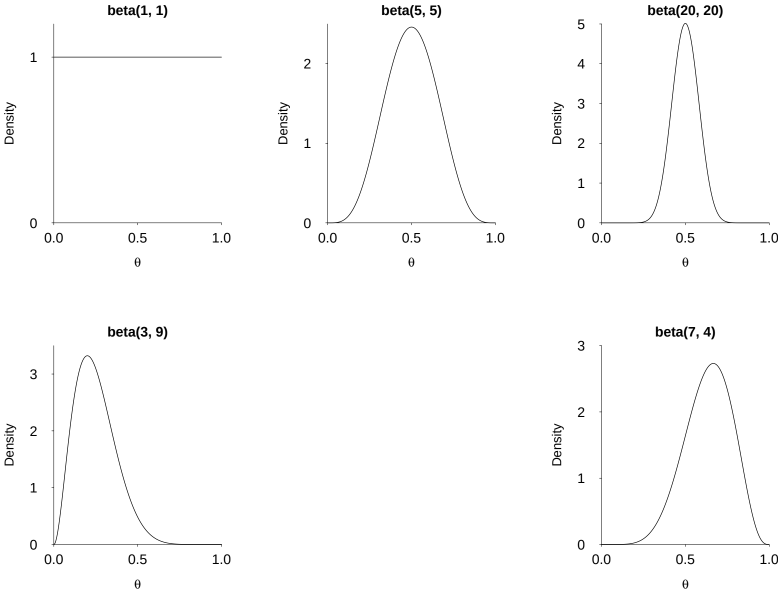
Prior

How to choose a prior?

- Tricky.
  (But, for that matter, frequentism is also full of choices to be made.
  At least choice is made more explicit in Bayes.)
- There are two schools of thought:
  - Objective: Choose 'default' (vague, diffuse) priors, previously calibrated, that apply to a wide range of similar analyses.
  - Subjective: Choose priors that reflect one's own belief, or the current state of the art.

For the coin bias where $\theta = p(\text{heads})$, what shall the prior distribution be?
We focus on the beta family, to keep tradition (it is JASP's default for the binomial model).

Prior


beta(1, 1)
Density
1
0
0.0
0.5
1.0
θ
beta(5, 5)
Density
2
1
0
0.0
0.5
1.0
θ
beta(20, 20)
Density
5
4
3
2
1
0
0.0
0.5
1.0
θ
beta(3, 9)
Density
3
2
1
0
0.0
0.5
1.0
θ
beta(7, 4)
Density
3
2
1
0
0.0
0.5
1.0
θ

Likelihood

$$\boxed{\text{Bayes' rule: } p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})}}$$

$p(\text{data}|\theta)$: The statistical model. Allows expressing the probability distribution for the data, at any fixed θ value.

Do note that:

- For any fixed θ, $p(\text{data}|\theta)$ is a probability distribution for the data.
- However, $p(\text{data}|\theta)$ is not a probability distribution once θ is allowed to vary!!!

And for Bayesians, parameters are random...
(Yeah, I know.)

Conclusion:

- $p(\text{data}|\theta)$ is not a probability distribution once θ varies.
- Some books write $\mathcal{L}(\theta|\text{data})$ instead of $p(\text{data}|\theta)$ to emphasize that the likelihood is a (not probability!) function of θ for fixed data.
  We don't do it: Notation '$p(\text{data}|\theta)$' is just too common.

Likelihood

Let's see that likelihood $\neq$ probability with an example:
Probability of observing 0, 1, or 2 heads in $N = 2$ throws, for various θ.

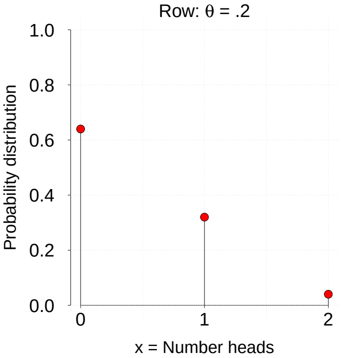
$$\boxed{p = \binom{N}{x}\theta^x(1-\theta)^{N-x}}$$

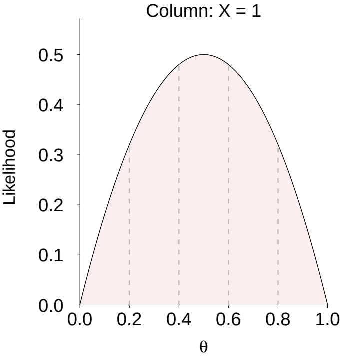
θ	Number of heads			Sum
	0	1	2	
0	1	0	0	1
.2	0.64	0.32	0.04	1
.4	0.36	0.48	0.16	1
.6	0.16	0.48	0.36	1
.8	0.04	0.32	0.64	1
1	0	0	1	1
Sum	2.2	1.6	2.2	

Observe that:

- For a fixed θ (i.e., per row), the sum of the probabilities across all possible data is 1.
- For fixed data (i.e., per column), the sum of the probabilities across θ is not 1 (and notice that we only entertained a few θ values in $[0, 1]$!...).

Likelihood





The likelihood function (right) is not a probability distribution (area under the curve $= .333 \neq 1$).

Likelihood

How do we choose a statistical model (i.e., a likelihood function)?

- Actually, the same way we do it in frequentism.
  E.g.: The binomial model for the coin bias, the normal model for IQ, . . .
- Let's demystify something:

  All models are wrong

  (but some are useful; Box, 1978).
- Models are simple and crude attempts, based on a plethora of assumptions, to explain the complex world we live in.

What one must ask is whether the model at hand is useful to:

- Explain the observed data.
- Make predictions for future observations.
- Build insight, possibly leading to improved models.
- Quantify uncertainty (yes, do embrace uncertainty!).

Marginal likelihood

$$\boxed{\text{Bayes' rule: } p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})}}$$

$p(data)$: Also referred to as the weighted *likelihood.*
Represents the probability of the data across all possible values of θ.

- $p(\text{data})$ is fully determined by the prior and likelihood. It is not chosen.
- It is a marginal (i.e., across θ) probability distribution for the data.
- It does not depend on θ.
- It is a normalizing constant, allowing the posterior to be a proper distribution (i.e., with total probability 1).
- It is hard to compute except for a few simple cases.

Often the Bayes' rule is written without $p(\text{data})$ explicitly, like this:

$$p(\theta|\text{data}) \propto p(\theta)p(\text{data}|\theta),$$

or, in words,

posterior is proportional to prior times likelihood.

Marginal likelihood

$$\boxed{\text{Bayes' rule: } p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})}}$$

Formula to compute marginal likelihood:

- When θ is discrete (not the coin bias example):

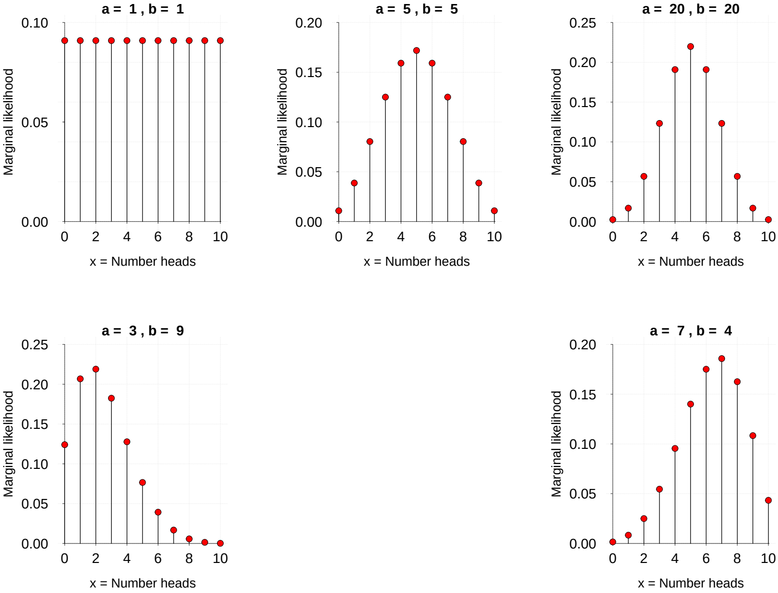
$$p(\text{data}) = \sum_{\text{All } \theta} p(\theta)p(\text{data}|\theta).$$

- When θ is continuous (hold on to your chairs...):

$$p(\text{data}) = \int_{\text{All } \theta} p(\theta)p(\text{data}|\theta)d\theta.$$

Think of the integral as the 'continuous analogue of the sum'.
Do not worry about it: You are not expected to compute integrals!!

Coin bias (θ continuous): $N = 10$, various beta(a, b) priors, binomial model



Coin bias, θ discrete: $N = 2$, binomial model

For simplicity, assume that $\theta = p(\text{heads})$ can only assume three value: 0, .5, 1. In this case the marginal likelihood is

$$p(\text{data}) = \sum_{\text{All } \theta} p(\theta) p(\text{data}|\theta).$$

Likelihood

We again use the binomial model:

$$p(X = x|N, \theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}$$

Prior

Assume we suspect the coin to be biased, although we can't tell if it favors heads or tails. One possible prior distribution is as follows:

$$p(\theta) = \begin{cases} .45, & \theta = 0 \\ .1, & \theta = .5 \\ .45, & \theta = 1 \end{cases}.$$

Coin bias, θ discrete: $N = 2$, binomial model

θ	Prior	Likelihood			Prior×Likelihood		
		0	1	2	0	1	2
0	.45	1	0	0	0.45	0	0
.5	.1	0.25	0.5	0.25	0.025	0.05	0.025
1	.45	0	0	1	0	0	0.45
				$p(\text{data}) =$	0.475	0.05	0.475

Note that:

- The sum of the likelihood values is 1, per row (i.e., the likelihood is a valid probability distribution for θ fixed).
- $p(\text{data})$ is indeed a valid probability distribution:

$$p(\text{data}) = \begin{cases} .475, & x = 0 \\ .05, & x = 1 \\ .475, & x = 2 \end{cases}, \text{ with } .475 + .05 + .475 = 1.$$

Posterior

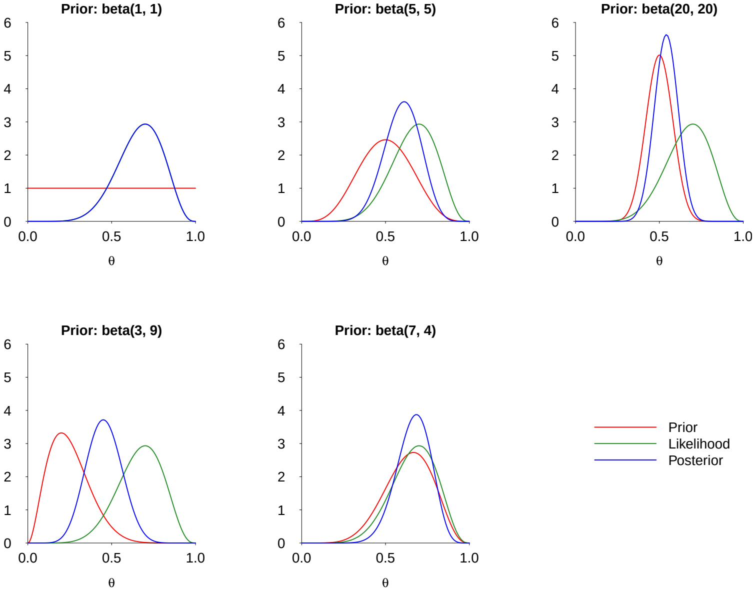
$$\boxed{p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})}}$$

The posterior is, arguably, the main goal of Bayesian inference.
It reflects the updated knowledge about θ, in light of the observed data.

The posterior is a compromise between the prior and the likelihood.
It incorporates our prior beliefs about θ and the info provided by the data.

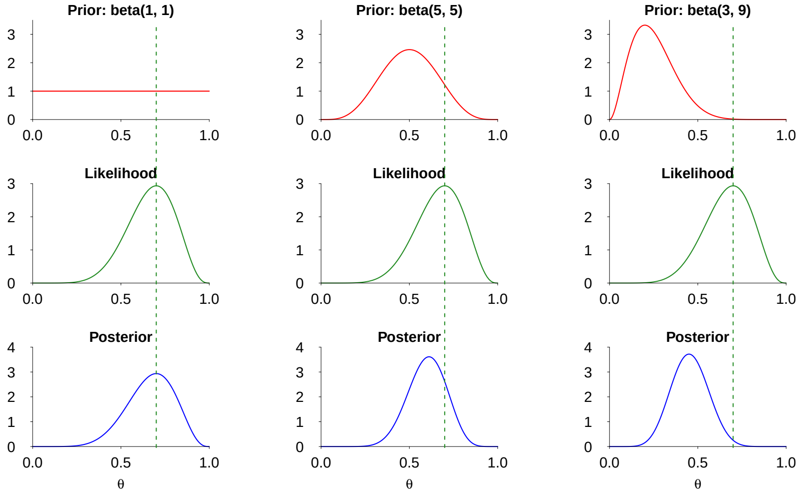
The Bayes' rule offers a rational, mathematically valid, means of combining both pieces of information.

Posterior: $N = 10$, $x = 7$

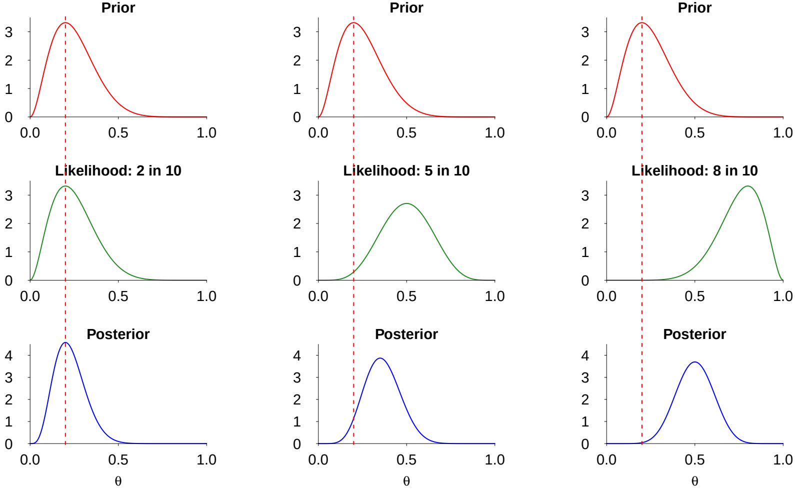


Posterior

How does the prior affect the posterior (fixed data: $N = 10$, $x = 7$)?

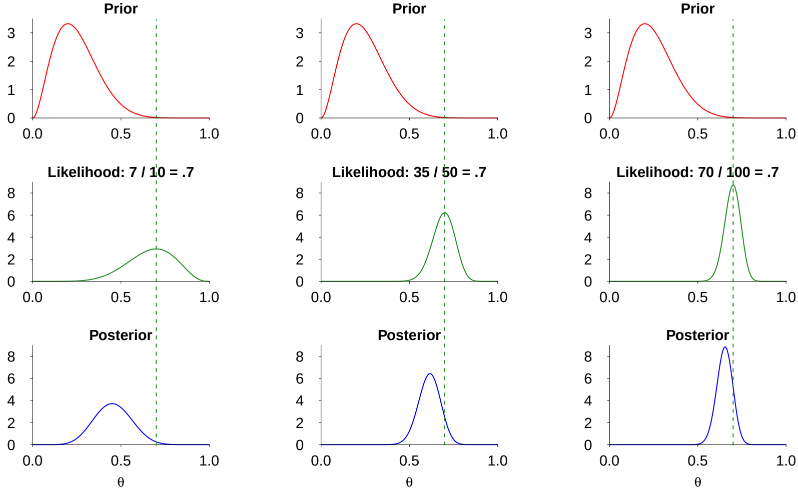


How does the sample mean affect the posterior (fixed prior and $N = 10$)?



Posterior

How does the sample size affect the posterior (fixed prior and $\frac{x}{N} = .7$)?



Posterior – When updating our belief is impossible

If $p(\theta) = 0$ then $p(\theta|\text{data}) = \frac{p(\theta)p(\text{data}|\theta)}{p(\text{data})} = 0$.
I.e., we cannot update completely dogmatic prior beliefs, no matter how compelling the data are against our initial belief.

Similarly, if your prior is

$$p(\theta) = \begin{cases} 1, & \theta = .5 \\ 0, & \text{otherwise} \end{cases},$$

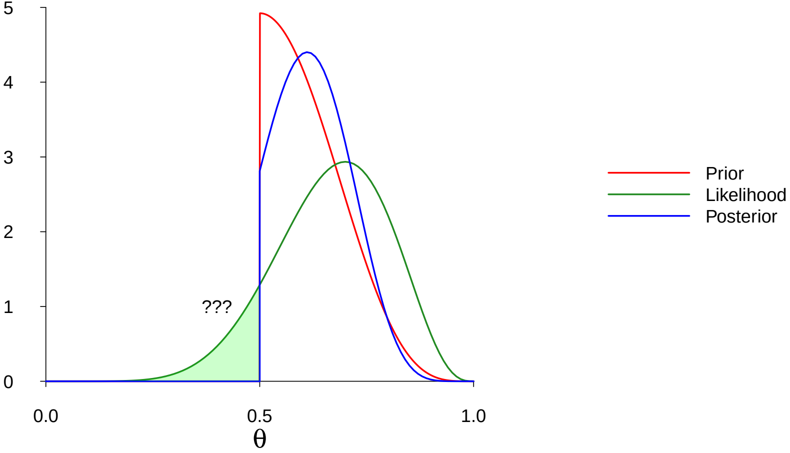
then $p(\theta|\text{data}) = p(\theta)$, no matter how much the evidence in the data implies otherwise.

General rule: Avoid dogmatic priors (Cromwell's rule).

Posterior – When updating our belief is impossible

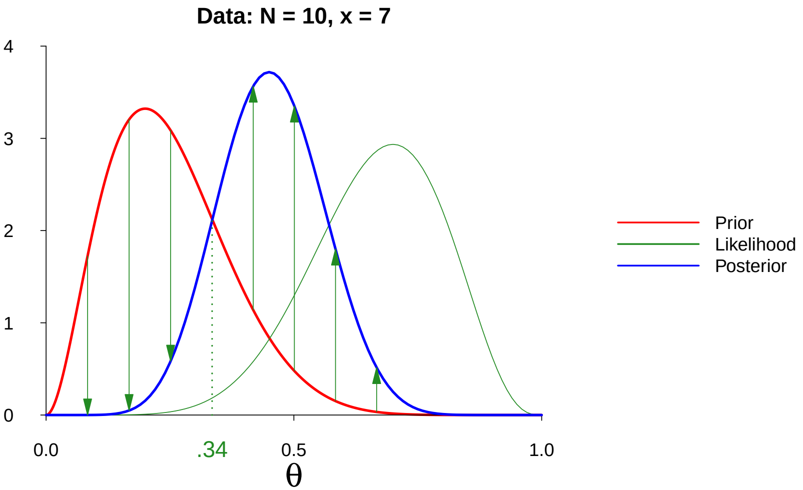
Data: $N = 10$, $x = 7$.

Prior: beta(5, 5) truncated in interval $(.5, 1)$.



Posterior – Reallocation of credibility

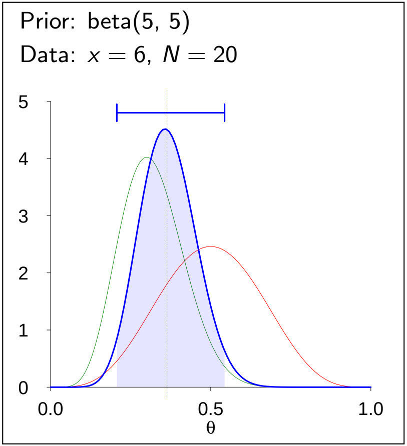
“Bayesian inference is reallocation of credibility across possibilities.”



Posterior

We can summarize the information in the posterior in multiple ways:

- Point estimates.
- Interval estimates.



Prior: beta(5, 5)

Data: $x = 6$, $N = 20$

- median = 0.36
- sd = 0.09
- 95% central credible interval = (0.21, 0.54)

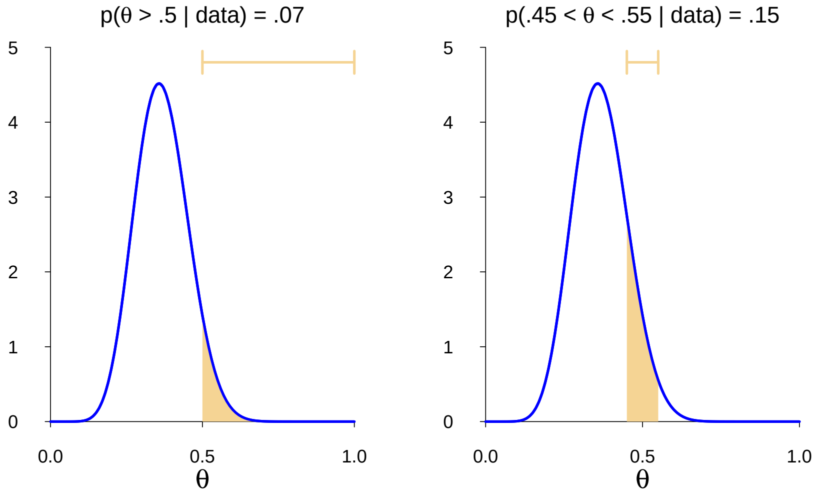
Interpretation:

There is a 95% probability that θ lies between .21 and .54.

Compare this interpretation with that from a 95% confident interval (very different)!!

Posterior

We can compute any posterior probabilities we want!



Posterior

Q: How do we do the math? This all sounds too hard to compute…
A: With computers. JASP and R are your friends!

Only for a small handful of models (i.e., combinations of prior and likelihood) do we know the closed form of the posterior.
The coin bias example is one such case (you need to know this!):

- Data: X successes in N independent trials.
- Likelihood: $X|\theta \sim \text{binomial}(N, \theta)$.
- Prior: $\theta \sim \text{beta}(a, b)$.
- Then posterior: $\theta|X \sim \text{beta}(a + X, b + N - X)$.
  So, add the number of successes X to the 1st parameter, and the number of failures $(N - X)$ to the 2nd parameter. That's it.

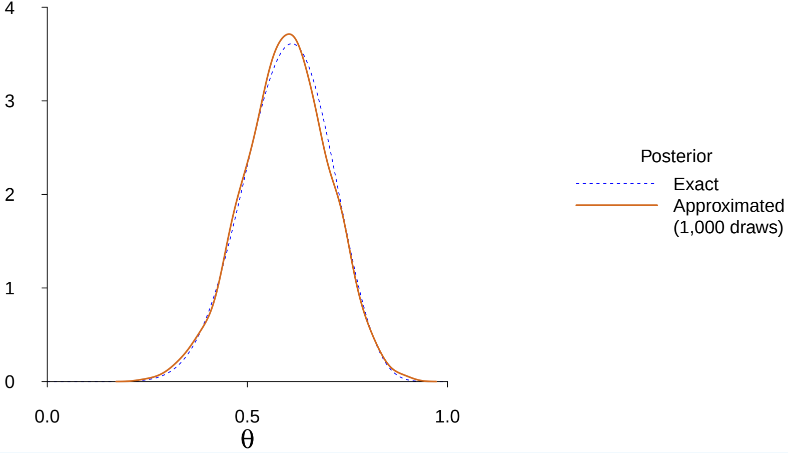
In general, we don't have a formula for $p(\theta|\text{data})$.
But computers help us to *sample* from $p(\theta|\text{data})$. If we draw a large sample from $p(\theta|\text{data})$, then we have a fairly good approximation of the posterior!
This is what MCMC (Markov Chain Monte Carlo) is all about. But not today.

Sampling from the posterior

Data: $N = 10$, $x = 7$.

Prior: beta(5, 5).

Therefore, posterior = beta($5 + 7$, $5 + 10 - 7$) = beta(12, 8).



For the next lecture

Contents:

- Good statistics, bad statistics

Read:

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2001). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*, 1359-1366. doi:10.1177/0956797611417632

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*, 524-532. doi:10.1177/0956797611430953