PSYC 573 Bayesian Data Analysis (2024 Fall): Course Notes

Hok Chio (Mark) Lai

2024-08-13

Table of contents

Preface

There will be some math in this notes. Don't worry if you feel the math is challenging; for applied focused students, it is much more important to understand the concepts of Bayesian methods than to understand the mathematical symbols, as they usually can be handled by the software.

Part I

Week 1

1 Introduction

1.1 History of Bayesian Statistics

Here is a nice brief video that covers some of the 250+ years of history of Bayesian statistics:

https://www.youtube.com/watch?v=BcvLAw-JRss

If you are interested in learning more about the story, check out the nice popular science book, "The theory that would not die," by McGrayne (2011)

1.1.1 Thomas Bayes (1701–1762)

You may find a biography of Bayes from https://www.britannica.com/biography/Thomas-Bayes. There is also a nice story in the book by Lambert (2018). He was an English Presbyterian minister. The important work he wrote that founded Bayesian statistics was "An Essay towards solving a Problem in the Doctrine of Chances", which he did not publish and was later discovered and edited by his friend, Richard Price, after Bayes's death [1]

1.1.2 Pierre-Simon Laplace (1749–1827)

Laplace, a French Mathematician, was an important figure in not just Bayesian statistics, but also in other areas of mathematics, astronomy, and physics. We actually know much more the work by Laplace than by Bayes, and Laplace has worked independently on the inverse probability problem (i.e., $P[\text{Parameter}|\text{Data}]$). Indeed, he was credited for largely formalizing Bayesian interpretation of probability and most of the machinery for Bayesian statistics, and making it a useful technique to be applied to different problems, despite the discipline being called "Bayesian." His other contributions include the methods of least square and the central limit theorem. See a short biography of him at https://www.britannica.com/biography/Pierre-Simon-marquis-de-Laplace.

[1] Price is another important figure in mathematics and philosopher, and have taken Bayes' theorem and applied it to insurance and moral philosophy.

1.1.3 20th Century

Until early 1920s, the *inverse probability* method, which is based on what is now called Bayes's Theorem, is pretty much the predominant point of view of statistics. Then a point of view later known as *frequentist* statistics arrived, and quickly became the mainstream school of thinking for statistical inferences, and is still the major framework for quantitative research. In the early 1920s, frequentist scholar, most notably R. A. Fisher and Jerzy Neyman, criticized Bayesian inference for the use of subjective elements in an objective discipline. In Fisher's word,

> The theory of inverse probability is founded upon an error, and must be wholly rejected—Fisher, 1925

Ironically, the term *Bayesian* was first used in one of Fisher's work. And interestingly, Fisher actually thought he "have been doing almost exactly what Bayes had done in the 18th century."[2]

Despite criticisms from frequentist scholars, Bayesian methods has been used by scholars in the Allies in World War II, such as Alan Turing, in an algorithm to break coded messages in the Enigma machine that the German Navy used to communicate. However, because of the more complex mathematics involved in Bayesian statistics, Bayesian statistics is limited to straight-forward problems and theoretical discussions until the early 1980s, when computing speed increases tremendously and makes *Markov Chain Monte Carlo*—the major algorithm for Bayesian estimation in modern Bayesian statistics—feasible. With the help of increased computing speed, Bayesian statistics has come back and been used as an alternative way of thinking, especially given growing dissatisfaction towards the misuse of frequentist statistics by some scholars across disciplines. Bayesian estimation methods have also been applied to many new research questions where frequentist approaches work less well, as well as in big data analytics and machine learning.

1.2 Motivations for Using Bayesian Methods

Based on my personal experience, Bayesian methods is used quite often in statistics and related departments, as it is consistent and coherent, as contrast to frequentist where a new and probably ad hoc procedure needed to be developed to handle a new problem. For Bayesian, as long as you can formulate a model, you just run the analysis the same way as you would for simpler problems, or in Bayesian people's word "turning the Bayesian crank," and likely the difficulties would be more technical than theoretical, which is usually solved with better computational speed.

Social and behavioral scientists are relatively slow to adopt the Bayesian method, but things have been changing. In a recently accepted paper by Van De Schoot et al. (2017), the authors

[2] See the paper by John Aldrich on this.

reviewed papers in psychology between 1990 and 2015 and found that whereas less than 10% of the papers in 1990 to 1996 mentioned "Bayesian", the proportion increased steadily and was found in close to 45% of the psychology papers in 2015. Among studies using Bayesian methods, more than 1/4 cited computational problems (e.g., nonconvergence) in frequentist methods as a reason, and about 13% cited the need to incorporate prior knowledge into the estimation process. The other reasons included the flexibility of Bayesian methods for complex and nonstandard problems, and the use of techniques traditionally attached to Bayesian such as missing data and model comparisons.

1.2.1 Problem with classical (frequentist) statistics

The rise of Bayesian methods is also related to the statistical reform movement in the past two decades. The problem is that applied researchers are obsessed with $p < .05$ and often misinterpreted a small p-value as something that it isn't (read Gigerenzer, 2004). Some scholars coined the term p-hacking to refer to the practice of obtaining statistical significance by choosing to test the data in a certain way, either consciously or subconsciously (e.g., dichotomizing using mean or median, trying the same hypothesis using different measures of the same variable, etc). This is closely related to the recent "replication crisis" in scientific research, with psychology being in the center under close scrutiny.

Bayesian is no panacea to the problem. Indeed, if misused it can give rise to the same problems as statistical significance. My goal in this class is to help you appreciate the Bayesian tradition of embracing the uncertainty in your results, and adopt rigorous model checking and comprehensive reporting rather than relying merely on a p-value. I see this as the most important mission for someone teaching statistics.

1.3 Comparing Bayesian and Frequentist Statistics

Attributes	Frequentist	Bayesian
Interpretation of probability	Frequentist	Subjectivist
Uncertainty	How estimates vary in repeated sampling from the same population	How much prior beliefs about parameters change in light of data
What's relevant?	Current data set + all that might have been observed	Only the data set that is actually observed
How to proceed with analyses	MLE; ad hoc and depends on problems	"Turning the Bayesian crank"

1.4 Software for Bayesian Statistics

The following summarizes some of the most popular Bayesian software. Currently, JAGS and STAN are the most popular. General statistical programs like SPSS, SAS, and Stata also have some support for Bayesian analyses as well.

- WinBUGS
    - Bayesian inference Using Gibbs Sampling
    - Free, and most popular until late 2000s. Many Bayesian scholars still use WinBUGS
    - No further development
    - One can communicate from R to WinBUGS using the package `R2WinBUGS`
- JAGS
    - Just Another Gibbs Sampler
    - Very similar to WinBUGS, but written in C++, and support user-defined functionality
    - Cross-platform compatibility
    - One can communicate from R to JAGS using the package `rjags` or `runjags`
- STAN
    - Named in honour of Stanislaw Ulam, who invented the Markov Chain Monte Carlo method
    - Uses new algorithms that are different from Gibbs sampling
    - Under very active development
    - Can interface with R through the package `rstan`, and the R packages `rstanarm` and `brms` automates the procedure for fitting models in STAN for many commonly used models

Part II

Week 2

2 Probability

2.1 History of Probability

2.1.1 Games of chance

Correspondence between French Mathematicians (Pierre de Fermat and Blaise Pascal) on gambling problem by Antoine Gombaud, Chevalier de Méré. The problem is roughly of the form[1]:

> Imagine two people playing a multi-round game. In each round, each person has an equal chance of winning. The first person who wins six rounds will get a huge cash prize. Now, consider a scenario in which A and B have played six rounds, where A has won five and B has won one. At that time, the game had to be stopped due to a thunderstorm. Since neither A and B have reached six wins, instead of giving the prize to either one of them, they agree to divide up the prize. What would be a fair way to do so?

The discussion led to the formalization of using mathematics to solve the problem. Basically, one way is to say if A has a 97% chance to win the prize eventually and B has a 3% chance, then A should get 97% of the prize.

2.2 Different Ways to Interpret Probability

There are multiple perspectives for understanding probability.[2] What you've learned in your statistics training is likely based on the *frequentist* interpretation of probability (and thus frequentist statistics), whereas what you will learn in this class have the foundation on the *subjectivist* interpretation of probability. Understanding the different perspectives on probability is helpful for understanding the Bayesian framework.

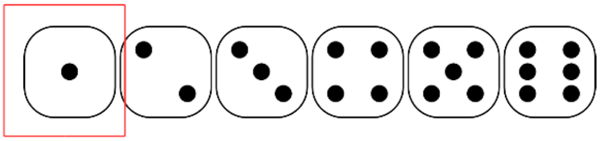
> 💡 You don't need to commit to one interpretation of probability in order to conduct Bayesian data analysis.

[1] see the exact form at https://en.wikipedia.org/wiki/Problem_of_points

[2] See http://plato.stanford.edu/entries/probability-interpret/ for more information

2.2.1 Classical Interpretation

This is an earlier perspective and is based on counting rules. The idea is that probability is equally distributed among all "indifferent" outcomes. "Indifferent" outcomes are those where a person does not have any evidence to say that one outcome is more likely than another. For example, when one throws a die, one does not think that a certain number is more likely than another unless one knows that the die is biased. In this case, there are six equally likely outcomes, so the probability of each outcome is 1 / 6.



2.2.2 Frequentist Interpretation

The frequentist interpretation states that probability is essentially the long-run relative frequency of an outcome. For example, to find the probability of getting a "1" when throwing a die, one can repeat the experiment many times, as illustrated below:

Trial	Outcome
1	2
2	3
3	1
4	3
5	1
6	1
7	5
8	6
9	3
10	3

And we can plot the relative frequency of "1"s in the trials:

As you can see, with more trials, the relative frequency approaches 1 / 6. It's the reason why in introductory statistics, many of the concepts require you to think in terms of repeated sampling (e.g., sampling distribution, *p*-values, standard errors, confidence intervals), because

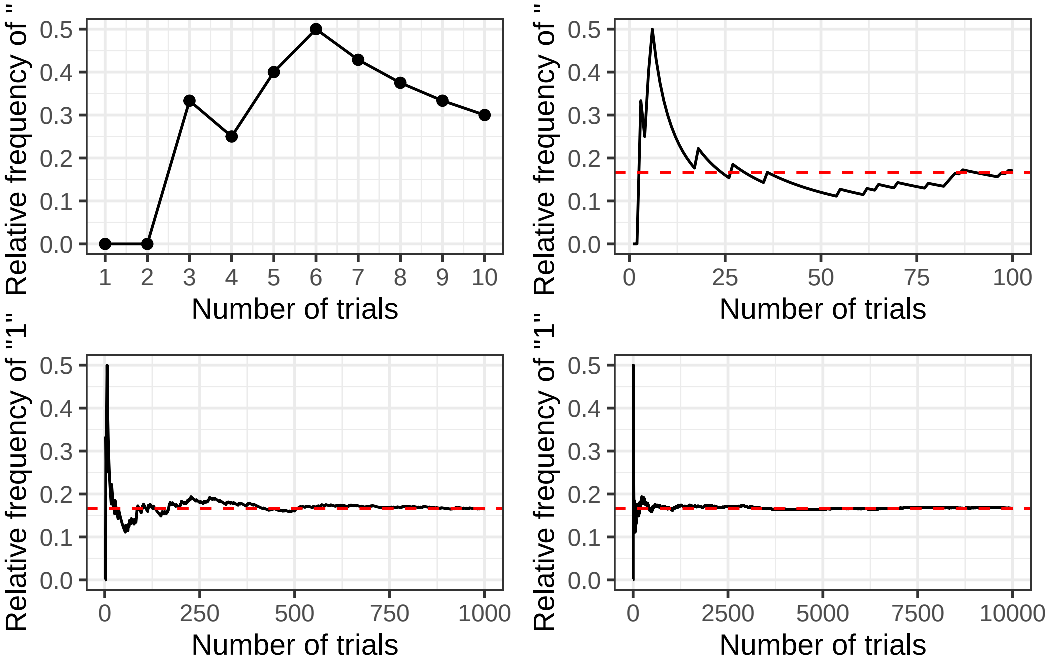


Figure 2.1: Relative frequency when repeatedly rolling a die.

probability in this framework is only possible when the outcome can be repeated. It's also the reason why we don't talk about something like:

- the probability of the null hypothesis being true, or
- the probability that the population mean is in the interval [75.5, 80.5],

because the population is fixed and cannot be repeated. Only the samples can be repeated, so probability in frequentist statistics is only about samples.

2.2.2.1 Problem of the single case

Because of the frequentist's reference to long-run relative frequency, under this framework, it does not make sense to talk about the probability of an event that cannot be repeated. For example, it does not make sense to talk about

- the probability that the Democrats/Republicans will win the 2028 US Presidential Election, or
- the probability that the LA Rams winning the 2024 Super Bowl, or
- the probability that it will rain on Christmas Day in LA in 2024,

because all these are specific events that cannot be repeated. However, it is common for lay people to talk about probabilities or chances for these events.

2.2.3 Subjectivist Interpretation

The frequentist interpretation is sometimes called the "objectivist view," as the reference of probability is based on empirical evidence of long-run relative frequency (albeit hypothetical in many cases). In contrast, the *subjectivist* view of probability is based on one's belief. For example, when I say that the probability of getting a "1" from rolling a die is 1 / 6, it reflects the state of my mind about the die. My belief can arise from different sources: Maybe I make the die and know it is a fair one; maybe I saw someone throwing the die 1,000 times, and the number of "1"s was close to 1,000 / 6, or maybe someone I trust and with authority says that the die has a 1-in-6 chance of showing a "1".

The "subjective" component has been criticized a lot by frequentist scholars, sometimes unfairly. To be clear, what "subjective" here means is that probability reflects the state of one's mind instead of the state of the world, and so it is totally fine that two people can have different beliefs about the same event. However, it does not mean that probability is arbitrary, as the beliefs are subjected to the constraints of the axioms of probability as well as the condition that the person possessing such beliefs is *rational*.[3] Therefore, if two persons are exposed to the same information, they should form similar, though likely not identical, beliefs about the event.

The subjective interpretation works perfectly fine with single events, as one can have a belief about whether it rains on a particular day or a belief about a particular election result.

2.2.3.1 Calibrating a subjective belief

In order to represent one's belief by probability, one needs to assign a nonzero value to every plausible outcome of an event. This has been the job of odds-makers for a long time. Indeed, a lot of the development in the field of probability has to do with coming up with a fair bet. The process of assigning probabilities to outcomes of an event according to one's belief is called *calibration*. For example, consider three possible outcomes for tomorrow's weather. For simplicity, consider three mutually exclusive possible outcomes: sunny, cloudy, and rainy.

To calibrate my belief, consider first if you bet \$10, and the return is (a) \$30 for sunny, (b) \$30 for cloudy, and (c) \$30 for rainy. Which one will you bet? If you're like me in LA, I'm pretty sure I'll bet on (a), as I think that it is more likely to have a sunny day. This means that setting $P(\text{sunny}) = P(\text{cloudy}) = P(\text{rainy}) = 1\ /\ 3$ is not a good reflection of my belief.

[3] In a purely subjectivist view of probability, assigning a probability P to an event does not require any justifications, as long as it follows the axioms of probability. For example, I can say that the probability of me winning the lottery and thus becoming the wealthiest person on earth tomorrow is 95%, which by definition would make the probability of me not winning the lottery 5%. Most Bayesian scholars, however, do not endorse this version of subjectivist probability and require justifications of one's beliefs (that has some correspondence to the world).

Now consider the bet with the returns (a) \$20 for sunny, (b) \$30 for cloudy, and (c) \$60 for rainy. This would reflect the belief that there is a 50% chance of a sunny day, 33.33% chance of a cloudy day, and 16.67% chance of a rainy day. Will you take the bet? This is an improvement from the last one, but I would still say a sunny day is a good bet, which suggests that the probability of 50% is too low for a sunny day. The idea is to continue iterating until it is hard to consider (a), (b), or (c) as a clear betting favorite. For me, this would end up being something like (a) \$16.7 for sunny, (b) \$33.3 for cloudy, and (c) \$100 for rainy, which would correspond to 60% sunny, 30% cloudy, and 10% rainy.

If it's hard for you to consider the gambling analogy, an alternative way is to consider how many times is a sunny day more likely than a non-sunny day, and how many times is a cloudy day more likely than a rainy day. For example, I may consider a sunny day to be twice as likely as a non-sunny day, which would give the probability of a sunny day to be 66.67%. Then, if I also think that a cloudy day is three times as likely as a rainy day, I would assign a probability of $33.33\% \times 3 / 4 = 25\%$ for a cloudy day, and a probability of $33.33\% \times 1 / 4 = 8.33\%$ for a rainy day.

The process of calibrating one's belief plays a key role in Bayesian data analysis, namely in the form of formulating a *prior* probability distribution.

2.3 Basics of Probability

Kolmogorov axioms

For an event A_i (e.g., getting a "1" from throwing a die)

- $P(A_i) \geq 0$ [All probabilities are non-negative]
- $P(A_1 \cup A_2 \cup \cdots) = 1$ [Union of all possibilities is 1]
- $P(A_1) + P(A_2) = P(A_1 \text{ or } A_2)$ [Addition rule]

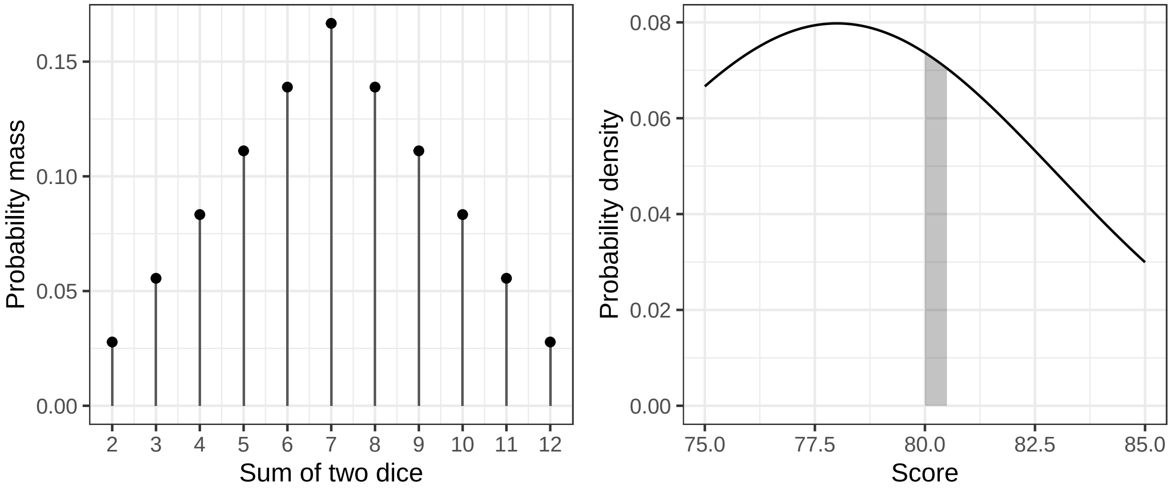
Consider two events, for example, on throwing a die,

- A: The number is odd
- B: The number is larger than or equal to 4

Assuming that die is (believed to be) fair, you can verify that the probability of A is $P(A) = 3 / 6 = 1 / 2$, and the probability of B is also $P(B) = 3 / 6 = 1 / 2$.

2.3.1 Probability Distributions

- Discrete event (e.g., outcome of throwing a die or an election): probability *mass*. The probability is nonzero, at least for some outcomes. The graph below on the left shows the probability mass of the sum of the numbers from two dice.
- Continuous event (e.g., temperature): probability *density*.[4] The probability is basically zero for any outcome. Instead, the probability density is approximated by $P(A \leq a \leq A + h)/h$ for a very small h.
  - For example, to find the probability density that a person's well-being score is 80, we first find the probability that a person scores between 80 and 80.5 (or 80 and 80.0005), and divide that probability by 0.5 (or 0.0005). See the shaded area of the graph below on the right.



(a) Probability mass

(b) Probability density

Figure 2.2: Examples of probability distributions

For this course, as in the textbook, we use $P(x)$ to mean both the probability mass for an outcome x when the event is discrete, and the probability density at an outcome x when the event is continuous.

[4] For many problems in the social and behavioral sciences, the measured variables are not truly continuous, but we still use continuous distributions to approximate them.

2.3.1.1 Example: Normal Distribution

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

```r
# Write a function to compute the density of an outcome x
# for a normal distribution
my_normal_density <- function(x, mu, sigma) {
    exp(- ((x - mu) / sigma) ^2 / 2) / (sigma * sqrt(2 * pi))
}
# For example, density at x = 36 in a normal distribution
# with mu = 50 and sigma = 10
my_normal_density(36, mu = 50, sigma = 10)
```

```
#> [1] 0.01497275
```

2.3.2 Summarizing a Probability Distribution

While it is useful to know the probability mass/density of every possible outcome, in many situations, it is helpful to summarize a distribution by some numbers.

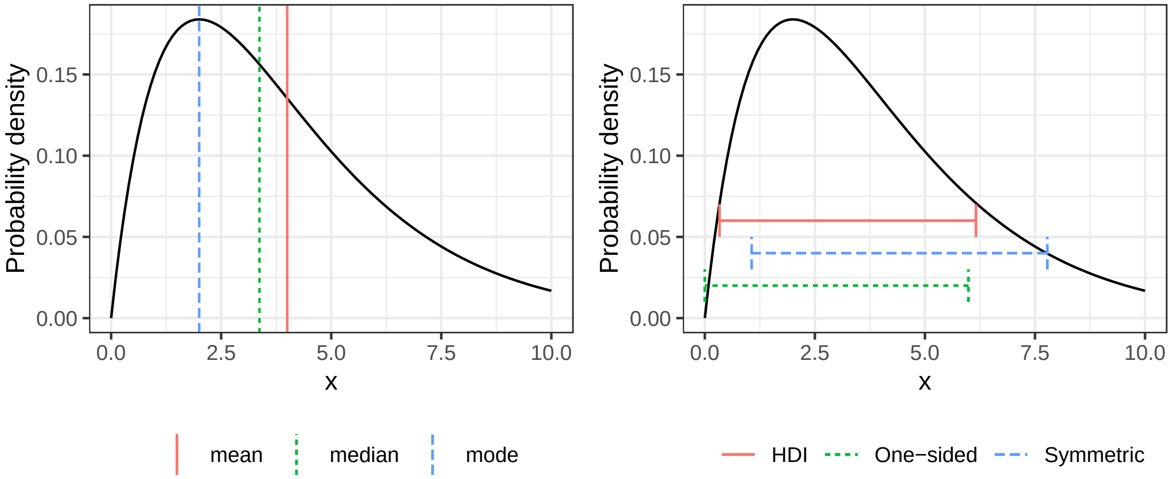
2.3.2.1 Central Tendency

- Mean: $E(X) = \int x \cdot P(x) dx$
- Median: 50th percentile; the median of X is Mdn_X such that $P(X \leq Mdn_X) = 1\ /\ 2$
- Mode: A value with maximum probability mass/density

See Figure 2.3a for examples.

2.3.2.2 Dispersion

- Variance: $V(X) = E[X - E(X)]^2$
  - Standard deviation: $\sigma(X) = \sqrt{V(X)}$
- Median absolute deviation (MAD): $1.4826 \times Mdn(|X - Mdn_X|)$



(a) Central Tendency

(b) Interval

Figure 2.3: Measures of central tendency and interval

2.3.2.3 Interval

Use $C(X)$ to denote an interval. A $W\%$ interval means $P(X \in C[X]) \approx W\%$

- One-sided interval: $C(X)$ is half-bounded
- Symmetric $W\%$ interval: $C(X) = [L(X), U(X)]$ is bounded, with $P(X < L[X]) = P(X > U[X]) \approx W\%/2$
  - Also called *equal-tailed* interval
- Highest density $W\%$ interval (HDI): $P(x_c) \geq P(x_o)$ for every x_c in $C(X)$ and every x_o outside $C(X)$. In general, the HDI is the shortest $W\%$ interval.

The plot in Figure 2.3b shows several 80% intervals.

2.3.2.4 Computing Summaries of Sample Distributions Using R

```r
# Simulate data from a half-Student's t distribution with
# df = 4, and call it sim_s
sim_s <- rt(10000, df = 4) # can be both positive and negative
sim_s <- abs(sim_s) # take the absolute values
ggplot(data.frame(x = sim_s), aes(x = x)) +
    geom_histogram(binwidth = 0.1)
```

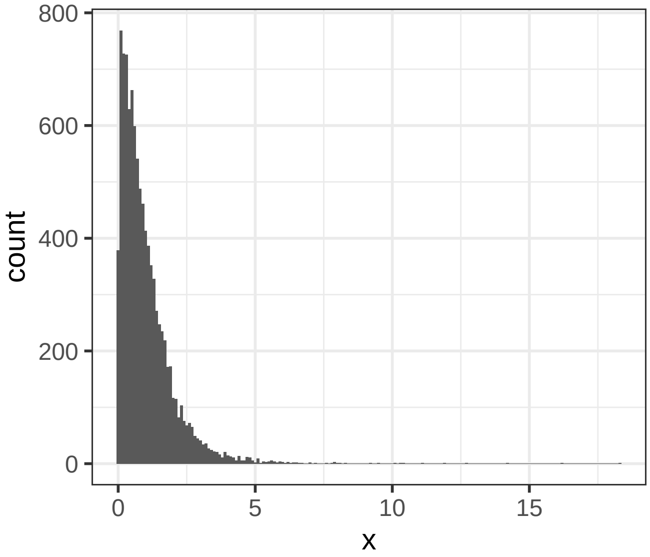


Figure 2.4: Simulated density of a half-Student's t distribution

```r
# Central tendency
# (note: the mode is difficult to compute for continuous
# variables, and rarely used in this course.)
c(mean = mean(sim_s),
  median = median(sim_s),
  mode = density(sim_s, bw = "SJ")$x[
      which.max(density(sim_s, bw = "SJ")$y)
  ])
```

```
#>      mean    median      mode
#> 1.0082926 0.7426326 0.1021776
```

```r
# Dispersion
c(
    variance = var(sim_s),
    sd = sd(sim_s),
    mad = mad(sim_s)
)
```

```
#>  variance        sd       mad
#> 1.0231058 1.0114869 0.6996741
```

```
# 80% Interval
c(`0%` = 0, quantile(sim_s, probs = .8)) # right-sided
```

```
#>     0%    80%
#> 0.0000 1.5598
```

```
c(quantile(sim_s, probs = .2), `100%` = Inf) # left-sided
```

```
#>       20%      100%
#> 0.2680906       Inf
```

```
quantile(sim_s, probs = c(.1, .9)) # equal-tailed/symmetric
```

```
#>       10%       90%
#> 0.1299027 2.1371636
```

```
HDInterval::hdi(sim_s)
```

```
#>        lower        upper
#> 0.0003616342 2.7992620765
#> attr(,"credMass")
#> [1] 0.95
```

2.3.3 Multiple Variables

- Joint probability: $P(X, Y)$
- Marginal probability:

$$P(X) = \int P(X, y) dy$$

$$P(Y) = \int P(x, Y) dx$$

  - The probability that outcome X happens, regardless of what values Y take.

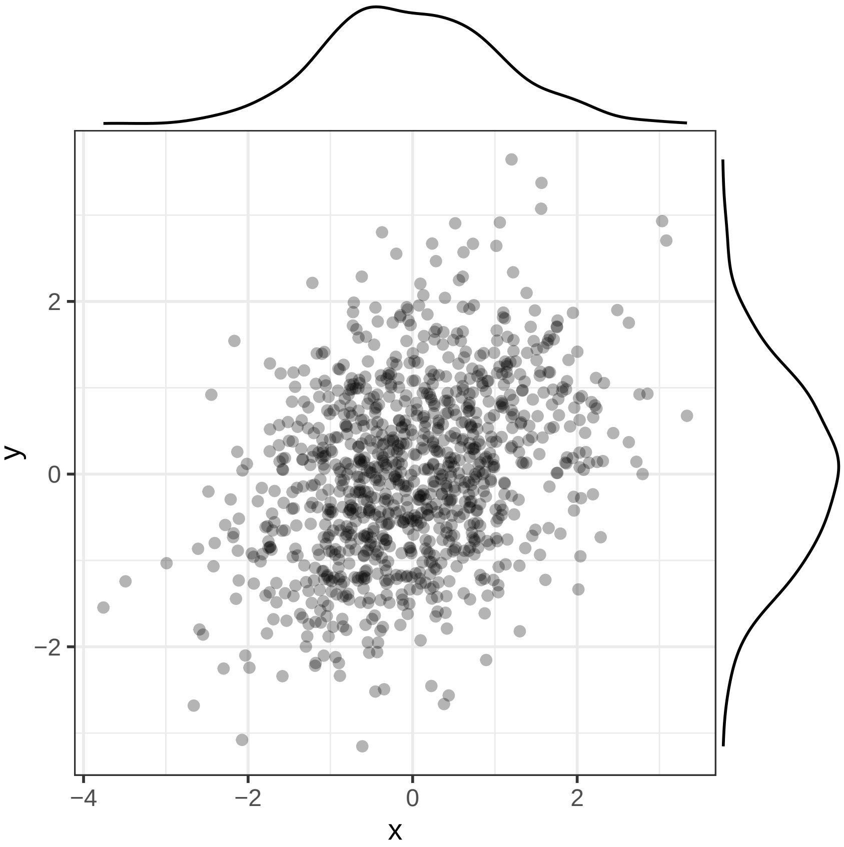


Figure 2.5: Joint and Marginal Distributions

2.3.3.1 Conditional Probability

Conditional probability is the probability of an event given some other information. In the real world, you can say that everything is conditional. For example, the probability of getting an odd number on throwing a die is 1/2 is conditional on the die being fair. We use $P(A \mid B)$ to represent the the conditional probability of event A given event B..

Continuing from the previous example, $P(A \mid B)$ is the conditional probability of getting an odd number, *knowing that the number is at least 4*. By definition, conditional probability is the probability that both A and B happen, divided by the probability that B happens.

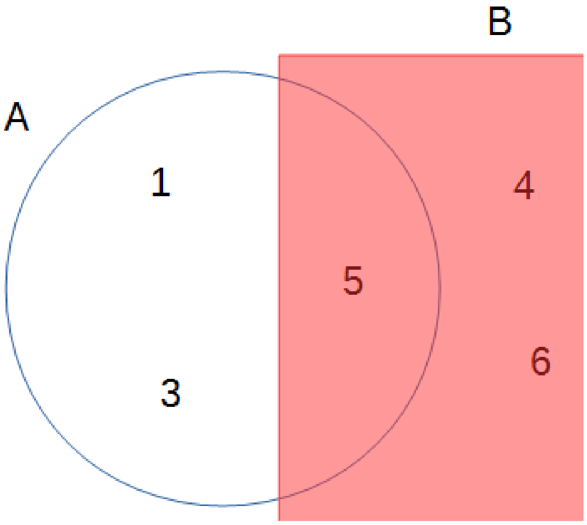
❗ **Conditional Probability**

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

In the example, $P(A, B) = 1 \ / \ 6$, because 5 is the only even number ≥ 4 when throwing a die. Thus,

$$\begin{aligned} P(A \mid B) &= 1/3 \\ &= \frac{P(A, B)}{P(B)} \\ &= \frac{1/6}{1/2} \end{aligned}$$

This picture should make it clear:



❗ Please recognize that $P(A \mid B) \neq P(B \mid A)$. For example, when throwing a die, $P(\text{number is six} \mid \text{even number}) = 1/3$, but $P(\text{even number} \mid \text{number is six})$ is 1.

2.3.3.2 Independence

❗ Two events, A and B, are independent if $P(A \mid B) = P(A)$

This means that any knowledge of B does not (or should not) affect one's belief about A. Consider the example:

- A: A die shows five or more
- B: A die shows an odd number

Here is the joint probability

	>= 5	<= 4
odd	1/6	2/6
even	1/6	2/6

So the conditional probability of $P(>= 5 \mid \text{odd}) = (1/6) \;/\; (1/2) = 1/3$, which is the same as $P(>= 5 \mid \text{even}) = (1/6) \;/\; (1/2) = 1/3$. Similarly it can be verified that $P(<= 4 \mid \text{odd}) = P(<= 4 \mid \text{even}) = 2/3$. Therefore, A and B are independent.

On the other hand, for the example

- A: A die shows four or more
- B: A die shows an odd number

the joint probabilities are

	>= 4	<= 3
odd	1/6	2/6
even	2/6	1/6

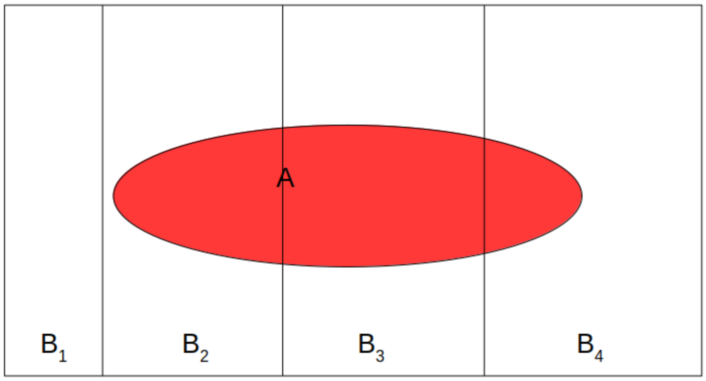
Obviously, A and B are not independent because once we know that the number is four or above, it changes the probability of whether it is an odd number or not.

❗ Independence can also be expressed as

If A and B are independent, $P(A, B) = P(A)P(B)$

2.3.4 Law of Total Probability

When we talk about conditional probability, like $B_1 = 4$ or above and $B_2 = 3$ or below, we can get $P(A \mid B_1)$ and $P(A \mid B_2)$ (see the figure below), we refer $P(A)$ as the *marginal probability*, meaning that the probability of A without knowledge of B.



If $B_1, B_2, \cdots, B_n$ are all mutually exclusive possibilities for an event (so they add up to a probability of 1), then

❗ **Law of Total Probability**

$$\begin{aligned} P(A) &= P(A, B_1) + P(A, B_2) + \cdots + P(A, B_n) \\ &= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n) \\ &= \sum_{k=1}^{n} P(A \mid B_k)P(B_k) \end{aligned}$$

3 Bayes's Theorem

The Bayes's theorem is, surprisingly (or unsurprisingly), very simple:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

More generally, we can expand it to incorporate the law of total probability to make it more applicable to data analysis. Consider B_i as one of the n many possible mutually exclusive events, then

$$\begin{aligned} P(B_i \mid A) &= \frac{P(A \mid B_i)P(B_i)}{P(A)} \\ &= \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)} \\ &= \frac{P(A \mid B_i)P(B_i)}{\sum_{k=1}^{n} P(A \mid B_k)P(B_k)} \end{aligned}$$

If B_i is a continuous variable, we will replace the sum by an integral,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\int_k P(A \mid B_k)P(B_k)}$$

The denominator is not important for practical Bayesian analysis, therefore, it is sufficient to write the above equality as

$$P(B_i \mid A) \propto P(A \mid B_i)P(B_i)$$

3.1 Example 1: Base rate fallacy (From Wikipedia)

3.1.1 Question

A police officer stops a driver *at random* and does a breathalyzer test for the driver. The breathalyzer is known to detect true drunkenness 100% of the time, but in 1% of the cases, it gives a false positive when the driver is sober. We also know that in general, for every 1,000 drivers passing through that spot, one is driving drunk. Suppose that the breathalyzer shows positive for the driver. What is the probability that the driver is truly drunk?

3.1.2 Solution

$P(\text{positive}|\text{drunk}) = 1$
$P(\text{positive}|\text{sober}) = 0.01$
$P(\text{drunk}) = 1/1000$
$P(\text{sober}) = 999/1000$

Using Bayes' Theorem,

$$\begin{aligned}P(\text{drunk}|\text{positive}) &= \frac{P(\text{positive}|\text{drunk})P(\text{drunk})}{P(\text{positive}|\text{drunk})P(\text{drunk}) + P(\text{positive}|\text{sober})P(\text{sober})} \\ &= \frac{1 \times 0.001}{1 \times 0.001 + 0.01 \times 0.999} \\ &= 100/1099 \approx 0.091\end{aligned}$$

So there is less than a 10% chance that the driver is drunk even when the breathalyzer shows positive.

You can verify that with a simulation using R:

```r
set.seed(4)
truly_drunk <- c(rep("drunk", 100), rep("sober", 100 * 999))
table(truly_drunk)
```

```
#> truly_drunk
#> drunk sober
#>   100 99900
```

```r
breathalyzer_test <- ifelse(truly_drunk == "drunk",
    # If drunk, 100% chance of showing positive
    "positive",
    # If not drunk, 1% chance of showing positive
    sample(c("positive", "negative"), 999000,
        replace = TRUE, prob = c(.01, .99)
    )
)
# Check the probability p(positive | sober)
table(breathalyzer_test[truly_drunk == "sober"])
```

```
#>
#> negative positive
#>    98903      997
```

```r
# 997 / 99900 = 0.00997998, so the error rate is less than 1%
# Now, Check the probability p(drunk | positive)
table(truly_drunk[breathalyzer_test == "positive"])
```

```
#>
#> drunk sober
#>   100   997
```

```r
# 100 / (100 + 997) = 0.0911577, which is only 9.1%!
```

3.2 Bayesian Statistics

`Bayesian statistics` is a way to estimate some parameter θ (i.e., some quantities of interest, such as the population mean, regression coefficient, etc) by applying the Bayes' Theorem.

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

There are three components in the above equality:

- $P(D|\theta)$, the probability that you observe data D, given the parameter θ; this is called the `likelihood` (Note: It is the likelihood of θ, but probability about y)
- $P(\theta)$, the probability distribution θ, without referring to the data D. This usually requires appeals to one's degree of belief, and so is called the `prior`
- $P(\theta|y)$, the updated probability distribution of θ, after observing the data D; this is called the `posterior`

On the other hand, classical/frequentist statistics focuses solely on the likelihood function.[1] In Bayesian statistics, the goal is to update one's belief about θ based on the observed data D.

[1] The likelihood function in classical/frequentist statistics is usually written as $P(y;\theta)$. You will notice that here I write the likelihood for classical/frequentist statistics to be different from the one used in Bayesian statistics. This is intentional: In frequentist conceptualization, θ is fixed, and it does not make sense to talk about the probability of θ. This implies that we cannot condition on θ, because conditional probability is defined only when $P(\theta)$ is defined.

3.3 Example 2: Locating a Plane

Consider a highly simplified scenario of locating a missing plane in the sea. Assume that we know the plane, before missing, happened to be flying on the same latitude, heading west across the Pacific, so we only need to find the longitude of it. We want to go out to collect debris (data) so that we can narrow the location (θ) of the plane down.

3.3.1 Prior

We start with our prior. Assume that we have some rough idea that the plane should be, so we express our belief in a probability distribution like the following:

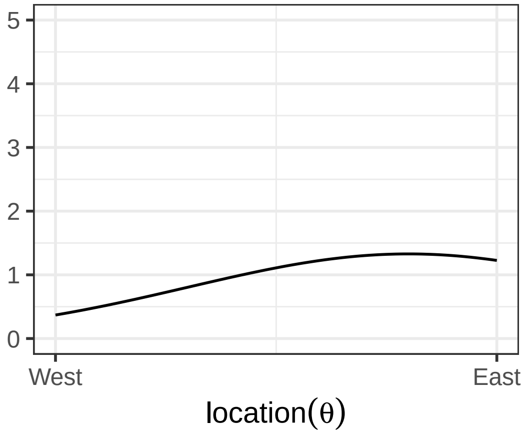


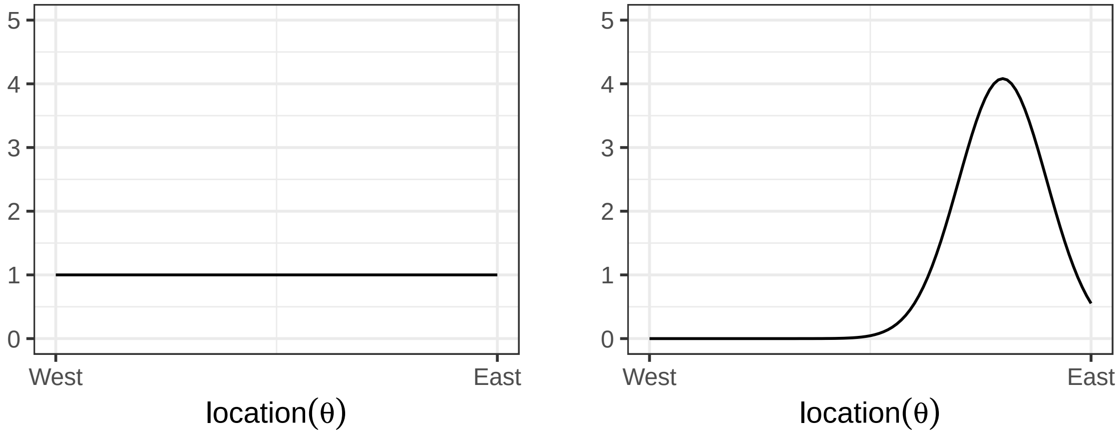
Figure 3.1: Prior distribution.

which says that our belief is that the plane is about twice more likely to be towards the east than towards the west. Below are two other options for priors (out of infinitely many), one providing virtually no information and the other encoding stronger information:

The prior is chosen to reflect the researcher's belief, so it is likely that different researchers will formulate a different prior for the same problem, and that's okay as long as the prior is reasonable and justified. Later we will learn that in regular Bayesian analyses, with moderate sample size, different priors generally make only negligible differences.

3.3.2 Likelihood

Now, assume that we have collected debris in the locations shown in the graph,



(a) Noninformative prior.

(b) Informative prior.

Figure 3.2: More options for prior distribution.

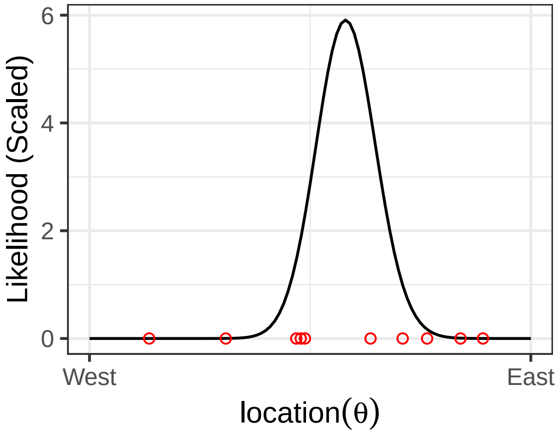


Figure 3.3

3.3.3 Posterior

Now, from Bayes's Theorem,

$$\text{Posterior Probability} \propto \text{Prior Probability} \times \text{Likelihood}$$

So we can simply multiply the prior probabilities and the likelihood to get the posterior probability for every location. A rescaling step is needed to ensure that the area under the curve will be 1, which is usually performed by the software.

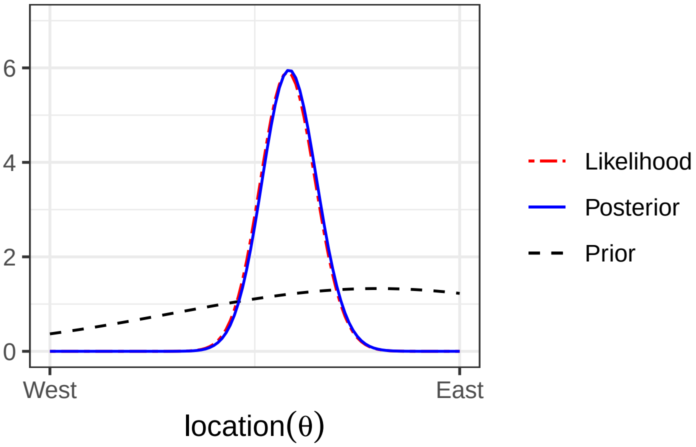


Figure 3.4

As illustrated below, the posterior distribution is a synthesis of (a) the prior and (b) the data (likelihood).

3.3.4 Influence of Prior

Figure 3.5 shows what happen with a stronger prior:

3.3.5 Influence of More Data

Figure 3.6 shows what happen with 20 more data points:

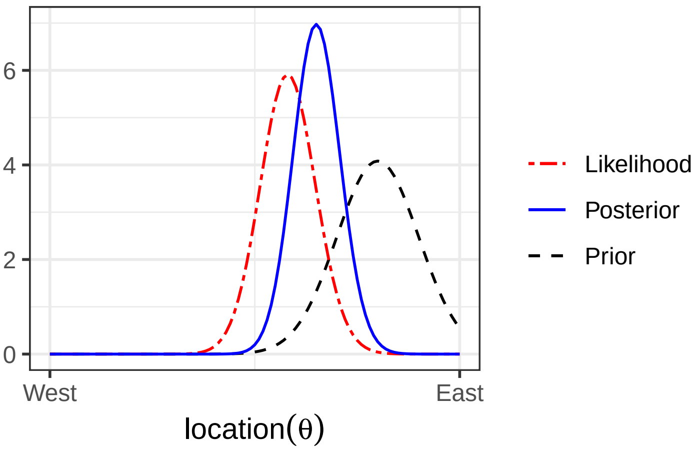


Figure 3.5

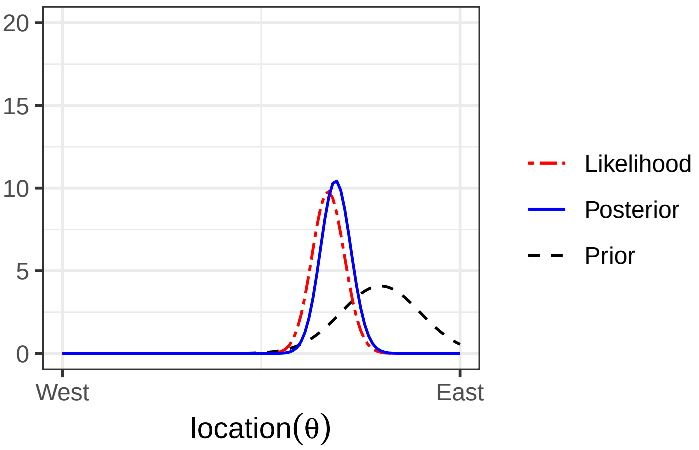


Figure 3.6

3.4 Data-Order Invariance

In many data analysis applications, researchers collect some data D_1, and then collect some more data D_2. An example would be researchers conducting two separate experiments to study the same research question. In Bayesian statistics, one can consider three ways to obtain the posterior:

1. Update the belief with D_1, and then with D_2
2. Update the belief with D_2, and then with D_1
3. Update the belief with both D_1 and D_2 simultaneously

Whether these three ways give the same posterior depends on whether **data-order invariance** holds. If the inference of D_1 does not depend on D_2, or vice versa, then all three ways lead to the same posterior. Specifically, if we have **conditional independence** such that

$$P(D_1, D_2 \mid \theta) = P(D_1 \mid \theta) P(D_2 \mid \theta),$$

then one can show all three ways give the same posterior (see p. 108 of Kruschke, 2015).

Exchangeability*

Exchangeability is an important concept in Bayesian statistics. Data are exchangeable when the joint distribution, $P(D_1, \dots, D_N)$, does not depend on the ordering of the data. A simple way to think about it is if you scramble the order of your outcome variable in your data set and still can obtain the same statistical results, then the data are exchangeable. An example situation where data are not exchangeable is

- D_1 is from year 1990, D_2 is from year 2020, and the parameter θ changes from 1990 to 2020

When data are exchangeable, the previously discussed conditional independence condition would generally hold.[2]

3.5 Bernoulli Likelihood

For binary data y (e.g., coin flip, pass/fail, diagnosed/not), an intuitive way to analyze is to use a Bernoulli model:

$$P(y = 1 \mid \theta) = \theta$$
$$P(y = 0 \mid \theta) = 1 - \theta'$$

[2] The de Finetti's theorem shows that when the data are exchangeable and can be considered an infinite sequence (i.e., not from a tiny finite population), then the data are conditionally independent given some θ.

which is more compactly written as

$$P(y \mid \theta) = \theta^y (1-\theta)^{(1-y)},$$

where $\theta \in [0, 1]$ is the probability of a "1". You can verify that the compact form is the same as the longer form.

3.5.1 Multiple Observations

When there are more than one y, say $y_1, \ldots, y_N$, that are conditionally independent, we have

$$\begin{aligned} P(y_1, \ldots, y_N \mid \theta) &= \prod_{i=1}^N P(y_i \mid \theta) \\ &= \theta^{\sum_{i=1}^N y_i} (1-\theta)^{\sum_{i=1}^N (1-y_i)}, \\ &= \theta^z (1-\theta)^{N-z} \end{aligned}$$

where z is the number of "1"s (e.g., number of heads in coin flips). Note that the likelihood only depends on z, not the individual ys. In other words, the likelihood is the same as long as there are z heads, regardless of when those heads occur.

Let's say $N = 4$ and $z = 1$. We can plot the likelihood in R:

```r
# Write the likelihood as a function of theta
lik <- function(th, num_flips = 4, num_heads = 1) {
    th ^ num_heads * (1 - th) ^ (num_flips - num_heads)
}
# Likelihood of theta = 0.5
lik(0.5)
```

```
#> [1] 0.0625
```

```r
# Plot the likelihood
ggplot(data.frame(th = c(0, 1)), aes(x = th)) +
    # `stat_function` for plotting a function
    stat_function(fun = lik) +
    # use `expression()` to get greek letters
    labs(x = expression(theta),
    y = "Likelihood with N = 4 and z = 1")
```

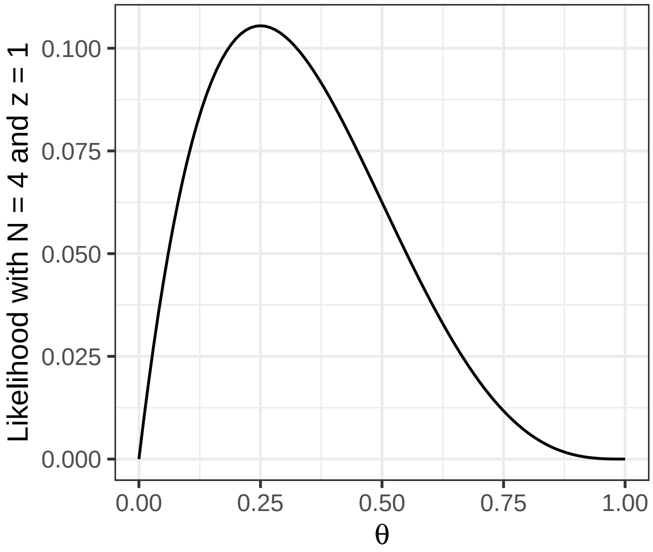


Figure 3.7: Binomial likelihood function with $N = 4$ and $z = 1$

3.5.2 Setting Priors

Remember again the relationship between the prior and the posterior:

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

The posterior distributions are mathematically determined once the priors and the likelihood are set. However, the mathematical form of the posterior is sometimes very difficult to deal with.

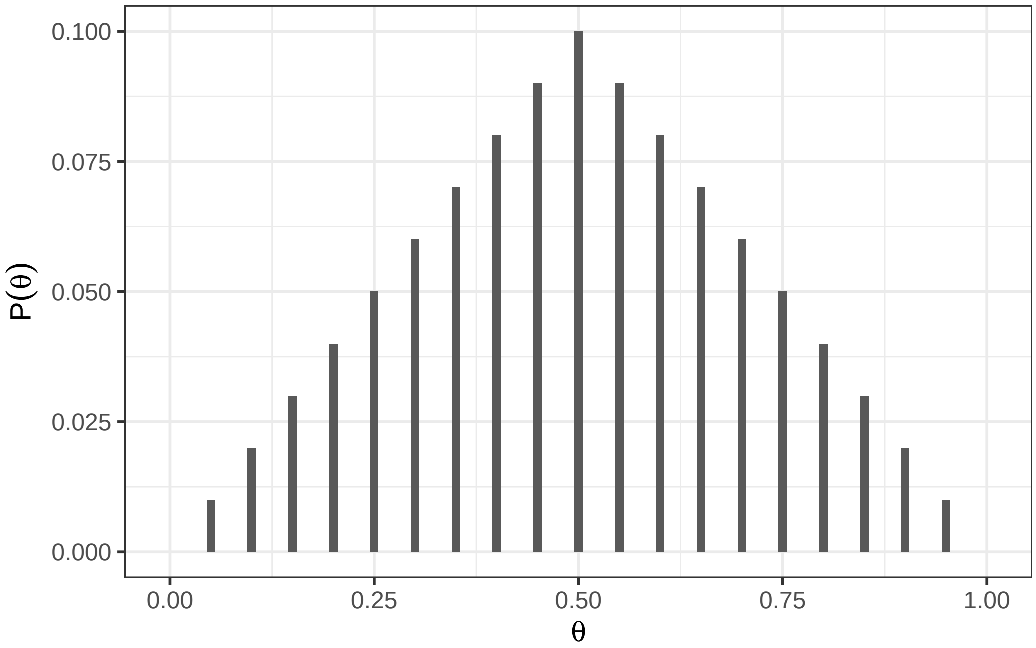
One straight forward, brute-force method is to discretize the parameter space into a number of points. For example, by taking $\theta = 0$, 0.05, 0.10, . . . , 0.90, 0.95, 1.00, one can evaluate the posterior at these 21 **grid points**.

Let's use a prior that peaks at 0.5 and linearly decreases to both sides. I assume that $\theta = 0.5$ is twice as likely as $\theta = 0.25$ or $\theta = 0.75$ to reflect my belief that the coin is more likely to be fair.

```r
# Define a grid for the parameter
grid_df <- data.frame(th = seq(0, 1, by = 0.05))
# Set the prior mass for each value on the grid
grid_df$pth <- c(0:10, 9:0)  # linearly increasing, then decreasing
# Convert pth to a proper distribution such that the value
# sum to one
```

```r
grid_df$pth <- grid_df$pth / sum(grid_df$pth)                              ①
# Plot the prior
ggplot(grid_df, aes(x = th, y = pth)) +
    geom_col(aes(x = th, y = pth),
        width = 0.01,
    ) +
    labs(y = expression(P(theta)), x = expression(theta))
```

① Note the line `grid_df$pth <- grid_df$pth / sum(grid_df$pth)`, which ensures that the probability values sum to one. This is a trick we will use to obtain the posterior probability.



Prior Predictive Distribution

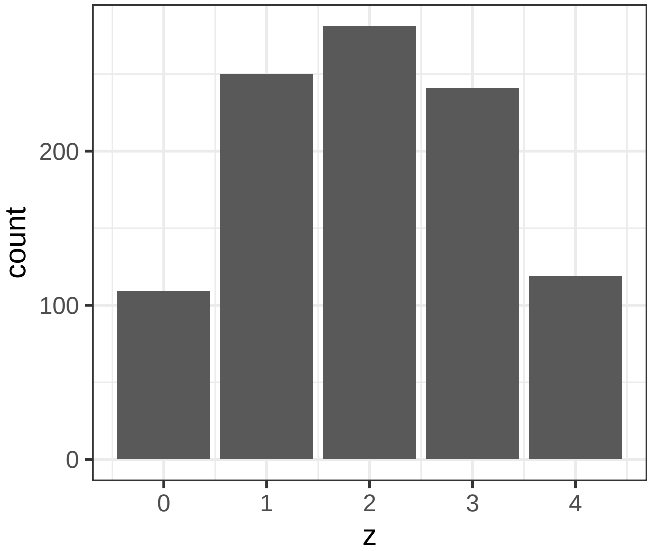
One way to check whether the prior is appropriate is to use the **prior predictive distribution**. Bayesian models are **generative** in the sense that they can be used to simulate data. The prior predictive distribution can be obtained by first simulating some θ values from the prior distribution and then simulating a data set for each θ.

```
# Draw one theta
num_trials <- 4  # number of draws
sim_th1 <- sample(grid_df$th, size = 1,
                  # based on prior probability
                  prob = grid_df$pth)
# Simulate new data of four flips based on model
sim_y1 <- rbinom(num_trials, size = 1, prob = sim_th1)

# Repeat many times
# Set number of simulation draws
num_draws <- 1000
sim_th <- sample(grid_df$th, size = num_draws, replace = TRUE,
                 # based on prior probability
                 prob = grid_df$pth)
# Use a for loop
# Initialize output
sim_y <- matrix(NA, nrow = num_trials, ncol = num_draws)
for (s in seq_len(num_draws)) {
    # Store simulated data in the sth column
    sim_y[, s] <- rbinom(num_trials, size = 1, prob = sim_th[s])
}
# Show the first 10 simulated data sets based on prior:
sim_y[, 1:10]
```

```
#>      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
#> [1,]    1    0    0    1    0    1    0    0    0     1
#> [2,]    0    0    0    1    1    0    1    0    1     1
#> [3,]    1    0    0    1    0    0    0    1    1     1
#> [4,]    1    0    1    1    1    0    0    1    1     0
```

```
# Show the distribution of number of heads
sim_heads <- colSums(sim_y)
ggplot(data.frame(z = sim_heads), aes(x = z)) +
    geom_bar()
```



The outcome seems to fit our intuition that it's more likely to be half heads and half tails, but there is a lot of uncertainty.

3.5.3 Summarizing the Posterior

```
grid_df <- grid_df %>%
    mutate(
        # Use our previously defined lik() function
        py_th = lik(th, num_flips = 4, num_heads = 1),
        # Product of prior and likelihood
        `prior x lik` = pth * py_th,
        # Scaled the posterior
        pth_y = `prior x lik` / sum(`prior x lik`)
    )
# Print a table
knitr::kable(grid_df)
```

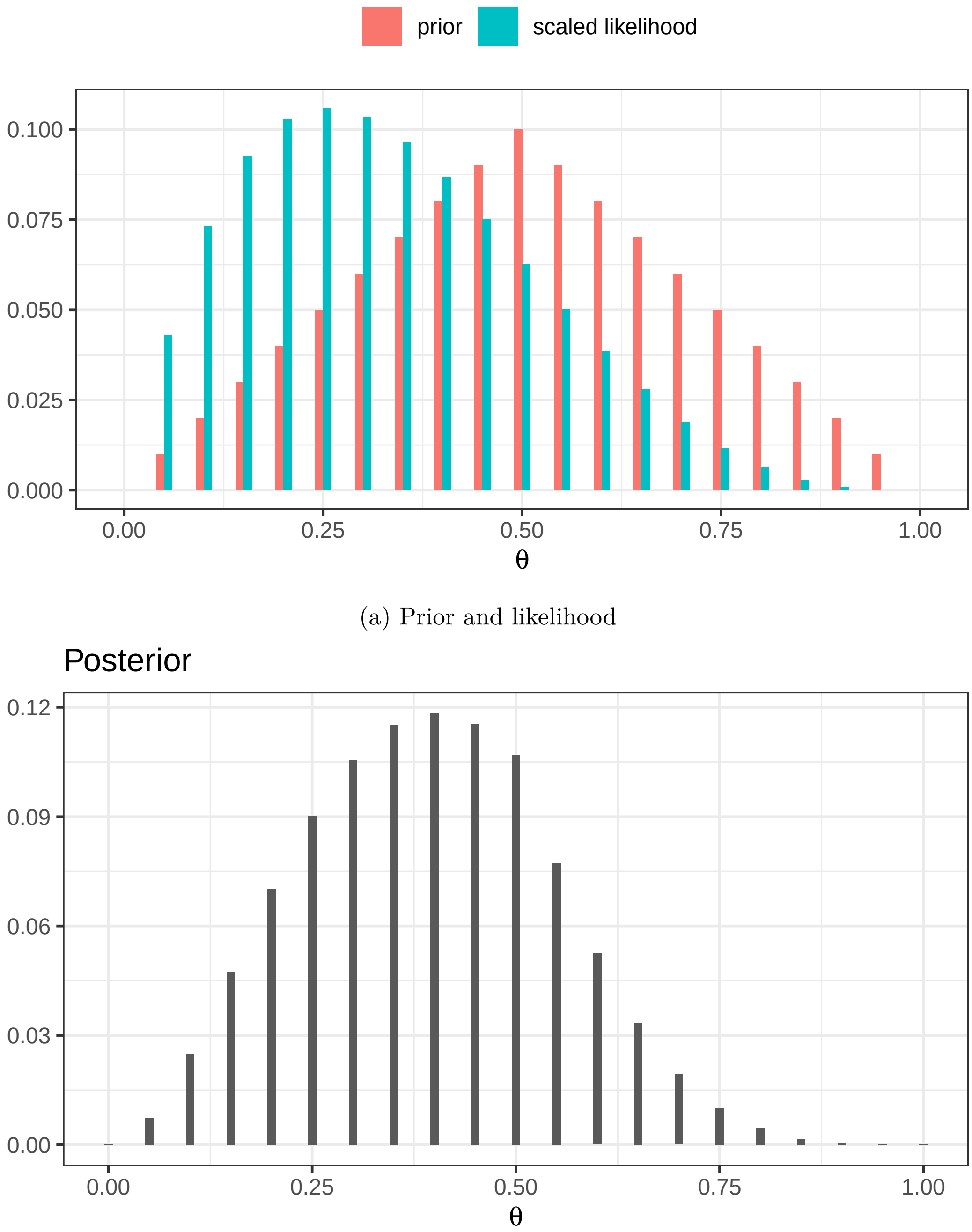
th	pth	py_th	prior x lik	pth_y
0.00	0.00	0.0000000	0.0000000	0.0000000
0.05	0.01	0.0428687	0.0004287	0.0073359
0.10	0.02	0.0729000	0.0014580	0.0249500
0.15	0.03	0.0921188	0.0027636	0.0472914

th	pth	py_th	prior x lik	pth_y
0.20	0.04	0.1024000	0.0040960	0.0700927
0.25	0.05	0.1054688	0.0052734	0.0902416
0.30	0.06	0.1029000	0.0061740	0.1056525
0.35	0.07	0.0961187	0.0067283	0.1151381
0.40	0.08	0.0864000	0.0069120	0.1182815
0.45	0.09	0.0748688	0.0067382	0.1153071
0.50	0.10	0.0625000	0.0062500	0.1069530
0.55	0.09	0.0501187	0.0045107	0.0771891
0.60	0.08	0.0384000	0.0030720	0.0525695
0.65	0.07	0.0278687	0.0019508	0.0333832
0.70	0.06	0.0189000	0.0011340	0.0194056
0.75	0.05	0.0117188	0.0005859	0.0100268
0.80	0.04	0.0064000	0.0002560	0.0043808
0.85	0.03	0.0028687	0.0000861	0.0014727
0.90	0.02	0.0009000	0.0000180	0.0003080
0.95	0.01	0.0001187	0.0000012	0.0000203
1.00	0.00	0.0000000	0.0000000	0.0000000

```r
# Plot the prior/likelihood and the posterior
ggplot(data = grid_df, aes(x = th)) +
    geom_col(aes(x = th - 0.005, y = pth, fill = "prior"),
        width = 0.01,
    ) +
    geom_col(aes(x = th + 0.005, y = py_th / sum(py_th),
        fill = "scaled likelihood"), width = 0.01,
    ) +
    labs(fill = NULL, y = NULL, x = expression(theta)) +
    theme(legend.position = "top")
ggplot(data = grid_df, aes(x = th)) +
    geom_col(aes(x = th, y = pth_y), width = 0.01) +
    labs(
        fill = NULL, y = NULL, title = "Posterior",
        x = expression(theta)
    )
```

Figure 3.8b shows the posterior distribution, which represents our updated belief about θ. We can summarize it by simulating θ values from it and compute summary statistics:

```r
# Define a function for computing posterior summary
summ_draw <- function(x) {
```



(a) Prior and likelihood

(b) Posterior

Figure 3.8: Bernoulli posterior distribution

```
    c(
        mean = mean(x),
        median = median(x),
        sd = sd(x),
        mad = mad(x),
        `ci.1` = quantile(x, prob = .1, names = FALSE),
        `ci.9` = quantile(x, prob = .9, names = FALSE)
    )
}
# Sample from the posterior
post_samples <- sample(
    grid_df$th,
    size = 1000, replace = TRUE,
    prob = grid_df$pth_y
)
summ_draw(post_samples)
```

```
#>      mean    median        sd       mad      ci.1      ci.9
#> 0.3848000 0.4000000 0.1538429 0.1482600 0.2000000 0.6000000
```

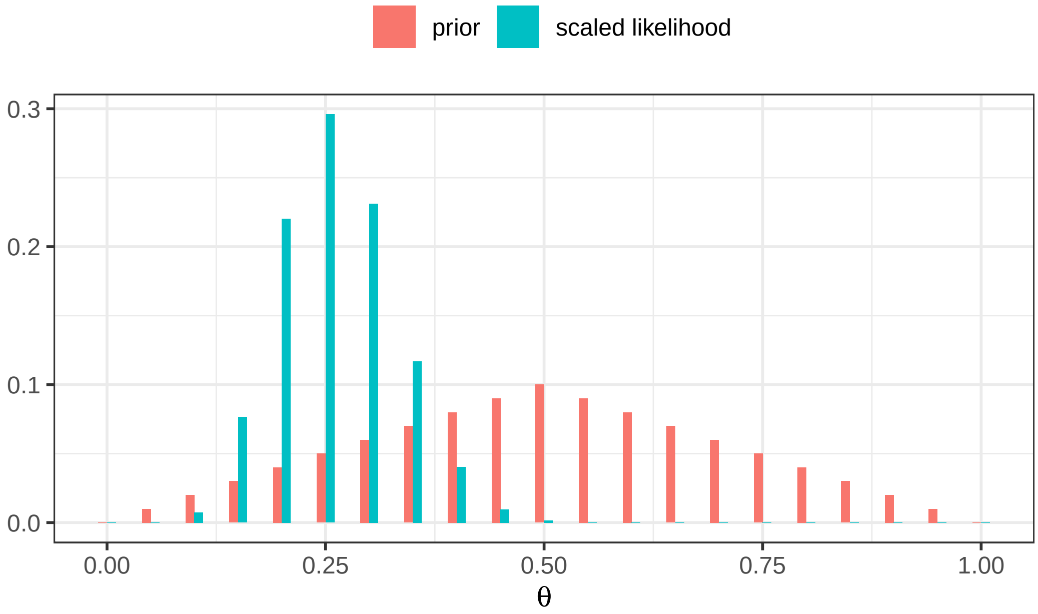
```
# Alternatively, use the `posterior` package
data.frame(theta = post_samples) |>
    posterior::summarize_draws()
```

```
#> # A tibble: 1 x 10
#>   variable  mean median    sd   mad    q5   q95  rhat ess_bulk ess_tail
#>   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
#> 1 theta    0.385    0.4 0.154 0.148  0.15  0.65  1.00    1030.     721.
```

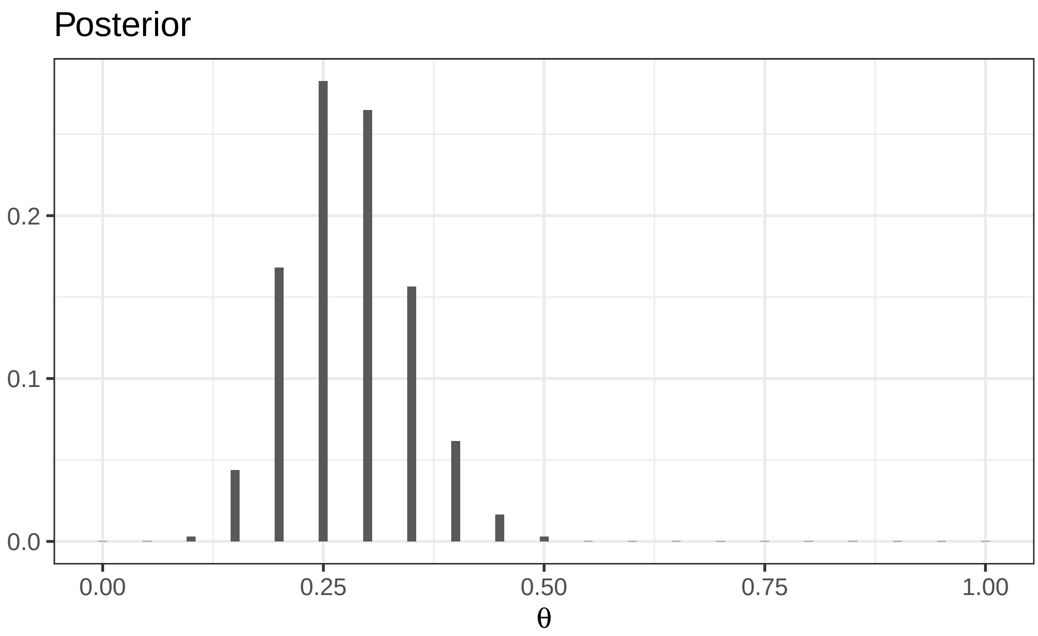
3.5.4 Influence of Sample Size

If, instead, we have $N = 40$ and $z = 10$, the posterior will be more similar to the likelihood.

```
grid_df2 <- grid_df %>%
    mutate(
        # Use our previously defined lik() function
        py_th = lik(th, num_flips = 40, num_heads = 10),
        # Product of prior and likelihood
        `prior x lik` = pth * py_th,
        # Scaled the posterior
        pth_y = `prior x lik` / sum(`prior x lik`)
    )
# Plot the prior/likelihood and the posterior
ggplot(data = grid_df2, aes(x = th)) +
    geom_col(aes(x = th - 0.005, y = pth, fill = "prior"),
        width = 0.01,
    ) +
    geom_col(aes(x = th + 0.005, y = py_th / sum(py_th),
        fill = "scaled likelihood"), width = 0.01,
    ) +
    labs(fill = NULL, y = NULL, x = expression(theta)) +
    theme(legend.position = "top")
```



```
ggplot(data = grid_df2, aes(x = th)) +
    geom_col(aes(x = th, y = pth_y), width = 0.01) +
    labs(
        fill = NULL, y = NULL, title = "Posterior",
        x = expression(theta)
    )
```



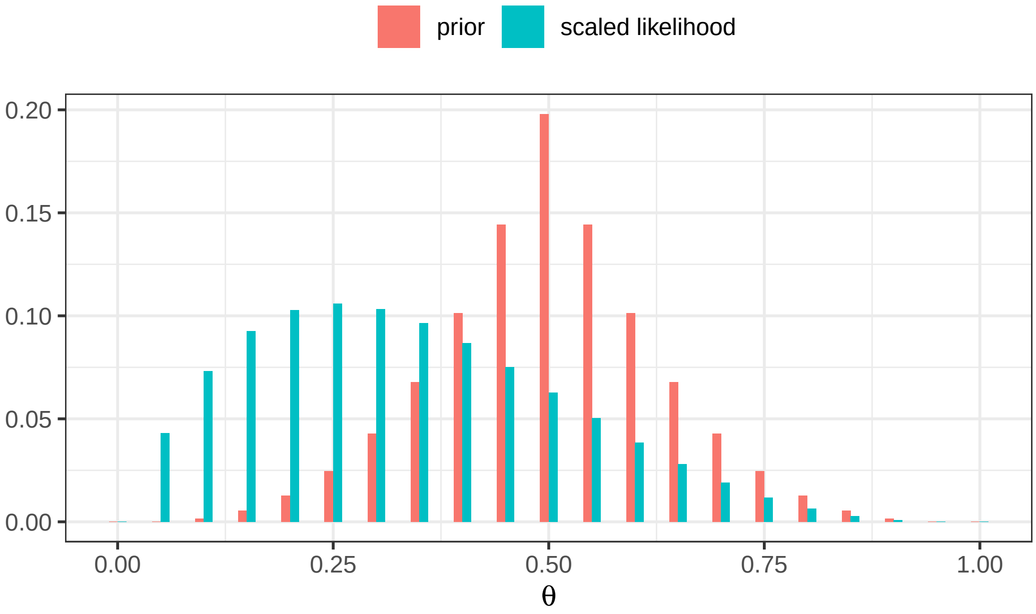
```
# Sample from the posterior
post_samples <- sample(
    grid_df2$th,
    size = 1000, replace = TRUE,
    prob = grid_df2$pth_y
)
summ_draw(post_samples)
```

```
#>       mean     median         sd        mad       ci.1       ci.9
#> 0.28085000 0.30000000 0.06542215 0.07413000 0.20000000 0.35000000
```

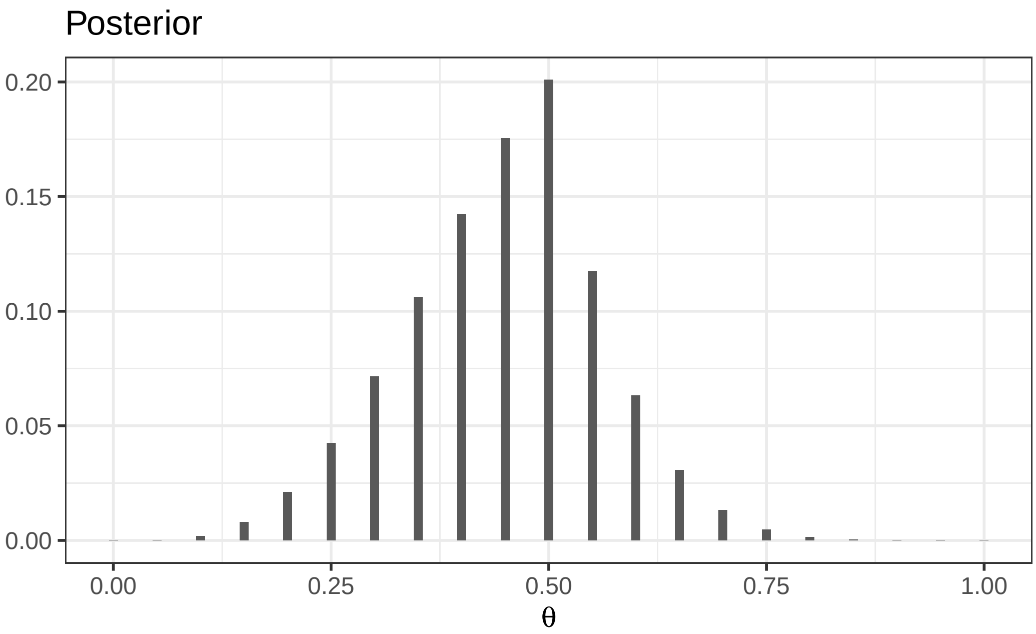
3.5.5 Influence of Prior

If we have a very strong prior concentrated at $\theta = .5$, but still with $N = 40$ and $z = 10$, the posterior will be more similar to the prior.

```
grid_df3 <- grid_df %>%
    mutate(
        # stronger prior
        pth = pth ^ 3,
        # scale the prior to sume to 1
        pth = pth / sum(pth),
        # Use our previously defined lik() function
        py_th = lik(th, num_flips = 4, num_heads = 1),
        # Product of prior and likelihood
        `prior x lik` = pth * py_th,
        # Scaled the posterior
        pth_y = `prior x lik` / sum(`prior x lik`)
    )
# Plot the prior/likelihood and the posterior
ggplot(data = grid_df3, aes(x = th)) +
    geom_col(aes(x = th - 0.005, y = pth, fill = "prior"),
        width = 0.01,
    ) +
    geom_col(aes(x = th + 0.005, y = py_th / sum(py_th),
        fill = "scaled likelihood"), width = 0.01,
    ) +
    labs(fill = NULL, y = NULL, x = expression(theta)) +
    theme(legend.position = "top")
```



```
ggplot(data = grid_df3, aes(x = th)) +
    geom_col(aes(x = th, y = pth_y), width = 0.01) +
    labs(
        fill = NULL, y = NULL, title = "Posterior",
        x = expression(theta)
    )
```



```
# Sample from the posterior
post_samples <- sample(
    grid_df3$th,
    size = 1000, replace = TRUE,
    prob = grid_df3$pth_y
)
summ_draw(post_samples)
```

```
#>      mean    median        sd       mad      ci.1      ci.9
#> 0.4493000 0.4500000 0.1096656 0.0741300 0.3000000 0.6000000
```

```
# Alternatively, use the `posterior` package
data.frame(theta = post_samples) |>
    posterior::summarize_draws()
```

```
#> # A tibble: 1 x 10
#>   variable  mean median    sd    mad    q5   q95  rhat ess_bulk ess_tail
#>   <chr>    <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
#> 1 theta    0.449   0.45 0.110 0.0741  0.25   0.6  1.00    1001.     899.
```

3.5.6 Remark on Grid Approximation

In this note, we discretized θ into a finite number of grid points to compute the posterior, mainly for pedagogical purposes. A big limitation is that our posterior will have no density for values other than the chosen grid points. While increasing the number of grid points (e.g., 1,000) can give more precision, the result is still not truly continuous. A bigger issue is that the computation breaks down when there is more than one parameter; if there are p parameters, with 1,000 grid points per parameter, one needs to evaluate the posterior probability for $1,000^p$ grid points, which is not feasible even with modern computers. So more efficient algorithms, namely Markov chain Monte Carlo (MCMC) methods, will be introduced as we progress in the course.

References

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE.

McGrayne, S. B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale university press.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100