Bayesian Data Analysis
Thinking Probabilistically

Min-Te Sun, Ph.D.

1

Reallocation of Credibility Across Possibilities

- On a typical day at your location, what is the probability that it is cloudy? Suppose you are told it is raining, now what is the probability that it is cloudy?
  - p(cloudy) < p(cloudy|raining)
- Suppose instead you are told that everyone outside is wearing sunglasses.
  - p(cloudy) > p(cloudy|sunglasses)
- Bayes' rule is merely the mathematical relation between the prior allocation of credibility and the posterior reallocation of credibility conditional on data.

2

Statistics

- Statistics – collecting, organizing, analyzing, and interpreting data.
- Two main statistical methods are used in data analysis:
  - **Exploratory Data Analysis** (**EDA**): This is about numerical summaries, such as the mean, mode, standard deviation, and interquartile ranges (this part of EDA is also know as **descriptive statistics**). EDA is also about visually inspecting the data, using tools you may be already familiar with, such as histograms and scatter plots.
  - **Inferential statistics**: This is about making statements beyond the current data. We may want to understand some particular phenomenon, or maybe we want to make predictions for future (as yet unobserved) data points, or we need to choose among several competing explanations for the same observations. Inferential statistics is the set of methods and tools that will help us to answer these types of questions.

3

Statistic Approach is Better Because

- **Ontological –** Statistics is a form of modeling unified under the mathematical framework of probability theory. Using a probabilistic approach provides a unified view of what may seem like very disparate methods; statistical methods and **machine learning** (**ML**) methods look much more similar under the probabilistic lens.
- **Technical –** Modern software, such as PyMC3, allows practitioners, just like you and me, to define and solve models in a relative easy way. Many of these models were unsolvable just a few years ago or required a high level of mathematical and technical sophistication.

4

Working with Data

- Data is an essential ingredient in statistics and data science.
- Data comes from several sources, such as experiments, computer simulations, surveys, and field observations.
- There is a whole branch of statistics dealing with data collection, known as **experimental design**.
  - Sometimes gathering data is not cheap. For example, while it is true that the **Large Hadron Collider** (**LHC**) produces hundreds of terabytes a day, its construction took years of manual and intellectual labor.

5

Data Uncertainty

- As a general rule, we can think of the process generating the data as stochastic, because there is ontological, technical, and/or epistemic uncertainty, that is, the system is intrinsically stochastic, there are technical issues adding noise or restricting us from measuring with arbitrary precision, and/or there are conceptual limitations veiling details from us.
- We always need to interpret data in the context of models, including mental and formal ones. Data does not speak but through models.

6

Assumption on Data Collection

- In this class, we will assume that
  - we already have collected the data.
  - Our data will also be clean and tidy, something rarely true in the *real world*.

7

Sample Space

- Definition: a set of possible outcomes
  - For coin flips, the sample space is head and tail
- If coin is "fair" when the probability of coming up heads (denoted as θ) is 0.5
- "Degree of Belief" about a parameter is denoted as $p(\theta)$
  - We might believe that $p(\theta = 0.5) = 0.99$ if the coin was minted by the federal government
- Both "probability" of head or tail outcome and "degree of belief " in biases refer to sample spaces

8

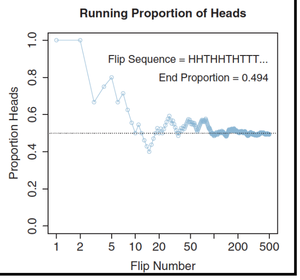
"Outside" or "Inside" the Head

- Probability of head in coin flips is "outside" the head.
  - Can be "sampled"
- Belief about the fairness of a coin is "inside" the head.
  - Can not be "sampled"
- Nevertheless, the mathematical properties of probabilities are the same in their essentials.

9

Approximating the Probability by the Long-Run Relative Frequency

- For a fair coin, it seems obvious that we should get about 50% heads in any long sequence of flips. However…
  - How many times of flips is adequate?
  - The proportion of heads is near 0.5 but not necessarily exactly equal to 0.5 at the end of the long sequence.



10

Deriving a Long-Run Relative Frequency

- The sample space of the coin consists of two possible outcomes, head and tail. By the assumption of fairness, we know that each outcome is equally likely. Therefore, the long-run relative frequency of heads should be exactly ½.
- This can be extended to other simple situations, such as a fair dice.

11

Subjective Belief

- How strongly do you believe that a coin minted by the US government is fair?
- How about a coin purchased at a magic shop?
- To specify our subjective beliefs, we have to specify how likely we think each possible outcome is.
  - hard to pin down mushy intuitive beliefs
  - How to "calibrate" subjective beliefs?

12

Simple Examples to Quantify "Belief"

- How strongly do you believe that there will be a snowstorm that closes the interstate highways near Indianapolis next New Year's Day?
- In a sack we put 10 marbles: 5 red, and 5 white. We shake the sack and then draw a marble at random.
  - The probability of getting a red marble is, of course, 5/10 = 0.5.

13

Which Do You Prefer?

- Set 1
  - Gamble A: You get $100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
  - Gamble B: You get $100 if you draw a red marble from a sack of marbles with 5 red and 5 white marbles.
- Set 2
  - Gamble A: You get $100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
  - Gamble C: You get $100 if you draw a red marble from a sack of marbles with 1 red and 9 white marbles.

14

Describing a Subjective Belief Mathematically

- We might believe that the average American woman is 5′4″ tall.
  - Too many possible outcomes in sample space!
  - We might instead describe your degree of belief by a bell-shaped curve that is highest at 5′4″ and drops off symmetrically above and below that most-likely height.
  - Change the width and center of the curve until it seems to best capture your subjective belief
  - This is normal (or Gaussian) distribution!

15

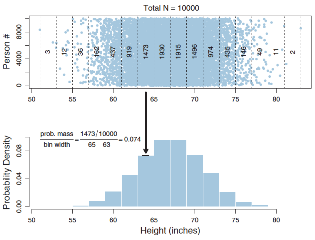
Probabilities Assign Numbers to Possibilities

- A probability, whether it's outside the head or inside the head, is just a way of assigning numbers to a set of mutually exclusive possibilities.
- Probabilities need to satisfy three properties
  1. A probability value must be nonnegative (i.e., zero or positive).
  2. The sum of the probabilities across all events in the entire sample space must be 1.0.
  3. For any two mutually exclusive events, the probability that one *or* the other occurs is the *sum* of their individual probabilities.

16

Discrete Probability Distributions: Probability Mass

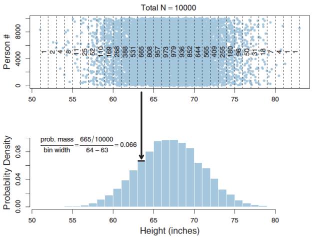
- For continuous outcome spaces, we can *discretize* the space into a finite set of mutually exclusive and exhaustive "bins."
- In particular, the # of measurements that fall within the interval 63″ to 65″ is 1,473, which means that the estimated probability of falling in that interval is 1473/10000 = 0.1473.



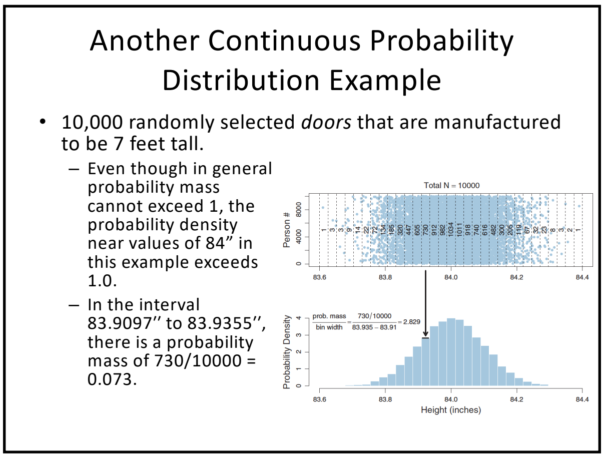
17

Continuous Probability Distributions: Probability Density

- A continuous outcome space (like a person's height)
  - It becomes problematic to talk about the probability of a specific value on continuum.
  - We can, however, talk about the probability mass of intervals.



18

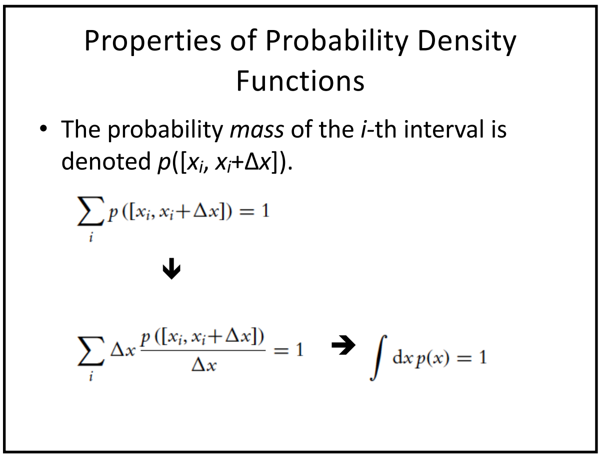


Another Continuous Probability Distribution Example

- 10,000 randomly selected *doors* that are manufactured to be 7 feet tall.
  - Even though in general probability mass cannot exceed 1, the probability density near values of 84” in this example exceeds 1.0.
  - In the interval 83.9097′′ to 83.9355′′, there is a probability mass of 730/10000 = 0.073.

Total N = 10000
Person #
0
4000
8000
1
3
3
9
14
22
72
134
185
320
447
605
730
912
982
1034
1011
918
740
616
482
300
205
119
67
32
23
8
3
2
1
83.6
83.8
84.0
84.2
84.4
prob. mass / bin width = 730/10000 / 83.935 − 83.91 = 2.829
Probability Density
0
1
2
3
4
83.6
83.8
84.0
84.2
84.4
Height (inches)

19



Properties of Probability Density Functions

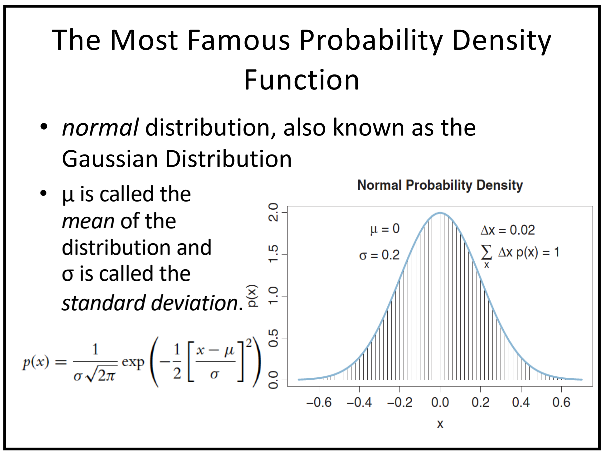
- The probability *mass* of the *i*-th interval is denoted *p*([*x*_i, *x*_i+Δ*x*]).

$$\sum_i p\left([x_i, x_i{+}\Delta x]\right) = 1$$

$$\Downarrow$$

$$\sum_i \Delta x \, \frac{p\left([x_i, x_i{+}\Delta x]\right)}{\Delta x} = 1 \quad \Rightarrow \quad \int \mathrm{d}x \, p(x) = 1$$

20



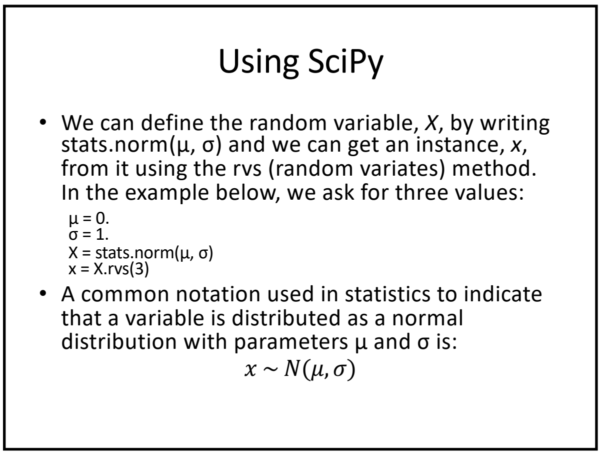
The Most Famous Probability Density Function

- *normal* distribution, also known as the Gaussian Distribution
- μ is called the *mean* of the distribution and σ is called the *standard deviation*.

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

Normal Probability Density
μ = 0
σ = 0.2
Δx = 0.02
∑ₓ Δx p(x) = 1
p(x)
0.0
0.5
1.0
1.5
2.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
x

21



Using SciPy

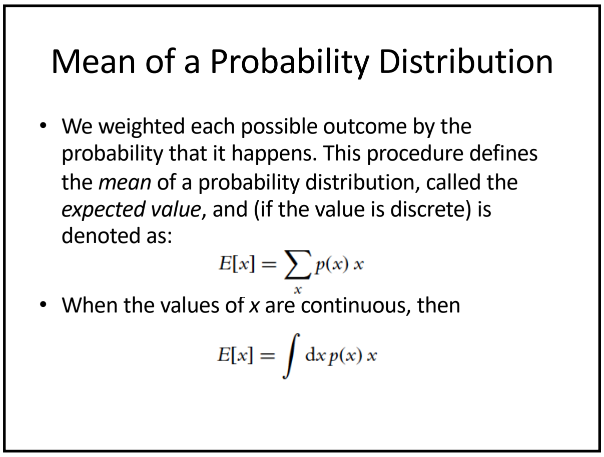
- We can define the random variable, *X*, by writing stats.norm(μ, σ) and we can get an instance, *x*, from it using the rvs (random variates) method. In the example below, we ask for three values:

```
μ = 0.
σ = 1.
X = stats.norm(μ, σ)
x = X.rvs(3)
```

- A common notation used in statistics to indicate that a variable is distributed as a normal distribution with parameters μ and σ is:

$$x \sim N(\mu, \sigma)$$

22



Mean of a Probability Distribution

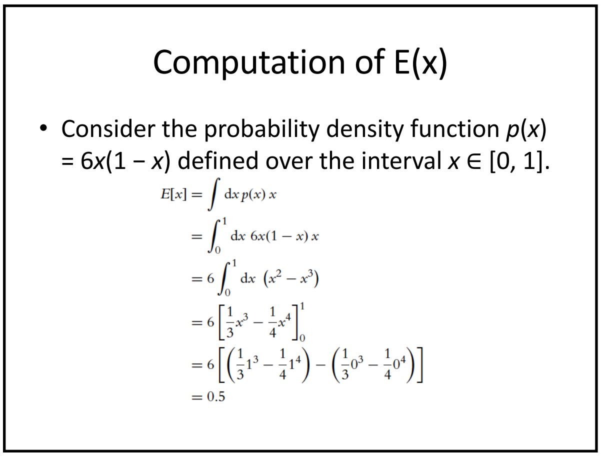
- We weighted each possible outcome by the probability that it happens. This procedure defines the *mean* of a probability distribution, called the *expected value*, and (if the value is discrete) is denoted as:

$$E[x] = \sum_x p(x)\, x$$

- When the values of *x* are continuous, then

$$E[x] = \int \mathrm{d}x \, p(x)\, x$$

23



Computation of E(x)

- Consider the probability density function *p*(*x*) = 6*x*(1 − *x*) defined over the interval *x* ∈ [0, 1].

$$\begin{aligned}
E[x] &= \int \mathrm{d}x \, p(x)\, x \\
&= \int_0^1 \mathrm{d}x \;\, 6x(1-x)\, x \\
&= 6 \int_0^1 \mathrm{d}x \;\, \left(x^2 - x^3\right) \\
&= 6 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_0^1 \\
&= 6 \left[\left(\frac{1}{3}1^3 - \frac{1}{4}1^4\right) - \left(\frac{1}{3}0^3 - \frac{1}{4}0^4\right)\right] \\
&= 0.5
\end{aligned}$$

24

Variance of a Probability Distribution

- The definition for "variance" is based on the squared difference between *x* and the mean, i.e., the mean squared deviation (MSD) of the *x* values from their mean.

$$\mathrm{var}_x = \int \mathrm{d}x\, p(x)\ (x - E[x])^2$$

- Please figure out the discrete version of variance by yourself.
- In a normal distribution, about 34% of the distribution lies between μ and μ+σ, but not to overgeneralize this to distributions with other shapes!

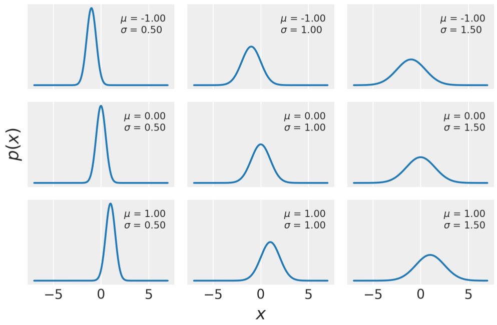
25

Thoughts on Mean and Standard Deviation

- The probability can be interpreted either as how much a value could be sampled from a generative process, or as how much credibility the value has relative to other values.
- When *p*(θ) represents credibility values of θ, instead of the probability of sampling θ, then the mean of *p*(θ) can be thought of as a value of θ that represents a typical credible value.
- The standard deviation of θ, which measures how wide the distribution is, can be thought of as a measure of uncertainty across candidate values.
  - If the standard deviation is small, then we believe strongly in values of θ near the mean.
  - If the standard deviation is large, then we are not very certain about what value of θ to believe in.

26

Gaussian Distributions with Different μ and σ



27

Two-Way Distributions

- The conjunction of two outcomes
  - EX: the probability of being dealt a card that is both a queen *and* a heart
  - Below is a more complicated example

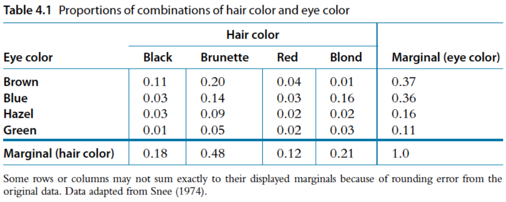


Table 4.1 Proportions of combinations of hair color and eye color

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Marginal (eye color)**
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

Some rows or columns may not sum exactly to their displayed marginals because of rounding error from the original data. Data adapted from Snee (1974).

28

Joint and Marginal Probability

- In each cell, the table indicates the *joint probability* of particular combinations of eye color and hair color.
- We may be interested in the probabilities of the eye colors overall, collapsed across hair colors. These probabilities are indicated in the right margin of the table, and they are therefore called *marginal* probabilities.

29

Marginal Probabilities

- If attributes are discrete:

$$p(e) = \textstyle\sum_h p(e,h)$$
$$p(h) = \textstyle\sum_e p(e,h)$$

- If attributes are continuous:

$$p(r) = \int \mathrm{d}c\, p(r,c)$$
$$p(c) = \int \mathrm{d}r\, p(r,c)$$

30

Conditional Probability

- The probability of one outcome, given that we know another outcome is true.
- Conditional on that the person has blue eyes, what is the probability that the person has blond hair?
  - The total (i.e., marginal) amount of blue-eyed people is 0.36, and that 0.16 of the population has blue eyes and blond hair. Therefore, of the 0.36 with blue eyes, the fraction 0.16/0.36 = 45% has blond hair.
- The general probability of having blond hair is 0.21, but when we learn that a person from this group has blue eyes, then the credibility of that person having blond hair increases to 0.45. This reallocation of credibility across the possible hair colors *is* Bayesian inference!

31

Computation of Conditional Probability

- Notation: $p(h|e)$

$$p(h|e) = p(e,h)/p(e)$$

$$\Downarrow$$

$$p(h|e) = p(e,h)/p(e) = p(e,h)/\sum_h p(e,h)$$

$$\Downarrow$$

$$p(h|e) = p(e,h)/p(e) = p(e,h)/\sum_{h^*} p(e,h^*)$$

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal (eye color)
Blue	0.03/0.36 = 0.08	0.14/0.36 = 0.39	0.03/0.36 = 0.08	0.16/0.36 = 0.45	0.36/0.36 = 1.0

32

Conditional Probability

- When attributes are discrete:

$$p(c|r) = \frac{p(r,c)}{\sum_{c^*} p(r,c^*)} = \frac{p(r,c)}{p(r)}$$

- When attributes are continuous

$$p(c|r) = \frac{p(r,c)}{\int \mathrm{d}c\, p(r,c)} = \frac{p(r,c)}{p(r)}$$

33

Thoughts on Conditional Probability

- It is important to recognize that, in general, $p(r|c)$ is *not* equal to $p(c|r)$.
- When we say "the probability of *x* given *y*" we do *not* mean that *y* has already happened and *x* has yet to happen.
  - All we mean is that we are restricting our calculations of probability to a particular subset of possible outcomes.
  - A better gloss of $p(x|y)$ is, "among all joint outcomes with value *y*, this proportion of them also has value *x*."

34

Another Example
Disease Diagnose (1/3)

- Consider trying to diagnose a rare disease. Suppose that in the general population, the probability of having the disease is only one in a thousand. We denote the true presence or absence of the disease as the value of a parameter, θ, that can have the value θ = ☹ if disease is present in a person, or the value θ = ☺ if the disease is absent. The base rate of the disease is therefore denoted *p*(θ = ☹) = 0.001. This is our prior belief that a person selected at random has the disease.

35

Disease Diagnose (2/3)

- Suppose that there is a test for the disease that has a 99% hit rate, which means that if a person has the disease, then the test result is positive 99% of the time. We denote a positive test result as *T* = +, and a negative test result as *T* = −. The observed test result is the datum that we will use to modify our belief about the value of the underlying disease parameter. The hit rate is expressed formally as *p*(*T* = +|θ = ☹) = 0.99. Suppose also that the test has a false alarm rate of 5%. This means that 5% of the time when the disease is absent, the test falsely indicates that the disease is present. We denote the false alarm rate as *p*(*T* = +|θ = ☺) = 0.05.

36

Disease Diagnose (3/3) Question for You

- Suppose we sample a person at random from the population, administer the test, and it comes up positive. What is the posterior probability that the person has the disease?
- A more formal way (i.e., mathematically) to ask this question: what is *p*(θ = ☹| *T* = +)?

37

Calculation (1/2)

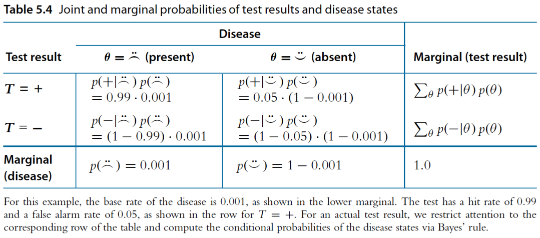


Table 5.4 Joint and marginal probabilities of test results and disease states

	Disease		
Test result	**θ = ☹ (present)**	**θ = ☺ (absent)**	**Marginal (test result)**
$T = +$	$p(+|$☹$)\,p($☹$)$ $= 0.99 \cdot 0.001$	$p(+|$☺$)\,p($☺$)$ $= 0.05 \cdot (1 - 0.001)$	$\sum_\theta p(+|\theta)\,p(\theta)$
$T = -$	$p(-|$☹$)\,p($☹$)$ $= (1 - 0.99) \cdot 0.001$	$p(-|$☺$)\,p($☺$)$ $= (1 - 0.05) \cdot (1 - 0.001)$	$\sum_\theta p(-|\theta)\,p(\theta)$
Marginal (disease)	$p($☹$) = 0.001$	$p($☺$) = 1 - 0.001$	1.0

For this example, the base rate of the disease is 0.001, as shown in the lower marginal. The test has a hit rate of 0.99 and a false alarm rate of 0.05, as shown in the row for $T = +$. For an actual test result, we restrict attention to the corresponding row of the table and compute the conditional probabilities of the disease states via Bayes' rule.

38

Calculation (2/2)

- Then, we have

$$p(\theta = ☹ | T = +) = \frac{p(T = + | \theta = ☹)\, p(\theta = ☹)}{\sum_\theta p(T = + | \theta)\, p(\theta)}$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot (1 - 0.001)}$$

$$= 0.019$$

39

Key Application of Bayes' Rule

- When the row variable represents data values and the column variable represents parameter values
  - A model of data specifies the probability of particular data values given the model's structure and parameter values.
- In other words, a model specifies:

$$p(\text{data values} \mid \text{parameters values})$$
$$\text{along with the prior, } p(\text{parameters values})$$

- We then use Bayes' rule to convert that to:

$$p(\text{parameters values} \mid \text{data values})$$

40

Bayes' Way to Look at Table

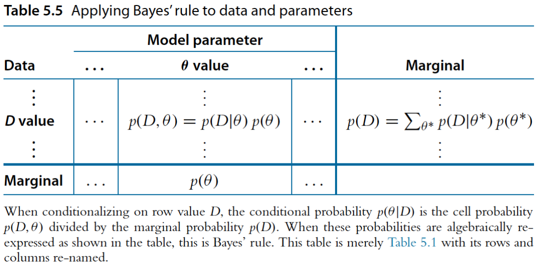


Table 5.5 Applying Bayes' rule to data and parameters

	Model parameter			
Data	**. . .**	**θ value**	**. . .**	**Marginal**
⋮ ***D* value** ⋮	$\cdots$	⋮ $p(D,\theta) = p(D|\theta)\,p(\theta)$ ⋮	$\cdots$	⋮ $p(D) = \sum_{\theta*} p(D|\theta^*)\,p(\theta^*)$ ⋮
Marginal	. . .	$p(\theta)$	. . .	

When conditionalizing on row value D, the conditional probability $p(\theta|D)$ is the cell probability $p(D,\theta)$ divided by the marginal probability $p(D)$. When these probabilities are algebraically re-expressed as shown in the table, this is Bayes' rule. This table is merely Table 5.1 with its rows and columns re-named.

41

Names for Factors of Bayes' Rule

- Factors of Bayes' Rule:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(D|\theta)}_{\text{likelihood}} \; \underbrace{p(\theta)}_{\text{prior}} \,/\, \underbrace{p(D)}_{\text{evidence}}$$

- Denominator (i.e., *evidence*, also known as *marginal likelihood*):

$$p(D) = \sum_{\theta*} p(D|\theta^*)p(\theta^*)$$

42

Bayes' Rule for Continuous Variables

- Only change is that the marginal likelihood changes from the sum in Equation 5.8 to an integral:

$$p(D) = \int d\theta^* p(D|\theta^*)p(\theta^*)$$

43

Independence of Attributes

- In general, when the value of *y* has no influence on the value of *x*, we know that the probability of *x* given *y* simply is the probability of *x* in general.
  - *p*(*x*|*y*) = *p*(*x*) for all values of *x* and *y*.
  - Because *p*(*x*|*y*) = *p*(*x*, *y*)/*p*(*y*), we have *p*(*x*, *y*) = *p*(*x*)*p*(*y*) for all values of *x* and *y* when *x* and *y* are independent.
  - Are the eye color and the hair color independent?

44

Independently and Identically Distributed Variables

- Many models assume that successive values of random variables are all sampled from the same distribution and those values are independent of each other. In such a case, we will say that the variables are **independently and identically distributed** (**iid**) variables for short.

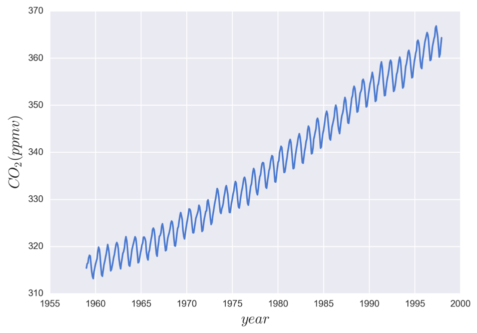
45

Non-iid Variables

- A common example of non-iid variables are **temporal series**, where a temporal dependency in the random variable is a key feature that should be taken into account. Take, for example, the data coming from http://cdiac.esd.ornl.gov. This data is a record of atmospheric CO_2 measurements from 1959 to 1997.

46

Increasing Concentration of Atmospheric CO_2



47

Bayesian Modeling

- Models are simplified descriptions of a given system or process that, for some reason, we are interested in.
  - Those descriptions are deliberately designed to capture only the most relevant aspects of the system and not to explain every minor detail. This is one reason a more complex model is not always a better one.
- There are many different kinds of models.
  - In this class, we will restrict ourselves to Bayesian models.

48

Bayesian Modeling Process

1. Given some data and some assumptions on how this data could have been generated, we design a model by combining building blocks known as **probability distributions**. Most of the time these models are crude approximations, but most of the time is all we need.
2. We use Bayes' theorem to add data to our models and derive the logical consequences of combining the data and our assumptions. We say we are conditioning the model on our data.
3. We criticize the model by checking whether the model makes sense according to different criteria, including the data, our expertise on the subject, and sometimes by comparing several models.

- In general, we will find ourselves performing these three steps in an iterative non-linear fashion.

49

Probabilistic Models

- Bayesian models are also known as **probabilistic models** because they are built using probabilities.
- Why probabilities?
  - Because probabilities are the correct mathematical tool to model uncertainty, so let's take a walk through the garden of forking paths.

50

Prior Distribution

- The prior distribution should reflect what we know about the value of the parameter before seeing the data, . If we know nothing, like Jon Snow, we could use flat priors that do not convey too much information. In general, we can do better than flat priors, as we will learn in this book. The use of priors is why some people still talk about Bayesian statistics as subjective, even when priors are just another assumption that we made when modeling and hence are just as subjective (or objective) as any other assumption, such as likelihoods.

51

Likelihood

- The likelihood is how we will introduce data in our analysis. It is an expression of the plausibility of the data given the parameters. In some texts, you will find people call this term sampling model, statistical model, or just model. We will stick to the name likelihood and we will call the combination of priors and likelihood model.

52

Posterior Distribution

- The posterior distribution is the result of the Bayesian analysis and reflects all that we know about a problem (given our data and model). The posterior is a probability distribution for the θ parameters in our model (not a single value). This distribution is a balance between the prior and the likelihood. There is a well-known joke: A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule. Anyway, this joke captures the idea of a posterior being somehow a compromise between prior and likelihood. Conceptually, we can think of the posterior as the updated prior in the light of (new) data. In fact, the posterior from one analysis can be used as the prior for a new analysis. This makes Bayesian analysis particularly suitable for analyzing data that becomes available in sequential order.

53

Marginal Likelihood

- The last term is the **marginal likelihood**, also known as **evidence**. Formally, the marginal likelihood is the probability of observing the data averaged over all the possible values the parameters can take (as prescribed by the prior). Anyway, for most of this class, we will not care about the marginal likelihood, and we will think of it as a simple normalization factor. We can do this because when analyzing the posterior distribution, we will only care about the relative values of the parameters and not their absolute values.

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

54

Ordering of Data

- Bayes’ rule gets us from a prior belief, $p(\theta)$, to a posterior belief, $p(\theta|D)$, when we take into account some data D.
- Now suppose we observe some *more* data, which we’ll denote D'. We can then update our beliefs again, from $p(\theta|D)$ to $p(\theta|D',D)$.
- Question: Does our final belief depend on whether we update with D first and D' second, or update with D' first and D second?

55

Answer

- It depends!
  - It depends on the model function that defines the likelihood, $p(D|\theta)$.
- In many models, the probability of data, $p(D|\theta)$, does not depend in any way on *other* data.
  - That is, the joint probability $p(D,D'|\theta)$ equals $p(D|\theta)\times p(D'|\theta)$.
  - In other words, the data probabilities are *independent.*

56

Reason

- Intuition: If the likelihood function has no dependence on data ordering, then the posterior should not have any dependence on data ordering.
- Formally: We simply write down Bayes’ rule and use the independence assumption that $p(D',D|\theta) = p(D'|\theta)\times p(D|\theta)$:

$$\begin{aligned}
p(\theta|D',D) &= \frac{p(D',D|\theta)\,p(\theta)}{\sum_{\theta^*} p(D',D|\theta^*)\,p(\theta^*)} && \text{Bayes' rule} \\
&= \frac{p(D'|\theta)p(D|\theta)\,p(\theta)}{\sum_{\theta^*} p(D'|\theta^*)p(D|\theta^*)\,p(\theta^*)} && \text{by assumption of independence} \\
&= \frac{p(D|\theta)p(D'|\theta)\,p(\theta)}{\sum_{\theta^*} p(D|\theta^*)p(D'|\theta^*)\,p(\theta^*)} && \text{multiplication is commutative} \\
&= p(\theta|D,D') && \text{Bayes' rule}
\end{aligned}$$

57

Single-Parameter Inference
The Coin Flipping Problem

- The coin-flipping problem, or the beta-binomial model, is a classical problem in statistics and goes like this: we toss a coin a number of times and record how many heads and tails we get. Based on this data, we try to answer questions such as, *is the coin fair?* Or, more generally, *how biased is the coin?*
- Note about terminology: When we refer to the “bias” in a coin, we will be referring to its underlying probability of coming up heads.
  - *When a coin is fair, it has a “bias” of 0.50.*
  - Other times, the term “bias” may be in its colloquial sense of a *departure from fairness*, as in “head biased” or “tail biased.”

58

Why Care about Coin Flips?

- Because coin flips are a surrogate for myriad other real-life (dichotomous) events that we do care about
  - For a given type of heart surgery, we may want to know what is the probability that patients survive more than one year.
  - For a survey question, we may want to know the probability of agree or disagree.
  - In a two-candidate election (A and B), before the election itself we want to estimate, from a poll, we may want to estimate the probability that candidate A will win.
  - Studying arithmetic ability by measuring accuracy on a multi-item exam, for which the item outcomes are correct or wrong.

59

General Model

- The first thing we will do is generalize the concept of bias. We will say that a coin with a bias of 1 will always land heads, one with a bias of 0 will always land tails, and one with a bias of 0.5 will land half of the time heads and half of the time tails. To represent the bias, we will use the θ parameter, and to represent the total number of heads for a N number of tosses, we will use the y variable. According to Bayes' theorem, we have to specify the prior, $p(\theta)$, and likelihood, $p(y|\theta)$.

60

Choosing The Likelihood

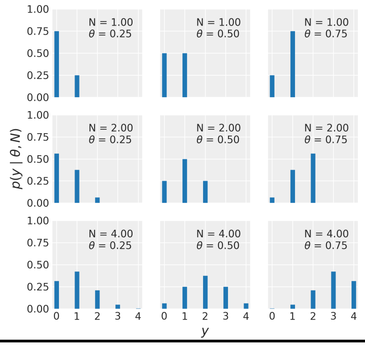
- Let's assume that only two outcomes are possible—heads or tails—and let's also assume that a coin toss does not affect other tosses, that is, we are assuming coin tosses are independent of each other. We will further assume all coin tosses come from the same distribution. Thus the random variable *coin toss* is an example of an **iid** variable. I hope you agree these are very reasonable assumptions to make for our problem. Given these assumptions a good candidate for the likelihood is the binomial distribution:

$$p(y|\theta, N) = \frac{N!}{y!\,(N-y)!}\theta^{y}(1-\theta)^{N-y}$$

- This is a discrete distribution returning the probability of getting heads (or in general, successes) out of *N* coin tosses (or in general, trials or experiments) given a fixed value of θ.

61

$p(y|\theta, N)$ Given Different values of θ and N



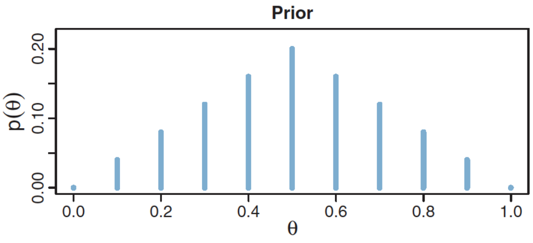
62

Binomial Distributions

- The binomial distribution is a reasonable choice for the likelihood. We can see that indicates how likely it is to obtain a head when tossing a coin (this is easier to see when *N* = 1, but is valid for any value of *N*)—just compare the value of θ with the height of the bar for *y* = 1 (heads).
- If we know the value of θ, the binomial distribution will tell us the expected distribution of heads. The only problem is that we do not know θ! But do not despair; in Bayesian statistics, every time we do not know the value of a parameter, we put a prior on it.

63

Using Computer for BDA Given Prior Distribution Below



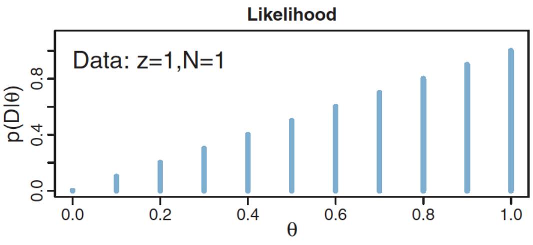
64

Collect Data and Applying Bayes' Rule

- Suppose that we flip the coin once and observe heads => *y* = 1 or, equivalently, *z* = 1 with *N* = 1
- *p*(*D*|θ) = θ

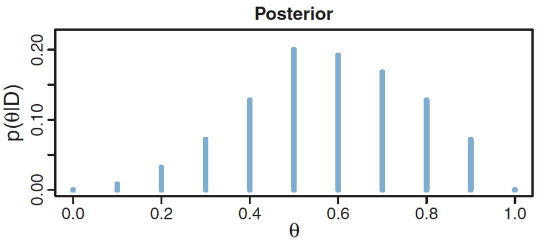
65

Likelihood of θ



66

Posterior of θ



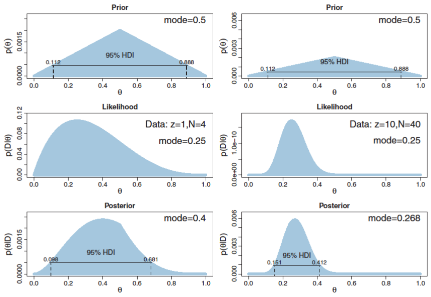
67

How Posterior is Obtained?

- From slide #21, we see that for these data the likelihood function becomes $p(D|\theta) = \theta$.
- At each candidate value of θ, the posterior probability is computed from Bayes' rule (slide #13) as the likelihood at θ times the prior at θ divided by $p(D)$.
  - For example, consider θ = 0.2, scanning vertically across panels. In the lower panel at θ = 0.2, the posterior probability is the likelihood from the middle panel at θ = 0.2 times the prior in the upper panel at θ = 0.2, divided by the sum, $\sum_{\theta^*} p(D|\theta^*)p(\theta^*)$.
- Notice that the overall contour of the posterior distribution is different from the prior distribution.

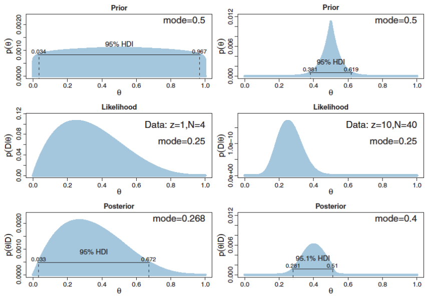
68

Influence of Number of Samples



69

Influence of Prior



70

Derivation of Posterior Use Beta Distribution as Prior $p(\theta)$

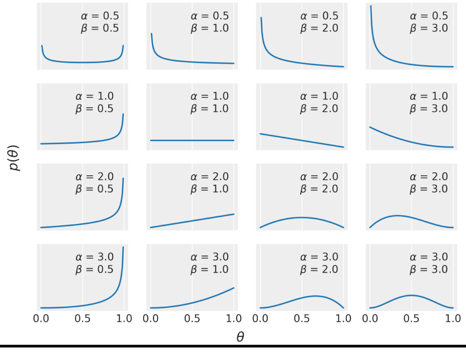
- As a prior, we will use a **beta distribution**, which is a very common distribution in Bayesian statistics and looks as follows:

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

- If we look carefully, we will see that the beta distribution looks similar to the binomial except for the first term, the one with all those Γ. The first term is a normalizing constant that ensures the distribution integrates to 1, and Γ represents gamma function. We can see from the preceding formula that the beta distribution has two parameters, α and β.

71

Beta Distributions Given Different Values of α and β



72

Conclusions After Observation

- As *a* gets bigger, the bulk of the distribution moves rightward over higher values of θ, but as *b* gets bigger, the bulk of the distribution moves leftward over lower values of θ.
- As *a* and *b* get bigger together, the beta distribution gets narrower.
- The variables *a* and *b* are called the *shape parameters* of the beta distribution because they determine its shape.
  - The shape parameters can have any positive real value.

73

We Use Beta Distribution for Prior of θ Because

1. Beta distribution is restricted to be between 0 and 1, in the same way our θ parameter is. In general, we use the beta distribution when we want to model proportions of a binomial variable.
2. Its versatility. As we can see, the distribution adopts several shapes (all restricted to the [0, 1] interval), including a uniform distribution, Gaussian-like distributions, and U-like distributions.
3. Beta distribution is the **conjugate prior** of the binomial distribution.

74

Conjugate Prior

- A conjugate prior of a likelihood is a prior that, when used in combination with a given likelihood, returns a posterior with the same functional form as the prior. Untwisting the tongue, every time we use a beta distribution as the prior and a binomial distribution as the likelihood, we will get a beta as the posterior distribution.
  - There are other pairs of conjugate priors; for example, the Normal distribution is the conjugate prior of itself.
- For many years, Bayesian analysis was restricted to the use of conjugate priors. Conjugacy ensures mathematical tractability of the posterior, which is important given that a common problem in Bayesian statistics is ending up with a posterior we cannot solve analytically. This was a deal breaker before the development of suitable computational methods to solve probabilistic methods.

75

Specify Prior Beliefs Using Beta Distribution

- We can think of *a* and *b* in the prior as if they were previously observed data, in which there were *a* heads and *b* tails in a total of *n* = *a* + *b* flips.
  - If we have no prior knowledge other than the knowledge that the coin has a head side and a tail side, that's tantamount to having previously observed one head and one tail, which corresponds to *a* = 1 and *b* = 1. (uniform)
  - If we think that the coin is probably fair but we're not very sure, then we can imagine that the previously observed data had, say, *a* = 4 heads and *b* = 4 tails.

76

Setting *a* and *b* in Beta Distribution

- Often we think of our prior beliefs in terms of a central tendency and certainty about that central tendency.
  - In thinking about the probability of left handedness in the general population of people, we might think from everyday experience that it's around 10%. But if we are not very certain about that value, we might consider the equivalent previous sample size to be small, say, *n* = 10, which means that of 10 previously observed people, 1 of them was left handed.
  - In thinking about the probability that a government-minted coin comes up heads, we might believe that it is very nearly 50%, and because we are fairly certain, we could set the equivalent previous sample size to, say, *n* = 200, which means that of 200 previously observed flips, 100 were heads.
- Our goal is to convert a prior belief expressed in terms of central tendency and sample size into equivalent values of *a* and *b* in the beta distribution.

77

Central Tendency and Certainty

- It turns out that the mean of the beta(θ|*a*, *b*) distribution is μ = *a*/(*a* + *b*) and the mode is ω = (*a* − 1)/(*a* + *b* − 2) for *a* > 1 and *b* > 1.
  - When *a* = *b*, the mean and mode are 0.5.
  - When *a* > *b*, the mean and mode are greater than 0.5.
  - When *a* < *b*, the mean and mode are less than 0.5.
- The spread of the beta distribution is related to the "concentration" κ = *a*+*b*.
  - As κ gets larger, the beta distribution gets narrower or more concentrated.

$$a = \mu\kappa \quad \text{and} \quad b = (1-\mu)\kappa$$

$$a = \omega(\kappa - 2) + 1 \quad \text{and} \quad b = (1-\omega)(\kappa - 2) + 1 \quad \text{for } \kappa > 2$$

78

Establishing the Shape Parameters

- Either μ + κ or ω + κ
  - Mode can be more intuitive than the mean, especially for skewed distributions, because the mode is where the distribution reaches its tallest height.
- Another way: μ + σ (standard deviation)
  - must be careful with this approach, because the standard deviation must make sense in the context of a beta density
  - In particular, σ should typically be less than 0.28867, which is the standard deviation of a uniform distribution.

$$a = \mu \left(\frac{\mu(1-\mu)}{\sigma^2} - 1 \right) \quad \text{and} \quad b = (1-\mu) \left(\frac{\mu(1-\mu)}{\sigma^2} - 1 \right)$$

79

Derive Posterior Using Beta Prior

$$\begin{aligned}
p(\theta|z,N) &= p(z,N|\theta)p(\theta)/p(z,N) && \text{Bayes' rule} \\
&= \theta^z(1-\theta)^{(N-z)} \ \frac{\theta^{(a-1)}(1-\theta)^{(b-1)}}{B(a,b)} \Big/ p(z,N) \\
& && \text{by definitions of Bernoulli and beta distributions} \\
&= \theta^z \ (1-\theta)^{(N-z)} \ \theta^{(a-1)} \ (1-\theta)^{(b-1)} \Big/ \big[B(a,b)p(z,N)\big] && \text{by rearranging factors} \\
&= \theta^{((z+a)-1)} \ (1-\theta)^{((N-z+b)-1)} \Big/ \big[B(a,b) \ p(z,N)\big] && \text{by collecting powers} \\
&= \theta^{((z+a)-1)} \ (1-\theta)^{((N-z+b)-1)} \Big/ B(z+a, N-z+b) && (6.8)
\end{aligned}$$

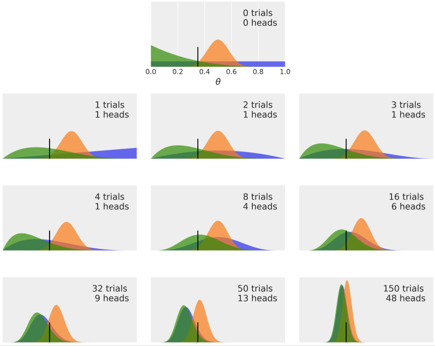
80

Note on Derivation

- The last step in the derivation, from *B*(*a*, *b*)*p*(*z*,*N*) to *B*(*z* + *a*,*N* − *z* + *b*), was not made via some elaborate covert analysis of integrals.
- The transition was made simply by thinking about what the normalizing factor for the numerator must be.
  - The numerator, $\theta^{((z+a)-1)}(1-\theta)^{((N-z+b)-1)}$, is the numerator of a beta(θ|*z* + *a*, *N* − *z* + *b*) distribution.
  - For the function on the previous slide to be a probability distribution, *as it must be*, the denominator *must* be the normalizing factor for the corresponding beta distribution, which is *B*(*z*+*a*,*N*−*z*+*b*) by definition of the beta function.

81

Computing and Plotting Posterior $p(y,N|\theta)$



82

Figure Explanation (1/4)

- On the first subplot of the Figure on previous slide, we have zero trials, thus the three curves represent our priors:
  - The uniform (blue) prior. This represent all the possible values for the bias being equally probable *a priori*.
  - The Gaussian-like (orange) prior is centered and concentrated around *0.5*, so this prior is compatible with information indicating that the coin has more or less about the same chance of landing heads or tails. We could also say this prior is compatible with the *belief* that most coins are fair. While *belief* is commonly used in Bayesian discussions, we think is better to talk about models and parameters that are informed by data.
  - The skewed (green) prior puts the most weight on a tail-biased outcome.

83

Figure Explanation (2/4)

- The rest of the subplots show posterior distributions for successive trials. The number of trials (or coin tosses) and the number of heads are indicated in each subplot's legend. There is also a black vertical line at 0.35 representing the true value for . Of course, in real problems, we do not know this value, and it is here just for pedagogical reasons. This figure can teach us a lot about Bayesian analysis, so grab your coffee, tea, or favorite drink and let's take a moment to understand it:
  - The result of a Bayesian analysis is a posterior distribution, not a single value but a distribution of plausible values given the data and our model.
  - The most probable value is given by the mode of the posterior (the peak of the distribution).
  - The spread of the posterior is proportional to the uncertainty about the value of a parameter; the more spread out the distribution, the less certain we are.

84

Figure Explanation (3/4)

- Intuitively, we are more confident in a result when we have observed more data supporting that result. Thus, even when numerically , seeing four heads out of eight trials gives us more confidence that the bias is 0.5 than observing one head out of two trials. This intuition is reflected in the posterior, as you can check for yourself if you pay attention to the (blue) posterior in the third and sixth subplots; while the mode is the same, the spread (uncertainty) is larger in the third subplot than in the sixth subplot.
- Given a sufficiently large amount of data, two or more Bayesian models with different priors will tend to converge to the same result. In the limit of infinite data, no matter which prior we use, all of them will provide the same posterior. Remember that infinite is a limit and not a number, so from a practical point of view, we could get *practically* indistinguishably posteriors for a finite and rather *small* number of data points.

85

Figure Explanation (4/4)

- How fast posteriors converge to the same distribution depends on the data and the model. In the preceding figure, we can see that the posteriors coming from the blue prior (uniform) and green prior (biased towards tails) converge faster to *almost the same* distribution, while it takes longer for the orange posterior (the one coming from the concentrated prior). In fact, even after 150 trials, it is somehow easy to recognize the orange posterior as a different distribution from the two others.
- Something not obvious from the figure is that we will get the same result if we update the posterior sequentially than if we do it all at once. We can compute the posterior 150 times, each time adding one more observation and using the obtained posterior as the new prior, or we can just compute one posterior for the *150* tosses at once. The result will be exactly the same. This feature not only makes perfect sense, it also leads to a natural way of updating our estimations when we get new data, a situation common in many data-analysis problems.

86

Influence of Prior and How to Choose One

- From the preceding example, it is clear that priors can influence inferences. This is totally fine, priors are supposed to do this. Newcomers to Bayesian analysis (as well as detractors of this paradigm) are generally a little nervous about how to choose priors, because they do not want the prior to act as a censor that does not let the data speak for itself! That's OK, but we have to remember that data does not really speak; at best, data murmurs. Data only makes sense in the context of our models, including mathematical and mental models. There are plenty of examples in the history of science where the same data leads people to think differently about the same topics, and this can happen even if you base your opinions on formal models.

87

Weakly-Informative Priors

- Some people like the idea of using non-informative priors (also known as *flat*, *vague*, or *diffuse priors*); these priors have the least possible amount of impact on the analysis. While it is possible to use them, in general, we can do better. Throughout this book, we will follow the recommendations of Gelman, McElreath, Kruschke, and many others, and we will prefer **weakly-informative priors**. For many problems, we often know something about the values a parameter can take, we may know that a parameter is restricted to being positive, or we may know the approximate range it can take, or whether we expect the value to be close to zero or below/above some value. In such cases, we can use priors to put some weak information in our models without being afraid of being too pushy. Because these priors work to keep the posterior distribution within certain reasonable bounds, they are also known as **regularizing priors**. Using informative priors is also a valid option if we have good-quality information to define those priors. Informative priors are very strong priors that convey a lot of information. Depending on your problem, it could be easy or not to find this type of prior.

88

Take-Home Message

- For example, in structural bioinformatics, people have been using, in Bayesian and non-Bayesian ways, all the prior information they could get to study and especially predict the structure of proteins. This is reasonable because we have been collecting data from thousands of carefully-designed experiments for decades and hence we have a great amount of trustworthy prior information at our disposal. Not using it would be absurd!
- So, the take-home message is this: if you have reliable prior information, there is no reason to discard that information, including the nonsensical argument that being objective means throwing away valuable information. Imagine if every time an automotive engineer had to design a new car, they had to start from scratch and reinvent the combustion engine, the wheel, and for that matter, the whole concept of a car. That's not the way things should work.

89

Frequentist Statistics

- Knowing we can classify priors into categories according to their relative strength does not make us less nervous about choosing from them. Maybe it would be better to not have priors at all—that would make modeling easier, right? Well, not necessarily. Priors can make models behave better, have better generalization properties, and can help us convey useful information. Also, every model, Bayesian or not, has some kind of prior in one way or another, even if the prior is not set explicitly. In fact, many result from frequentist statistics, and can be seen as special cases of a Bayesian model under certain circumstances, such as flat priors.
  - One common frequentist method to estimate parameters is known as maximum likelihood; this methods avoids setting a prior and works just by finding the value of θ that maximizes the likelihood. This value is usually notated by adding a little hat on top of the symbol of the parameter we are estimating, such as $\hat{\theta}$ or sometimes θ_{mle} (or even both). $\hat{\theta}$ is a point estimate (a number) and not a distribution.

90

Computing θ that Maximizes the Likelihood

- For the coin-flipping problem we can compute this analytically:

$$\hat{\theta} = \frac{y}{N}$$

- If you go back to the figure in slide #101, you will be able to check for yourself that the mode of the blue posterior (the one corresponding to the uniform/flat prior) agrees with the values of , computed for each subplot. So, at least for this example, we can see that even when the maximum likelihood method does not explicitly invoke any prior, it can be considered a special case of a Bayesian model, one with a uniform prior.

91

Benefits of Priors

- We cannot really avoid priors, but if we include them in our analysis, we will get several benefits, including
  1. A distribution of plausible values and not only the most probable one.
  2. We get more transparent models, meaning they're easier to criticize, debug (in a broad sense of the word), and hopefully improve. Building models is an iterative process; sometimes the iteration takes a few minutes, sometimes it could take years. Sometimes it will only involve you, and sometimes it will involve people you do not even know. Reproducibility matters and transparent assumptions in a model contribute to it.
  3. We are free to use more than one prior (or likelihood) for a given analysis if we are not sure about any special one; exploring the effect of different priors can also bring valuable information to the table. Part of the modeling process is about questioning assumptions, and priors (and likelihood) are just that. Different assumptions will lead to different models and probably different results. By using data and our domain knowledge of the problem, we will be able to compare models and, if necessary, decide on a winner. (Chapter 5 – *Model Comparison,* will be devoted to this issue.)

92

Communicating a Bayesian Analysis

- Creating reports and communicating results is central to the practice of statistics and data science. In this section, we will briefly discuss some of the peculiarities of this task when working with Bayesian models. In future classes, we will keep looking at examples about this important matter.

93

Model Notation and Visualization

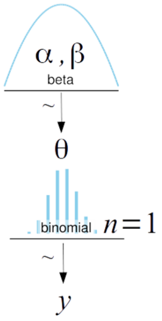
- If you want to communicate the results of an analysis, you should also communicate the model you used. A common notation to succinctly represent probabilistic models is:

$$\theta \sim \text{Beta}(\alpha, \beta)$$

$$\text{y} \sim \text{Bin}(n = 1, p = \theta)$$

94

Kruschke's Diagrams



95

Diagram Explanation

- On the first level, we have the prior that generates the values for θ, then the likelihood, and on the last line the data, *y*. Arrows indicate the relationship between variables, and the ~ symbol indicates the stochastic nature of the variables.

96

Summarizing Posterior

- The result of a Bayesian analysis is a posterior distribution, and all the information about the parameters given a dataset and a model is contained in the posterior distribution. Thus, by summarizing the posterior, we are summarizing the logical consequences of a model and data. A common practice is to report, for each parameter, the mean (or mode or median) to have an idea of the location of the distribution and some measure, such as the standard deviation, to have an idea of the dispersion and hence the uncertainty in our estimate. The standard deviation works well for normal-like distributions but can be misleading for other type of distributions, such as skewed ones. So, an alternative is to use the following measure.

97

Highest-Posterior Density (a.k.a. Highest Density Interval)

- A commonly-used device to summarize the spread of a posterior distribution is to use a **Highest-Posterior Density** (**HPD**) interval. An HPD is the shortest interval containing a given portion of the probability density. One of the most commonly-used is the 95% HPD, often accompanied by the 50% HPD. If we say that the 95% HPD for some analysis is [2-5], we mean that according to our data and model, we think the parameter in question is between 2 and 5 with a probability of 0.95.
  - There is nothing special about choosing 95%, 50%, or any other value. They are just arbitrary commonly-used values. 95% is a default value.
  - Ideally, justifications should be context-dependent and not automatic.

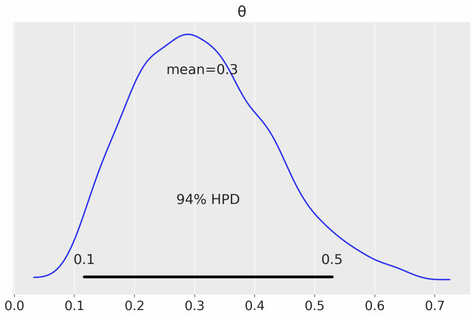
98

ArviZ

- ArviZ is a Python package for exploratory data analysis for Bayesian models. ArviZ has many functions to help us summarize the posterior, for example, az.plot_posterior can be used to generate a plot with the mean and HPD of a distribution.

99

An Example Figure from ArviZ



100

Two More Points on HPD

- ArviZ will use, by default, a value of 0.94 (corresponding to 94%) for HPD. You can change this by passing a different value to the credible_interval argument.
- Highest Posterior Density Intervals (used by Bayesian analysts) != Confidence Intervals (used by frequentists)

101

Posterior Predictive Checks

- One of the nice elements of the Bayesian toolkit is that once we have a posterior, it is possible to use the posterior, $p(\theta|y)$, to generate predictions, $\tilde{y}$, based on the data, y, and the estimated parameters, θ. The posterior predictive distribution is:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\, d\theta$$

102

- Thus, the posterior predictive distribution is an average of conditional predictions over the posterior distribution of . Conceptually (and computationally), we approximate this integral as an iterative two-step process:
  1. We sample a value of θ from the posterior, $p(\theta|y)$
  2. We feed that value of θ to the likelihood (or sampling distribution if you wish), thus obtaining a data point, $\tilde{y}$
- This process combines two sources of uncertainty: the parameters uncertainty; as captured by the posterior; and the sampling uncertainty; as captured by the likelihood.

103

Posterior Predictive Checks

- The generated predictions, $\tilde{y}$, can be used when we need to make predictions. But also we can use them to criticize the models by comparing the observed data, y, and the predicted data, $\tilde{y}$, to spot differences between these two sets, this is known as posterior predictive checks.
  - The main goal is to check for auto-consistency. The generated data and the observed data should look more or less similar, otherwise there was some problem during the modeling or some problem feeding the data to the model. But even in the absence of mistakes, differences could arise. Trying to understand the mismatch could lead us to improve models or at least to understand their limitations. Knowing which parts of our problem/data the model is capturing well and which it is not is valuable information even if we do not know how to improve the model.

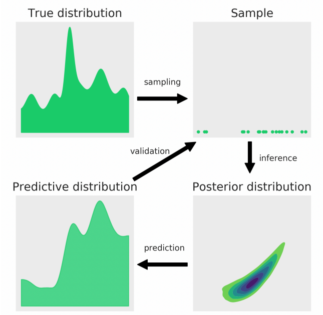
104

Summary

- We began our Bayesian journey with a very brief discussion about statistical modeling, probability theory, and the introduction of Bayes' theorem. We then used the coin-flipping problem as an excuse to introduce basic aspects of Bayesian modeling and data analysis. We used this classic example to convey some of the most important ideas of Bayesian statistics, such as using probability distributions to build models and represent uncertainties. We tried to demystify the use of priors and put them on an equal footing with other elements that are part of the modeling process, such as the likelihood, or even more *meta-questions*, such as why we are trying to solve a particular problem in the first place. We ended by discussing the interpretation and communication of the results of a Bayesian analysis.

105

Summarizing Bayesian Workflow



106

Figure Explanation

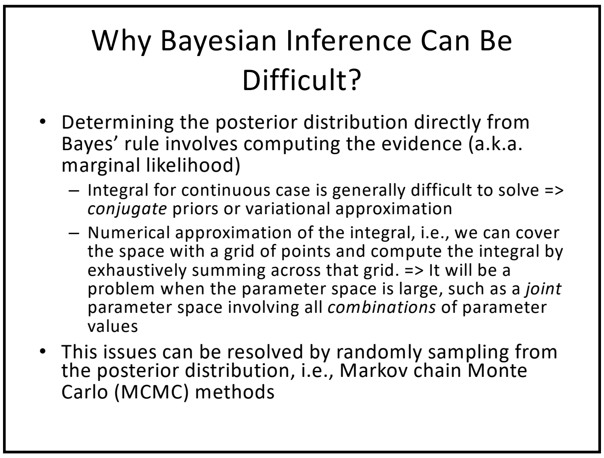
- There is a **True distribution** that in general is unknown, from which we get a finite **sample** (by doing an experiment, a survey, an observation, or a simulation). In order to learn something from the True distribution, given that we have only observed a **sample**, we build a probabilistic model (a prior and a likelihood). Using the model and the sample, we perform Bayesian Inference and obtain a **Posterior distribution**, which encapsulates all the information about a problem, given our model and data. From a Bayesian perspective, the posterior distribution is the main object of interest and everything else is derived from it, including predictions in the form of a **Posterior Predictive Distribution**.

107

How to Evaluate Our Model?

- As the Posterior distribution (and any other derived quantity from it) is a consequence of the model and data, the usefulness of Bayesian inferences are restricted by the quality of models and data. One way to evaluate our model is by comparing the Posterior Predictive Distribution with the finite sample we got in the first place. Notice that the Posterior distribution is a distribution of the parameters in a model (conditioned on the observed samples), while the Posterior Predictive Distribution is a distribution of the predicted samples (averaged over the posterior distribution). The process of model validation is of crucial importance not because we want to be sure we have *the right model*, but because we know we almost never have *the right model*. We check models to evaluate whether they are *useful enough* in a specific context and, if not, to gain insight into how to improve them.

108



Why Bayesian Inference Can Be Difficult?

- Determining the posterior distribution directly from Bayes' rule involves computing the evidence (a.k.a. marginal likelihood)
  - Integral for continuous case is generally difficult to solve => *conjugate* priors or variational approximation
  - Numerical approximation of the integral, i.e., we can cover the space with a grid of points and compute the integral by exhaustively summing across that grid. => It will be a problem when the parameter space is large, such as a *joint* parameter space involving all *combinations* of parameter values
- This issues can be resolved by randomly sampling from the posterior distribution, i.e., Markov chain Monte Carlo (MCMC) methods

109