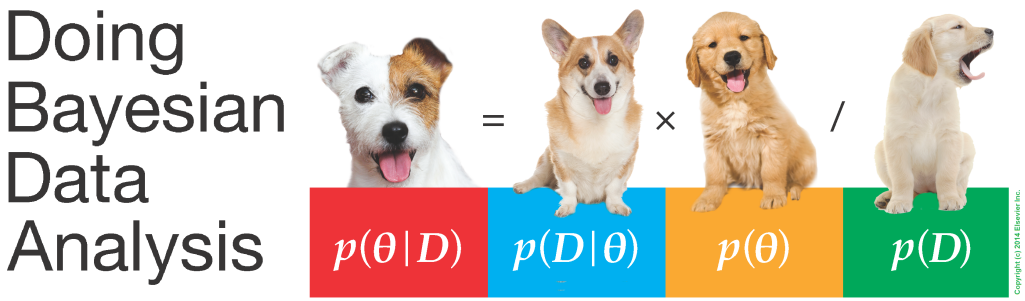
Probability Review for Data Science

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University
yuxiaohuang@gwu.edu

August 23, 2018

Reference



Picture courtesy of the book website

- This set of slices is an excerpt of the book by Professor John K. Kruschke, with some trivial changes by the creator of the slides
- Please find the reference to and website of the book below:
  - Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier
  - `https://sites.google.com/site/doingbayesiandataanalysis/`

Overview

1. Probability theory
2. Bayesian networks
3. Bayesian inference

The sample space

- The sample space is a set of events that are exhaustive and mutually exclusive
- The sample space may refer to:
  - data: when we flip a coin, we are sampling from the space of possible outcomes, {heads, tails}
  - parameters: when we grab a coin at random from a sack of coins, we are sampling from the space of possible biases, $[0, 1]$

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?
- **A:** Because coin flips represent every real-life event that has a binary outcome. **Q:** Any examples?

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?
- **A:** Because coin flips represent every real-life event that has a binary outcome. **Q:** Any examples?
- **A:**
  - for a heart surgery, whether it is successful or not
  - for a drug, whether it has side effect or not
  - for a survey question, whether the answer is correct or not
  - for a two-candidate election, whether one wins or not
  - ...

Two kinds of probabilities

- **Q:** Recall, what are the two kinds of sample space?

Two kinds of probabilities

- **Q:** Recall, what are the two kinds of sample space?
- **A:**
  - the sample space of the data (e.g., heads or tails)
  - the sample space of the parameters (e.g., bias)
- Each kind of sample space corresponds to a kind of probability
  - for data, the probability is over measurable outcomes that are "out there" in the world
  - for parameters, the probability is over unmeasurable beliefs that are "inside the head"

Probabilities assign numbers to possibilities

- A probability, no matter which kind it is, is just a way of assigning numbers to a set of exhaustive and mutually exclusive events
- A probability needs to satisfy three properties (Kolmogorov, 1956):
  - a probability value must be nonnegative
  - the sum of the probabilities across all events in the entire sample space must be 1
  - for any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities

Probability distributions

- A probability distribution is simply a list of all possible events and their corresponding probabilities
- There are two kinds of probability distribution
  - discrete distribution: e.g., probabilities over heads or tails
  - continuous distribution: e.g., probabilities over people's heights

Discrete distributions: probability mass

- When the sample space consists of discrete outcomes (e.g., heads or tails), the probability distribution is a list of probabilities of the outcomes
- The probability of a discrete outcome is called as a probability mass
- The sum of the probability masses across the sample space must be 1

Continuous distributions: rendezvous with density

- When the sample space consists of continuous outcomes (e.g., people's heights), we cannot use probability mass for a specific outcome. **Q:** Why?

Continuous distributions: rendezvous with density

- When the sample space consists of continuous outcomes (e.g., people's heights), we cannot use probability mass for a specific outcome. **Q:** Why?
- **A:**
  - because the probability mass for a specific outcome will be zero
  - e.g., the probability of someone's height being 67.21413908...
- Instead, we can:
  1. discretize the space into a finite set of mutually exclusive and exhaustive intervals
  2. calculate the probability mass in each interval
  3. use the ratio of probability mass to interval width
  4. this ratio is called the probability density

Probability density

- The top panel of Figure 4.2 (see next page) shows the discretized intervals and probability mass in each interval
- The second panel shows the probability density
- The third panel shows the narrower intervals and probability mass in each interval
- The bottom panel shows the probability density corresponding to the narrower intervals
- Generally, the narrower the intervals are, the more accurate the probability density is

Figure 4.2

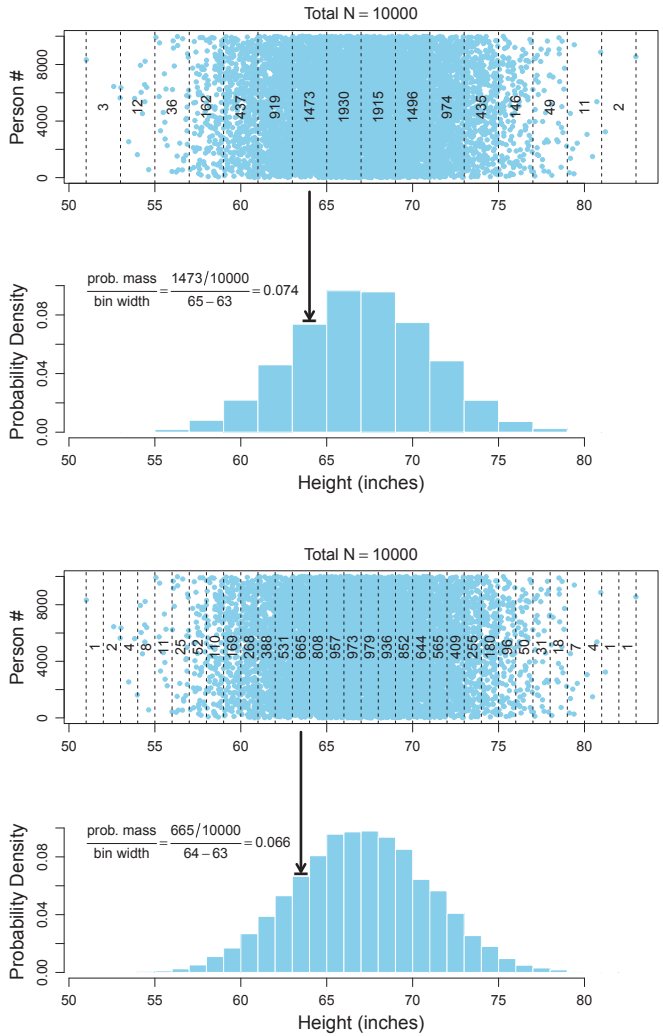


Figure 4.2: Examples of computing probability density. Within each main panel, the upper plot shows a scatter of 10,000 heights of randomly selected people, and the lower plot converts into probability density for the particular selection of bins depicted. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Probability density

- While probability mass cannot exceed 1, probability densities can
- The upper panel of Figure 4.3 (see next page) shows that most of the probability mass is concentrated around 84
- Consequently, the probability density near 84 exceeds 1.0, as shown in the lower panel
- This simply means that there is a high concentration of probability mass relative to the width of the interval

Figure 4.3

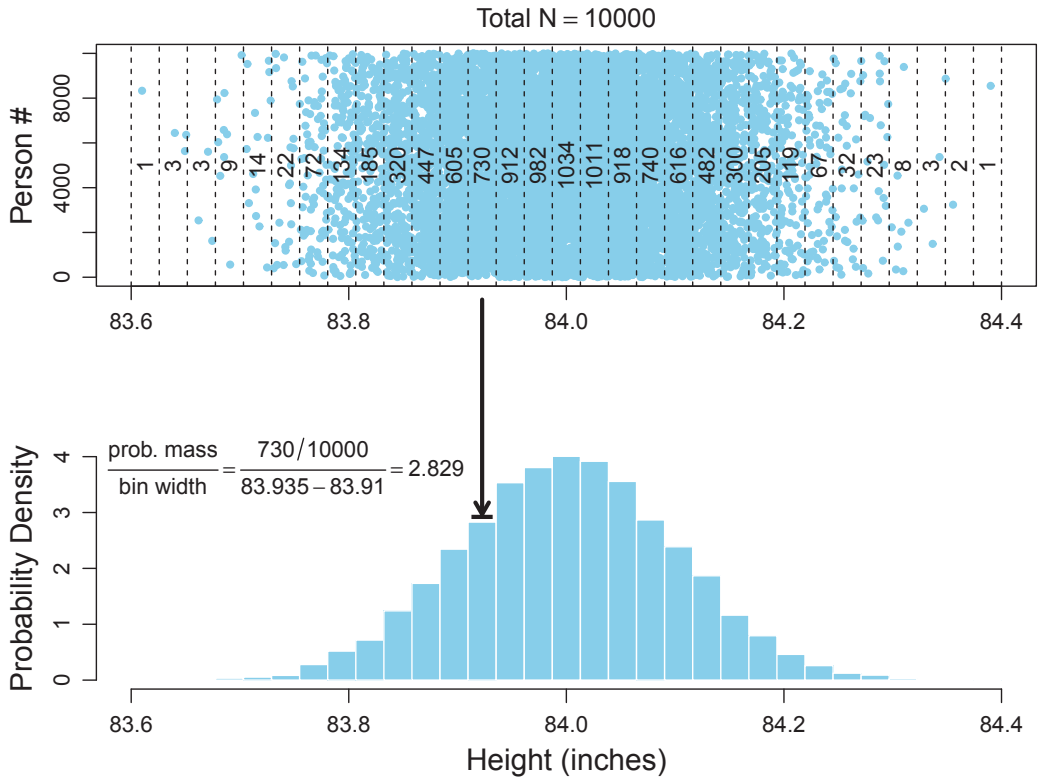


Figure 4.3: Example of probability density greater than 1.0. Here, all the probability mass is concentrated into a small region of the scale, and therefore the density can be high at some values of the scale. The annotated calculation of density uses rounded interval limits for display. (For this example, we can imagine that the points refer to manufactured doors instead of people, and therefore the *y*-axis of the top panel should be labelled "Door" instead of "Person.") Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Properties of probability density functions

- We need to define some notations first. Let:
  - x be the continuous variable
  - Δx be the width of an interval on x
  - i be an index for the intervals
  - $[x_i, x_i + \Delta x]$ be the interval between x_i and $x_i + \Delta x$
  - $p([x_i, x_i + \Delta x])$ be the probability mass of the ith interval
- Then the sum of those probability masses must be 1:

$$\sum_i p([x_i, x_i + \Delta x]) = 1$$

- We can rewrite the equation above in terms of the density of each interval, by dividing and multiplying by Δx:

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

Properties of probability density functions

- In the limit, as the interval width becomes infinitesimal, we denote:
  - summation as $\int$ instead of $\sum$
  - the width of the interval around x as $\mathrm{d}x$ instead of Δx
  - the probability density in the infinitesimal interval around x as $p(x)$
- Then the previous equation (in terms of density) can be rewritten as:

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

- By an abuse of notation, we use $p(x)$ to represent the probability mass when x is discrete
- Thus, what $p(x)$ represents depends on the context (x being discrete or continuous)

The normal probability density function

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is

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

- **Q:** What are μ and σ? what do they control?

The normal probability density function

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is

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

- **Q:** What are μ and σ? what do they control?
- **A:**
  - μ: the mean/mode/median, location parameter
  - σ: the standard deviation, scale parameter
- An example of the probability density is shown in Figure 4.4 (see next page), where x axis is divided into a dense comb of small intervals
- The figure also shows that the area under the curve is, in fact, 1

Figure 4.4

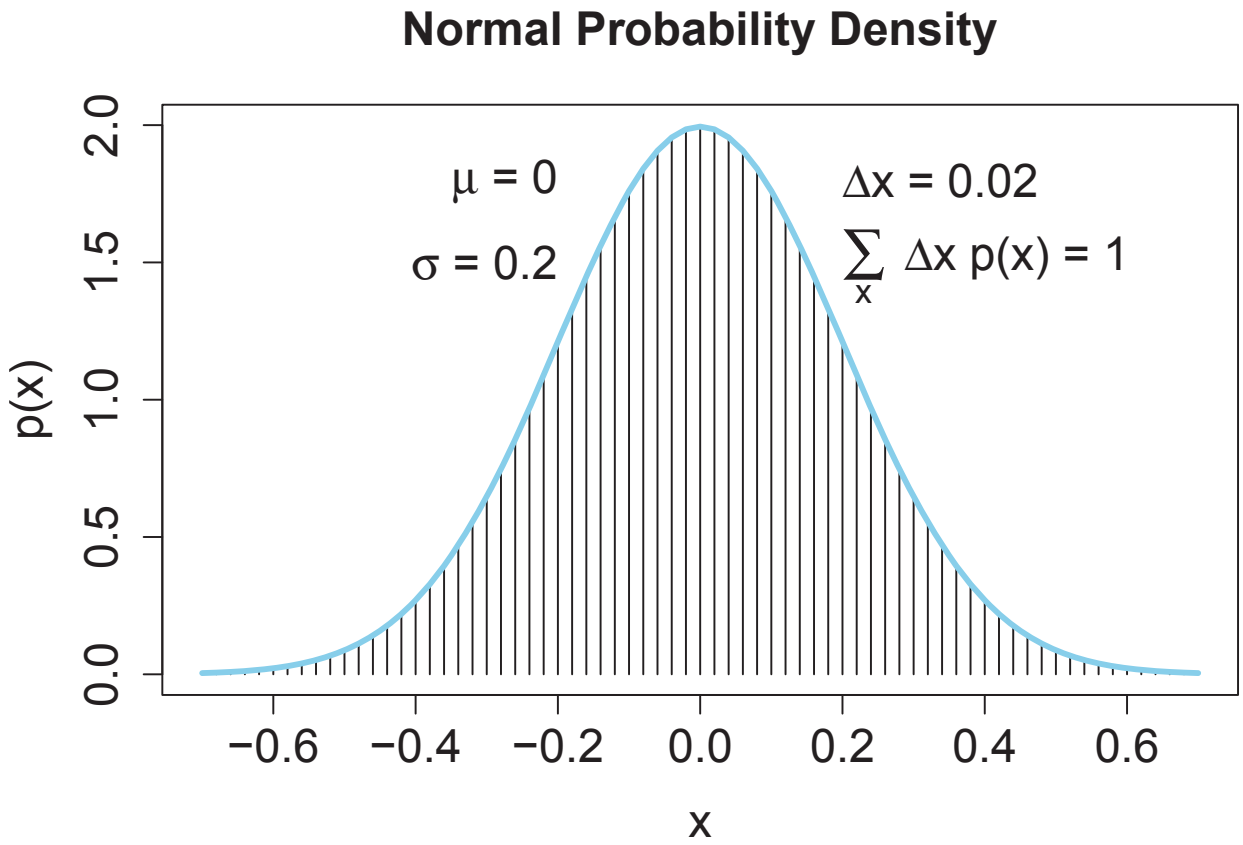


Figure 4.4: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Mean of a distribution

- The mean of a probability distribution is the long-run average of the values
- The mean is also called the expected value, denoted by $E[x]$
  - when x is discrete:

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

  - when x is continuous:

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

Variance of a distribution

- The variance of a probability distribution is a number that represents the dispersion of the distribution away from its mean
- The definition of variance, var_x, is the mean squared deviation (MSD) of the x values from their mean
  - when x is discrete:

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

  - when x is continuous:

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

- In other words, the variance is just the average value of $(x - E[x])^2$

Joint probability and marginal probability

- Table 4.1 (see next page) shows the probabilities of various combinations of people's eye color and hair color
- Each entry indicates the joint probability of particular combinations of eye color (e) and hair color (h), denoted by $p(e, h)$
- The right margin of the table shows the probabilities of the eye colors overall, collapsed across hair colors
- Such probabilities are called marginal probability, denoted by $p(e)$:

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

- The marginal probabilities of the hair colors, $p(h)$, are indicated on the lower margin of the table:

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

Table 4.1

Table 4.1: Proportions of combinations of hair color and eye color. 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). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal
	Black	Brunette	Red	Blond	(Eye Color)
Brown	.11	.20	.04	.01	.37
Blue	.03	.14	.03	.16	.36
Hazel	.03	.09	.02	.02	.16
Green	.01	.05	.02	.03	.11
Marginal (Hair Color)	.18	.48	.12	.21	1.0

Conditional probability

- We may want to know the probability of an outcome y, given some evidence x, $p(y|x)$
- This is sometimes called the conditional probability of y given x, and can be calculated as

$$p(y|x) = \frac{p(x,y)}{p(x)},$$

  that is, the number of instances where both x and y being true, out of the number where x being true
- Table 4.2 (see next page) shows the conditional probability of a hair color, given the eye color being blue

Table 4.2

Table 4.2: Example of conditional probability. Of the blue-eyed people in Table 4.1, what proportion have hair color h? Each cell shows $p(h|\text{blue}) = p(\text{blue}, h)/p(\text{blue})$ rounded to two decimal points. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal (Eye Color)
	Black	Brunette	Red	Blond	
Blue	.03/.36 = .08	.14/.36 = .39	.03/.36 = .08	.16/.36 = .45	.36/.36 = 1.0

Absolute independence

- We say x and y are absolutely independent, if

$$p(x, y) = p(x)p(y)$$

- Examples of absolute independence include:
  - eye color and height
  - hair color and weight
  - ...

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?
- **A:**

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

- **Q:** What is the overall size of the marginal probabilities?

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?
- **A:**

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

- **Q:** What is the overall size of the marginal probabilities?
- **A:** $2n$
- Absolute independence reduces the size from 2^n to $2n$!

Conditional Independence

- While powerful, absolute independence is rare
- It is more common to have conditional independence
- We say x and y are conditionally independent given z, if

$$p(x|z) = p(x|z, y)$$

- Examples of conditional independence include:
  - yellow finger and lung cancer, given the status of smoking
  - eye color and hair color, given the genes
  - ...

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?
- **A:**

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1})$$

- **Q:** What is the overall size of the conditional probabilities?

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?
- **A:**

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1})$$

- **Q:** What is the overall size of the conditional probabilities?
- **A:** $4n - 2$
- Conditional independence reduces the size from 2^n to $4n - 2$!

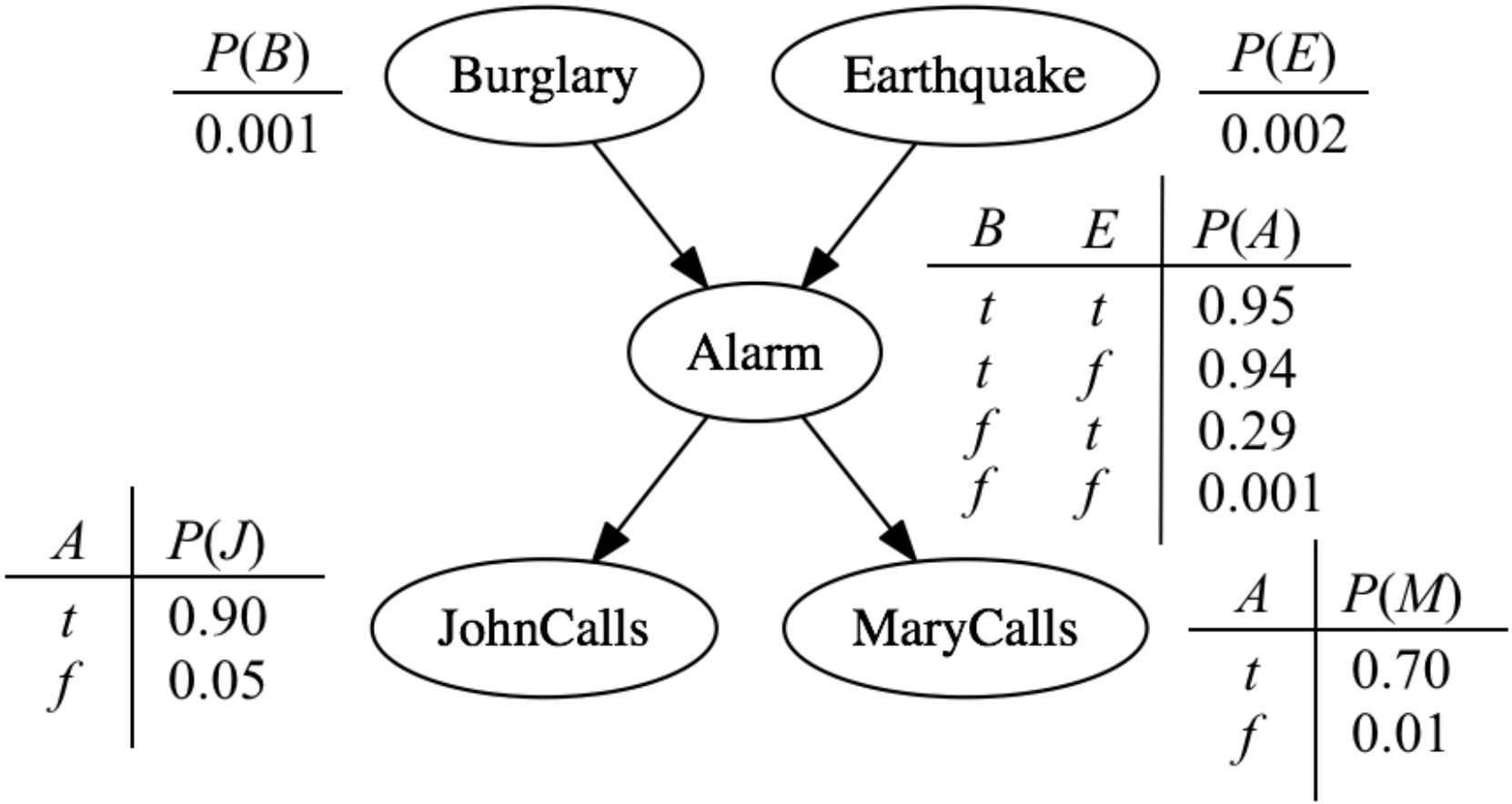
Bayesian networks

- Joint distribution
  - specifies joint distribution in a structured form
  - directed acyclic graph (aka DAG)
- Conditional Independence
  - nodes: random variables
  - edges: causation

Alarm network

- Nodes
  - B: a burglary occurs at your house
  - E: an earthquake occurs at your house
  - A: the alarm goes off due to B or E
  - J: John calls to report the alarm
  - M: Mary calls to report the alarm

The structure and conditional probability tables (CPTs)



Bayes' rule

- Bayes' rule is merely the mathematical relation between the prior allocation of credibility and the posterior reallocation of credibility conditional on data

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?
    - **A**:

$$p(\text{cloudy}) < p(\text{cloudy}|\text{raining})$$

  - suppose, instead, you have some data showing everyone outside is wearing sunglasses
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{sunglasses})$
    - **Q:** how does the data change your belief?

Example: reallocating posterior credibility

- Suppose you have some prior belief about cloudy weather, $P(\text{cloudy})$
  - suppose you have some data showing it is raining outside
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{raining})$
    - **Q:** how does the data change your belief?
    - **A**:

$$p(\text{cloudy}) < p(\text{cloudy}|\text{raining})$$

  - suppose, instead, you have some data showing everyone outside is wearing sunglasses
    - now you have a posterior belief about the weather being cloudy, given the data, $p(\text{cloudy}|\text{sunglasses})$
    - **Q:** how does the data change your belief?
    - **A:**

$$p(\text{cloudy}) > p(\text{cloudy}|\text{sunglasses})$$

Derived from definitions of conditional probability

- **Q:** Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?

Derived from definitions of conditional probability

- **Q:** Recall, based on the definition of conditional probability, what are $p(\theta|y)$ and $p(y|\theta)$?
- **A:**

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} \tag{1}$$

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)} \tag{2}$$

- We can write eq.(2) as

$$p(\theta, y) = p(y|\theta)p(\theta) \tag{3}$$

- Replace $p(\theta, y)$ in eq.(1) with that in eq.(3):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

- This is the Bayes' rule

Derived from definitions of conditional probability

- The denominator in Bayes' rule (eq.(4)) can be factorized as

$$p(y) = \begin{cases} \sum_\theta p(y|\theta)p(\theta), & \text{when } y \text{ is discrete} \\ \int_\theta \mathrm{d}\theta \; p(y|\theta)p(\theta), & \text{when } y \text{ is continuous} \end{cases}$$

- Based on the factorization of $p(y)$, we can write $p(\theta|y)$ as

$$p(\theta|y) = \begin{cases} \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}, & \text{when } y \text{ is discrete} \\ \frac{p(y|\theta)p(\theta)}{\int_\theta \mathrm{d}\theta \; p(y|\theta)p(\theta)}, & \text{when } y \text{ is continuous} \end{cases}$$

- This is how we usually use the Bayes' rule

Applied to parameters and data

- Bayes' rule can be used to estimate the parameters based on the data
- Let θ be the parameters and D the data, then Bayes' rule can be written as

$$p(\theta|D) = p(D|\theta)p(\theta)/p(D)$$

- The factors of Bayes' rule have specific names:

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

- The evidence is also called the marginal likelihood

Example: disease diagnosis

- Suppose the prior probability of having a disease (θ) is 0.001:

$$p(\theta = 1) = 0.001$$

- There is a test (y) 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:

$$p(y = 1|\theta = 1) = 0.99$$

- The test has a false alarm rate of 5%:

$$p(y = 1|\theta = 0) = 0.05$$

- **Q:** Suppose the test result is positive. What is the probability of having the disease?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?
- **A:**
  - we already know the prior probability ($p(\theta)$) and likelihood ($p(y|\theta)$)
  - we want to know the posterior probability ($p(\theta|y)$)
- **Q:** Can you solve this problem using Bayes' rule?

Example: disease diagnosis

- **Q:** What do we already know and what do we want to know?
- **A:**
  - we already know the prior probability ($p(\theta)$) and likelihood ($p(y|\theta)$)
  - we want to know the posterior probability ($p(\theta|y)$)
- **Q:** Can you solve this problem using Bayes' rule?
- **A:**

$$
\begin{aligned}
p(\theta = 1|y = 1) &= \frac{p(y = 1|\theta = 1)p(\theta = 1)}{\sum_\theta p(y = 1|\theta)p(\theta)} \\
&= \frac{p(y = 1|\theta = 1)p(\theta = 1)}{p(y = 1|\theta = 1)p(\theta = 1) + p(y = 1|\theta = 0)p(\theta = 0)} \\
&= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times (1 - 0.001)} \\
&= 0.019
\end{aligned}
$$

Data order invariance

- Suppose we first observe data D then D'. Bayes' rule gives us

$$p(\theta) \to p(\theta|D) \to p(\theta|D', D)$$

- Now suppose we observe data in a reversed order: first D' then D. Bayes' rule gives us

$$p(\theta) \to p(\theta|D') \to p(\theta|D, D')$$

- **Q:** Does our final belief depend on the order of the data? In other words, does the following equation hold?

$$p(\theta|D', D) = p(\theta|D, D')$$

Data order invariance

- Suppose we first observe data D then D'. Bayes' rule gives us

$$p(\theta) \to p(\theta|D) \to p(\theta|D', D)$$

- Now suppose we observe data in a reversed order: first D' then D. Bayes' rule gives us

$$p(\theta) \to p(\theta|D') \to p(\theta|D, D')$$

- **Q:** Does our final belief depend on the order of the data? In other words, does the following equation hold?

$$p(\theta|D', D) = p(\theta|D, D')$$

- **A:** Our final belief *does not* depend on the order of the data, when the data are independent:

$$p(D, D'|\theta) = p(D|\theta) \cdot p(D'|\theta)$$

Bayesian inference in one sentence

Bayesian inference reallocates belief from what you know to what you see.

Example: wet sidewalk

- Suppose we step outside one morning and find that the sidewalk is wet
- **Q:** What could be the causes?

Example: wet sidewalk

- Suppose we step outside one morning and find that the sidewalk is wet
- **Q:** What could be the causes?
- **A:**
  - recent rain
  - recent garden irrigation
  - a newly erupted underground spring
  - a broken sewage
  - a passerby who spilled a drink
  - ...
- Based on our previous knowledge, the prior credibilities (probabilities) of some causes are greater than those of the others. For example:
  - $P(\text{recent rain}) > P(\text{recent garden irrigation})$
  - $P(\text{recent rain}) > P(\text{a passerby who spilled a drink})$

Example: wet sidewalk

- Suppose we observe that the sidewalk is wet and so are the trees and parked cars
- **Q:** What does this observation tell us?

Example: wet sidewalk

- Suppose we observe that the sidewalk is wet and so are the trees and parked cars
- **Q:** What does this observation tell us?
- **A:**
  - $P(\text{recent rain} \mid \text{observation}) \uparrow$
  - $P(\text{other causes} \mid \text{observation}) \downarrow$

Example: wet sidewalk

- Suppose we observe that the sidewalk is wet and so are the trees and parked cars
- **Q:** What does this observation tell us?
- **A:**
  - $P(\text{recent rain} \mid \text{observation}) \uparrow$
  - $P(\text{other causes} \mid \text{observation}) \downarrow$
- Suppose we observe that the wetness was localized to a small area, and there was an empty drink cup nearby
- **Q:** What does this tell us?

Example: wet sidewalk

- Suppose we observe that the sidewalk is wet and so are the trees and parked cars
- **Q:** What does this observation tell us?
- **A:**
  - $P(\text{recent rain} \mid \text{observation}) \uparrow$
  - $P(\text{other causes} \mid \text{observation}) \downarrow$
- Suppose we observe that the wetness was localized to a small area, and there was an empty drink cup nearby
- **Q:** What does this tell us?
- **A:**
  - $P(\text{a passerby who spilled a drink} \mid \text{observation}) \uparrow$
  - $P(\text{other causes} \mid \text{observation}) \downarrow$

Example: Sherlock Holmes

Picture courtesy of wikipedia

Example: Sherlock Holmes

- Sherlock Holmes often said to his sidekick, Doctor Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, chap. 6)
- **Q:** What does it mean?

Example: Sherlock Holmes

- Sherlock Holmes often said to his sidekick, Doctor Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890, chap. 6)
- **Q:** What does it mean?
- **A:**
  - there are a set of potential causes for a crime, $\{\theta_1, \theta_2, \ldots, \theta_n\}$
  - if none of the causes, except for θ_i, can explain the evidence, y:

$$P(y|\theta_j) = 0 \quad \text{where} \quad j \neq i,$$

  - then no matter how unlikely θ_i seemed before observing the evidence, it must be the real cause given the evidence:

$$P(\theta_i|y) = 1 \quad \text{even when} \quad P(\theta_i) \ll 1$$

- Figure 2.1 (see next page) illustrates Holmes' reasoning

Figure 2.1

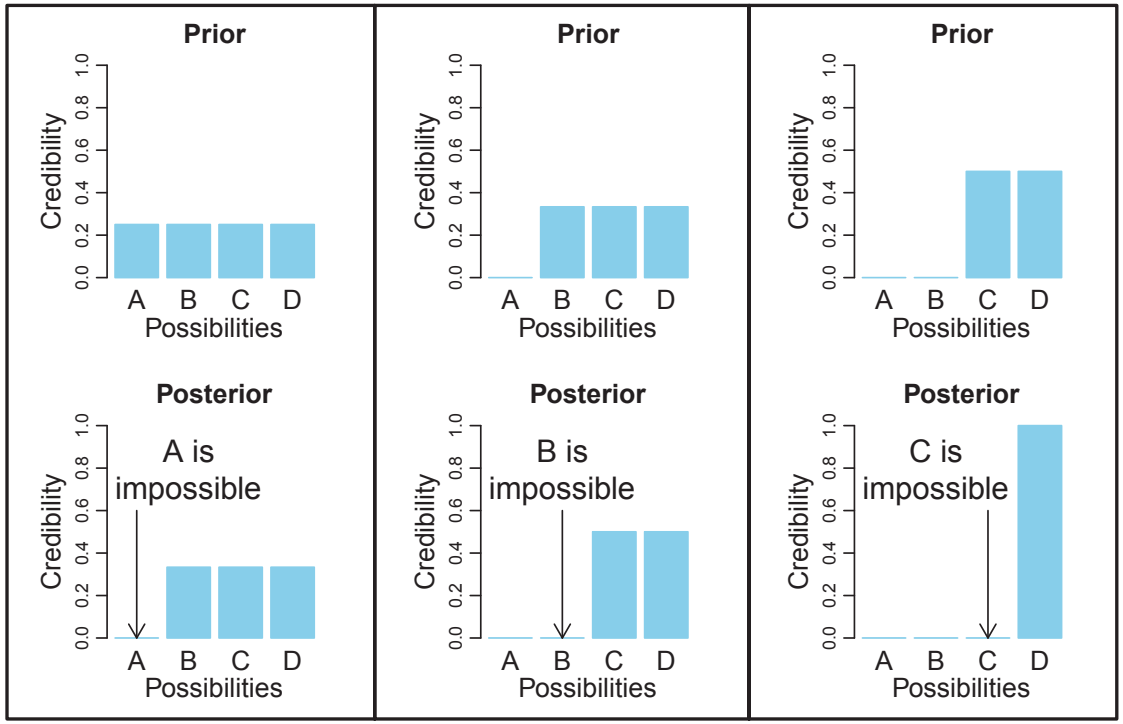


Figure 2.1: The upper-left graph shows the credibilities four possible causes for an outcome. The causes, labeled A, B, C and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset, hence all have prior credibility of 0.25. The lower-left graph shows the credibilities when one cause is learned to be impossible. The resulting posterior distribution is used as the prior distribution in the middle column, where another cause is learned to be impossible. The posterior distribution from the middle column is used as the prior distribution for the right column. The remaining possible cause is fully implicated by Bayesian re-allocation of credibility.

Example: Sherlock Holmes (continued)

- In the previous example, we found evidence y that ruled out A
- The evidence y can be, for example, footprints that are *absolutely* clear so that they *absolutely* cannot belong to A:

$$P(y|A) = 0$$

- Consequently, A can be *absolutely* ruled out based on y:

$$P(A|y) = 0$$

- **Q:** What if the footprints y are *not absolutely* clear? That is

$$P(y|A) > 0$$

Example: Sherlock Holmes (continued)

- In the previous example, we found evidence y that ruled out A
- The evidence y can be, for example, footprints that are *absolutely* clear so that they *absolutely* cannot belong to A:

$$P(y|A) = 0$$

- Consequently, A can be *absolutely* ruled out based on y:

$$P(A|y) = 0$$

- **Q:** What if the footprints y are *not absolutely* clear? That is

$$P(y|A) > 0$$

- **A:** If this were the case, A would not be ruled out:

$$P(A|y) > 0$$

Noisy data and probabilistic inference

- All scientific data have some degree of "noise" (e.g., footprints that are not absolutely clear)
- As a result, our decisions based on the data are not deterministic (e.g., ruling out A for sure)
- The beauty of bayesian inference is that, it makes probabilistic inference from the data, which reveals exactly how much to re-allocate probability (e.g., how likely A is the criminal given the blurry footprints)

Models and parameters

- Bayesian inference begins with a family of candidate models that characterize the trends and spreads in the data
- The parameters determine the exact shape of the models
- You can think of the models as devices (e.g., music player) that simulate data generation (e.g., music)
- You can think of the parameters as control knobs (e.g., volume control) on the devices

Two desiderata for a model

- First, a model should be comprehensible with meaningful parameters
  - normal distribution (as shown in Figure 2.4 on next page) has two parameters, mean and standard deviation
  - the mean controls the location of the distribution's central tendency, and thus sometimes called a location parameter
  - the standard deviation controls the width or dispersion of the distribution, and thus sometimes called a scale parameter
- Second, a model should fit the data well

Figure 2.4

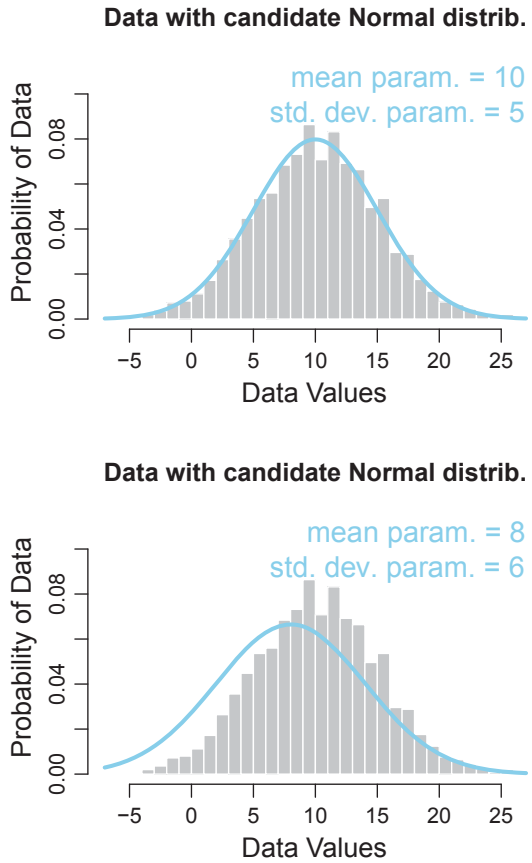


Figure 2.4: The two graphs show the same data histogram but with two different candidate descriptions by normal distributions. Bayesian analysis computes the relative credibilities of candidate parameter values. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Bayesian inference $\neq$ Causal inference

- Causal Inference aims to find the model that generates the data
  - smoking is a common cause of yellow finger and lung cancer
  - the relationship between yellow finger and lung cancer is correlational, but not causal
- Bayesian inference aims to find the model that fits the data
  - yellow finger can be used to predict lung cancer
  - however, yellow finger does not cause lung cancer
- The model that generates the data usually fits the data, not vice versa

The Five Steps of Bayesian Inference

1. Identify the data relevant to the research
2. Specify a model for the data
3. Specify a prior distribution for the parameters
4. Infer the posterior distribution of the parameters
5. Check whether the posterior distribution fits the data well
   - this is also known as "posterior predictive check"
   - if the posterior distribution does not fit the data, go back to step 2

Example: predicting a person's weight using their height

- Suppose we are interested in the relationship between weights and heights of people
- Particularly we would like to know:
  - by how much people's weights increase when heights increase
  - how certain we are about the magnitude of the increase
- **Q:** What are the five steps of bayesian inference?

Example: predicting a person's weight using their height

- Suppose we are interested in the relationship between weights and heights of people
- Particularly we would like to know:
  - by how much people's weights increase when heights increase
  - how certain we are about the magnitude of the increase
- **Q:** What are the five steps of bayesian inference?
- **A:**
  - identifying the data
  - defining a model
  - specifing a prior distribution
  - inferring the posterior distribution
  - checking the posterior distribution

Step 1: identifying the data

- Suppose we have collected heights and weights of 57 adults
- A scatter plot of the data is shown in the left panel of Figure 2.5 (see next page)

Figure 2.5

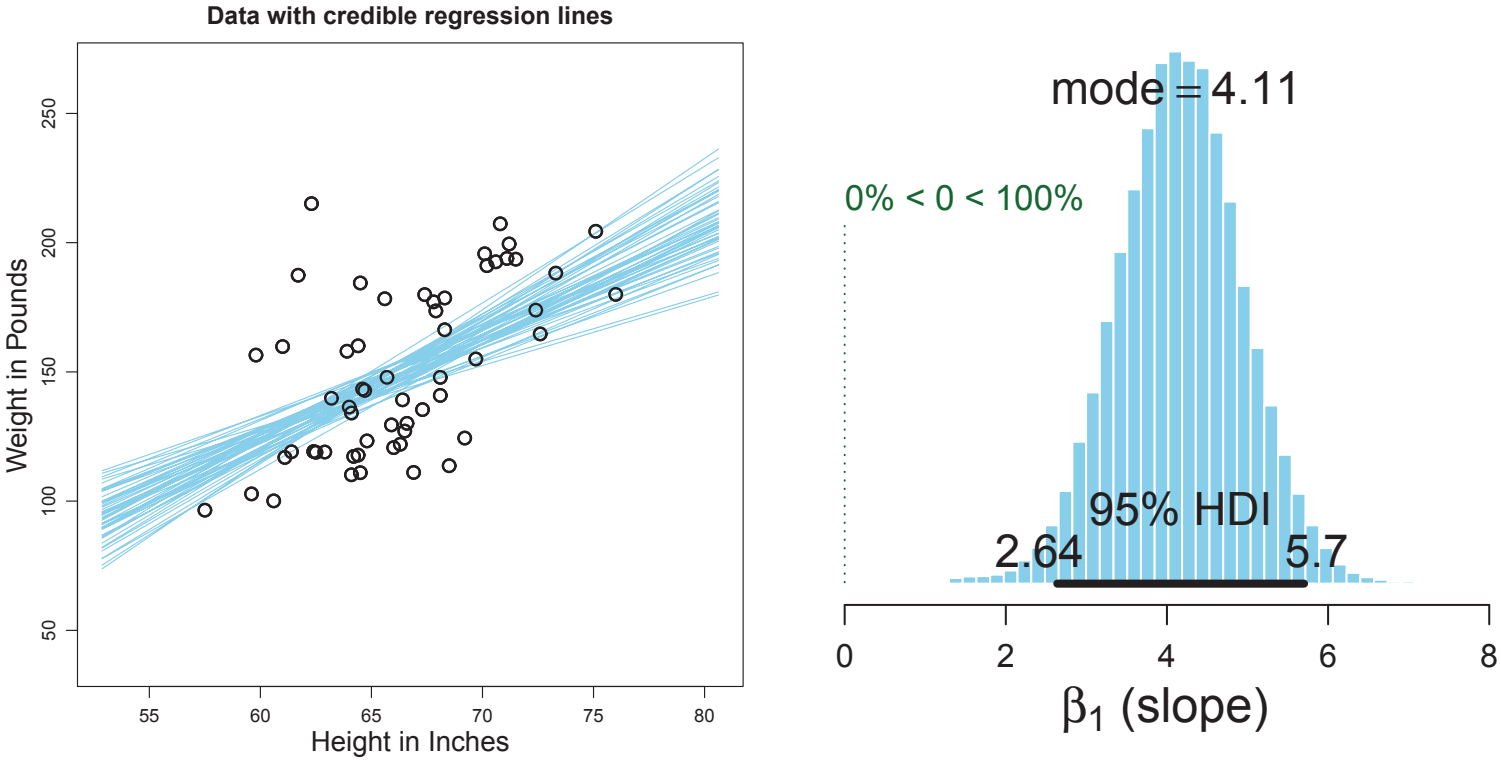


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Step 2: defining a model

- We aim to identify the relationship between weight and height
- We assume that weight is proportional to height
- Therefore, we will describe predicted weight ($\hat{y}$) as a multiplier (β_1) times height (x) plus a baseline (β_0):

$$\hat{y} = \beta_1 x + \beta_0$$

- As shown in the left panel of Figure 2.5 (see previous page), the model above is the form of a line (where β_1 is the slope and β_0 the intercept), and thus often called linear regression

Step 2: defining a model

- The previous model describes the linear relationship between height and weight
- However, since the actual weights may vary around the predicted ones, we need another model to capture this variation
- We assume that the actual weights y follows a normal distribution, with mean $\hat{y}$ (the predicted weights) and standard deviation σ:

$$y \sim N(\hat{y}, \sigma)$$

- **Q:** What does this model mean?

Step 2: defining a model

- The previous model describes the linear relationship between height and weight
- However, since the actual weights may vary around the predicted ones, we need another model to capture this variation
- We assume that the actual weights y follows a normal distribution, with mean $\hat{y}$ (the predicted weights) and standard deviation σ:

$$y \sim N(\hat{y}, \sigma)$$

- **Q:** What does this model mean?
- **A:**
  - the values of y near $\hat{y}$ are most probable
  - the decrease in probability around $\hat{y}$ is governed by σ

Step 2: defining a model

- Our complete model includes two models, a linear model and a normal distribution
- **Q:** What are the parameters the complete model has?

Step 2: defining a model

- Our complete model includes two models, a linear model and a normal distribution
- **Q:** What are the parameters the complete model has?
- **A:**
  - the slope, β_1
  - the intercept, β_0
  - the mean, $\hat{y}$
  - the standard deviation, σ

Step 2: defining a model

- Our complete model includes two models, a linear model and a normal distribution
- **Q:** What are the parameters the complete model has?
- **A:**
  - the slope, β_1
  - the intercept, β_0
  - ~~the mean, $\hat{y}$~~
  - the standard deviation, σ
- The mean ($\hat{y}$) is not a parameter, since it is determined by the linear model, given the slope (β_1), the intercept (β_0), and the height (x)

Step 3: specifying a prior distribution

- Generally, we might be able to:
  - inform the prior with previously conducted, and publicly verifiable, research on weights and heights of the target population
  - argue for a modestly informed prior based on consensual experience of social interactions
- For simplicity, in this example we will:
  - use two uniform distributions for the slope (β_1) and intercept (β_0), both of which across a vast range of possible values and centered at zero
  - use a uniform distribution for the standard distribution (σ), which extends from zero to a huge value

Step 4: inferring the posterior distribution

- The right panel of Figure 2.5 (see next page) shows the posterior distribution on the slope parameter, β_1
- **Q:** What can you see from the distribution?
  - what is the most credible value of the slope?
  - what does this most credible value mean?
  - what is the uncertainty in the slope?
  - is there a positive relationship between weight and height?

Figure 2.5

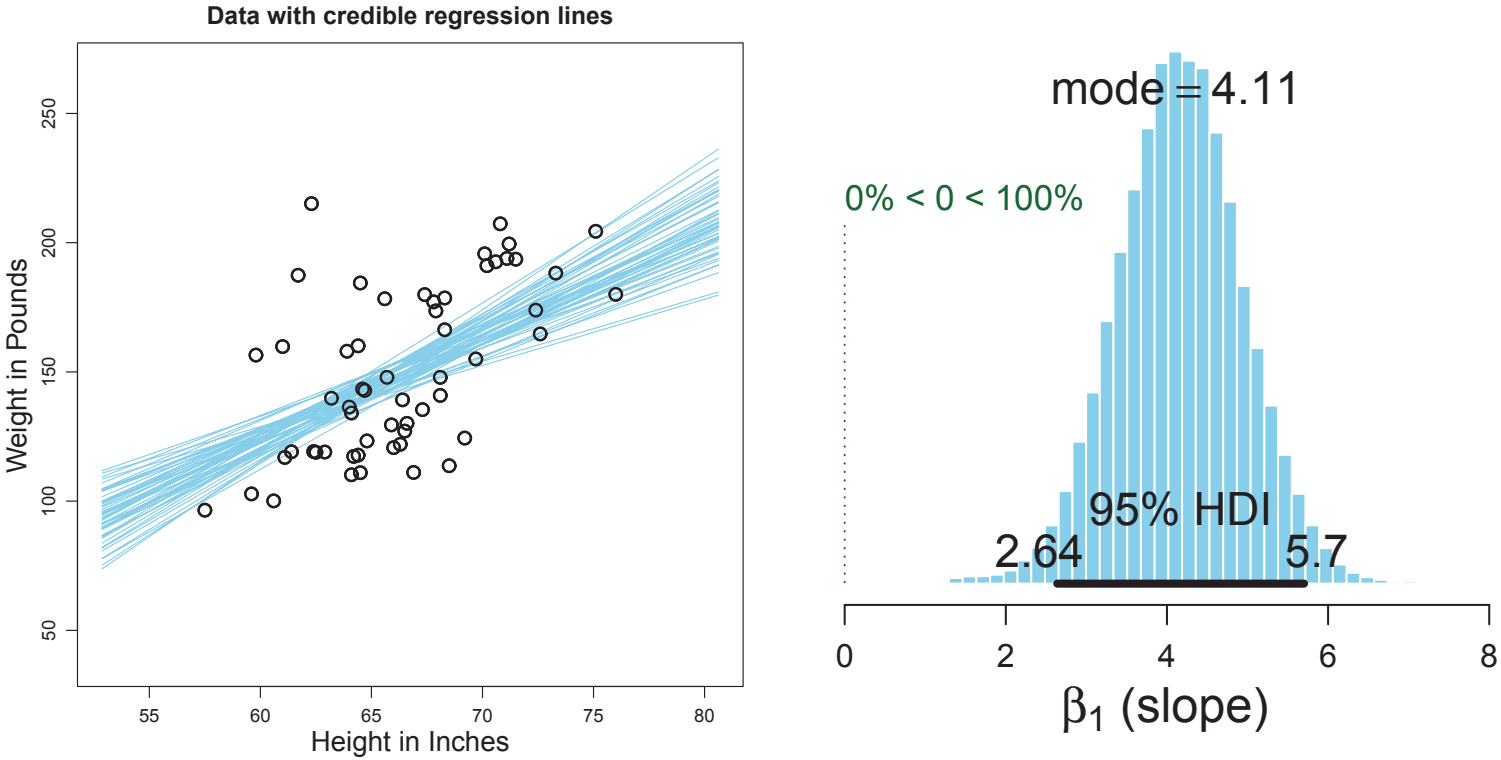


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Highest Density Interval (HDI)

- One way to summarize the uncertainty is by marking the span of values that are most credible
- This is called the highest density interval (HDI), where values within HDI are more credible (have higher probability "density") than values outside HDI
- The 95% HDI is marked by the black bar on the floor of the posterior distribution
- **Q:** Which is better? A wide HDI or narrow HDI?

Highest Density Interval (HDI)

- One way to summarize the uncertainty is by marking the span of values that are most credible
- This is called the highest density interval (HDI), where values within HDI are more credible (have higher probability "density") than values outside HDI
- The 95% HDI is marked by the black bar on the floor of the posterior distribution
- **Q:** Which is better? A wide HDI or narrow HDI?
- **A:** A narrow HDI, since it indicates stronger certainty

Credible regression lines

- One of the key ideas of bayesian Inference is to provide a distribution, rather than a point estimate, of the parameters
- This is why we care more about posterior distribution of the slope, instead of say, just the mode
- For the same reason, we care more about the credible regression lines:

$$\hat{y} = \beta_1 x + \beta_0,$$

  where β_1 and β_0 take values from the corresponding HDIs
- The left panel of Figure 2.5 (see next page) shows the credible regression lines

Figure 2.5

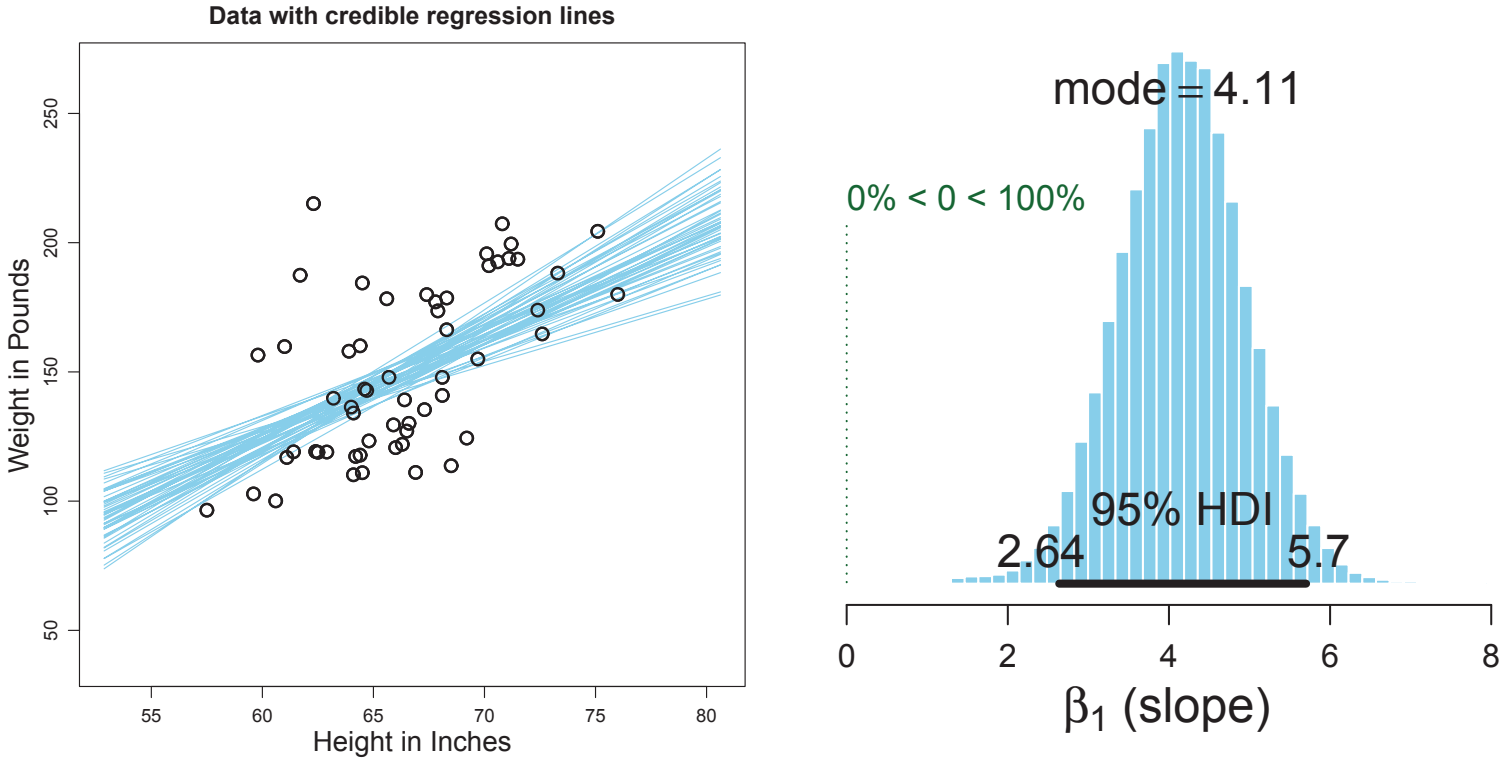


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Step 5: checking the posterior distribution

- The goal is to check whether the model (i.e., posterior distribution) fits the data reasonably well
- One method is to plot a distribution of predicted data from the model, with its most credible parameter values, against the actual data
- **Q:** What does it mean for this example?

Step 5: checking the posterior distribution

- The goal is to check whether the model (i.e., posterior distribution) fits the data reasonably well
- One method is to plot a distribution of predicted data from the model, with its most credible parameter values, against the actual data
- **Q:** What does it mean for this example?
- **A:**
  1. take the value of the three parameters, β_1, β_0, and σ, from their HDIs
  2. plug them into the model
  3. generate y values (weights) based on x values (heights)
- The comparison for this example is shown in Figure 2.6 (see next page), which shows that the model fits the data well

Figure 2.6

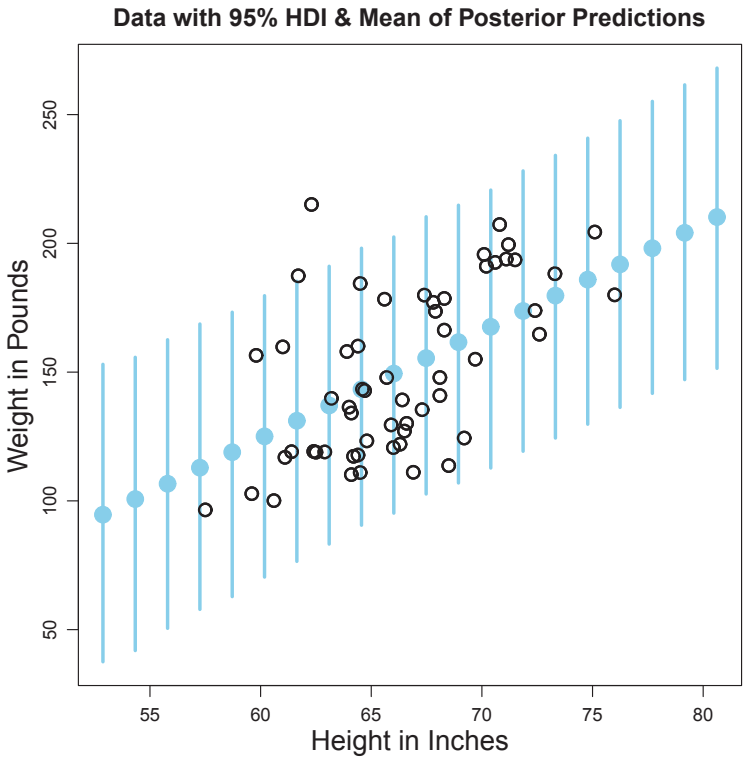


Figure 2.6: The data of Figure 2.5 are shown with posterior predicted weight values superimposed at selected height values. Each vertical bar shows the range of the 95% most credible predicted weight values, and the dot at the middle of each bar shows the mean predicted weight value. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.