Bayesian Methods for Data Science (DATS 6450 - 11)

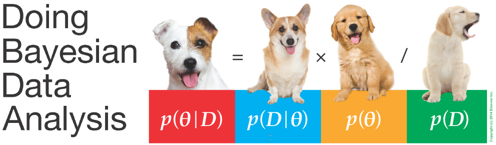
Introduction: Credibility, Models, and Parameters

Yuxiao Huang

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

September 4, 2019

Reference


Doing Bayesian Data Analysis
$p(\theta|D)$ = $p(D|\theta)$ × $p(\theta)$ / $p(D)$
Copyright (c) 2014 Elsevier Inc.

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. Bayesian inference is reallocation of credibility across possibilities

2. Possibilities are parameter values in descriptive models

3. The steps of bayesian data analysis

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

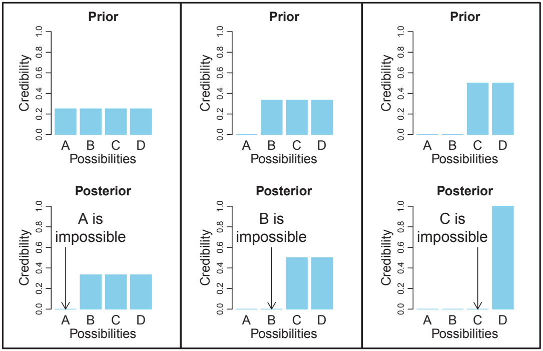


Figure 2.1: The upper-left graph shows the credibilities four possible causes for an outcome. The causes, labeled A, B, C and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset, hence all have prior credibility of 0.25. The lower-left graph 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. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

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

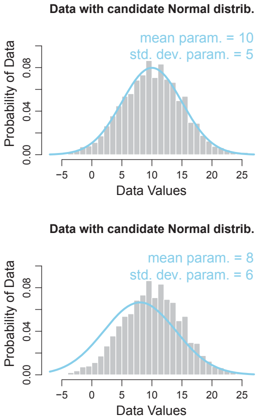


Figure 2.4: The two graphs show the same data histogram but with two different candidate descriptions by normal distributions. Bayesian analysis computes the relative credibilities of candidate parameter values. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS,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

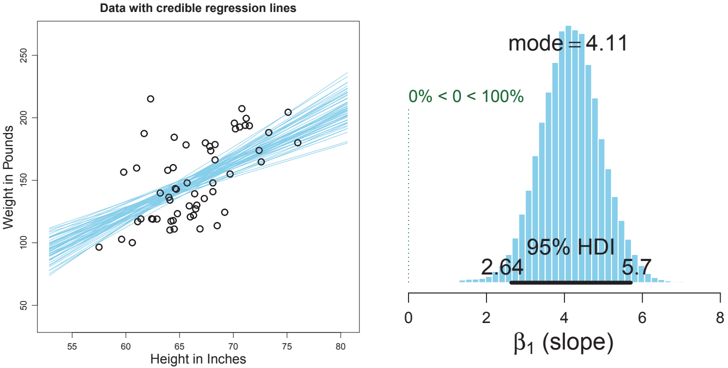


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). 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

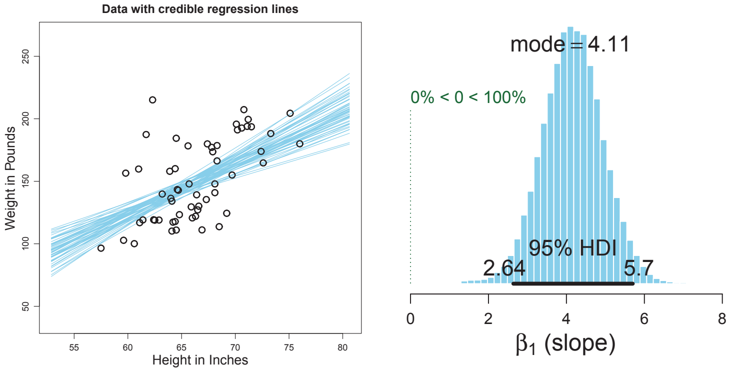


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). 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

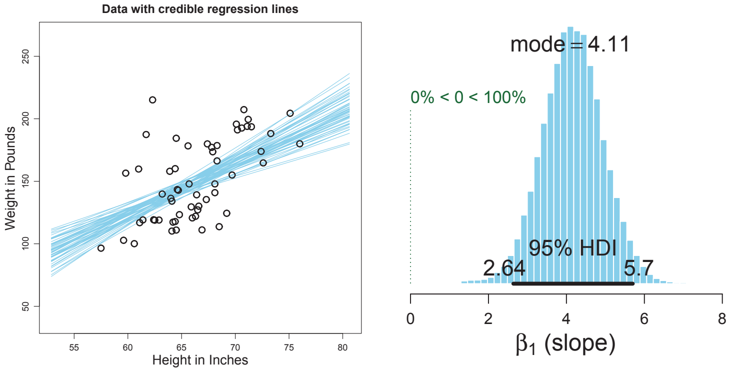


Figure 2.5: Data are plotted as circles in the scatter plot of the left panel. The left panel also shows a smattering of credible regression lines from the posterior distribution superimposed on the data. The right panel shows the posterior distribution of the slope parameter (i.e., β_1 in Eqn. 2.1). 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

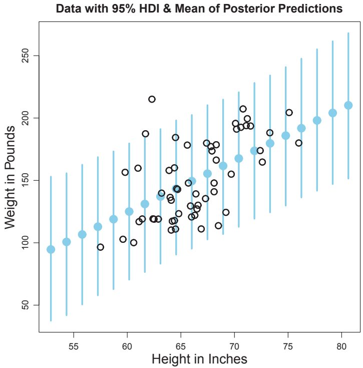


Figure 2.6: The data of Figure 2.5 are shown with posterior predicted weight values superimposed at selected height values. Each vertical bar shows the range of the 95% most credible predicted weight values, and the dot at the middle of each bar shows the mean predicted weight value. Copyright © Kruschke,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