Bayesian Statistics Review

Carlos Carrillo-Gallegos
Main Source: Doing Bayesian Data Analysis by John Kruschke

September 2022

1 Introduction

This document is my attempt to explain the fundamentals of Bayesian statistics to myself. It's not original material, of course. I'll be working through "Doing Bayesian Data Analysis" by Kruschke and trying to write it in my own words/in a way I can better comprehend myself. For many example concepts given in the book, I write my own examples in this document.

2 Credibility, Models, and Parameters

2.1 Bayesian Inference Is Reallocation Of Credibility Across Possibilities

Bayesian Inference is the process of using new information to adjust the probabilities (credibilities) of different explanations for a possible outcome.

For example, say you enter your room and see your laptop is not on your desk. There are many possible explanation that you may consider: A) someone broke in and stole it. B)Your friend moved it somewhere else in the apartment. C) You left it in the living room, not your room. All 3 possible explanations have a probability associated with them. The process of gathering more information to rule out some options is called **Bayesian Inference**.

Let's say you think there is an equal chance for all three events to be the explanation, such that A, B, and C all have a $\frac{1}{3}$ probability of occurring. This is called the **Prior Distribution**, as it is the initial condition from which the probabilities will change.

It turns out your friend wasn't at home since you last left, ruling out option B. That shifts the probability distribution, giving $\frac{1}{2}$ probability to both A) and C). This current distribution is known as the **Posterior Distribution**, because it is the distribution obtained after using newly obtained information. Moving forward, the Posterior will become the Prior for any information obtained afterwards.

You check your security footage and see nobody entered your home, thus ruling out option A). The Prior distribution is the former Posterior distribution

(the one with $\frac{1}{2}$ probability for A and C), and the Posterior distribution now has full probability for option C. The laptop must be in the living room.

This whole process is Bayesian Inference. If you had instead at first gone to check the living room before inspecting the other possibilities and found the laptop, that is know as **Bayesian Exoneration**. In this case, you eliminated any probability of A or B being correct by determining that option C was correct.

2.1.1 Inferences are Probabilistic

It's also important to note that real data is not as clean as the example above. In reality, it is unlikely to be able to fully rule out a possibility, and as such, **inferences are probabilistic**, meaning that we can have more confidence in one theory given more and more observations, but we're unlikely to be able to completely rule out all other possibilities.

2.2 Possibilities are Parameter Values in Descriptive Models

Consider a drug trail in which one group takes real blood pressure medication and another takes a placebo pill. We want to know the difference in typical blood pressure between the two groups, and we want to know how certain we are about our predictions. Each possible difference between the two groups' blood pressures is a possibility, and all of these possibilities are *describing the data*.We want to know which description is more or less *credible*.

Take figure one for example (from Kruschke's book). The histogram itself is the data we have, and the normal distribution is the attempt to describe the data with a function.

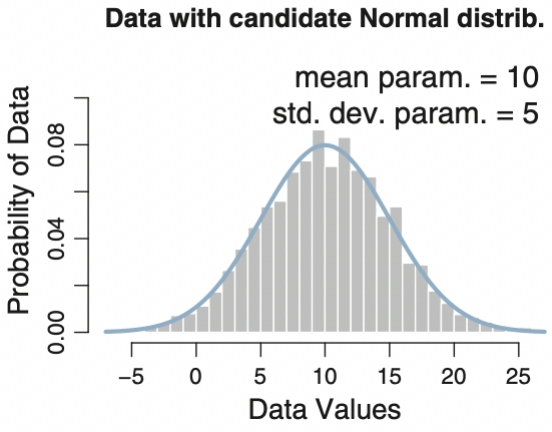

Figure 1: Normal distribution fitted to data

This, and all other fits to try and describe data, are functions governed by parameters. The parameters are associated with different possible explanations.

For a normal distribution, the parameters are the mean, which essentially controls where the fit is centered, and the standard deviation, which controls its width.

Two important factors to consider when judging parameters and their effectiveness are their effect and appearance. Firstly, the parameters should have a tangible meaning with respect to the distribution. The mean and standard deviation make sense as parameters for normal distribution because we can clearly see how they relate to the shape of the distribution. Secondly, the describing function should at least somewhat resemble the data.

2.3 Steps of Bayesian Data Analysis

From Kruschke:

1. Identify relevant data
2. Define a descriptive model (ex: a normal distribution) for the data, ensuring that the parameters are meaningful.
3. Specify a prior distribution.
4. Use Bayesian inference to change the weights (credibility) of the parameters.
5. Check that posterior function mimics data sufficiently well. If not, a different describing model is needed.

3 Probability

3.1 Sample Spaces and Degree of Belief

We define the **Sample Space** as the set of possible outcomes of an event. For example, the sample space of a coin flip is that it lands on heads, or on tails.

A bit of notation. We use θ to signify the probability of an event occurring, ie. the probability of a fair coin landing on heads is $\theta = 0.5$.

We also use $p(\theta)$ to signify **degree of belief**, or how confident we are in said probability. Our degree in belief would change if we were uncertain about the coin being fair, for instance.

We define **long-run relative frequency** as the amount of times we would expect an outcome if the event were to occur several times over. In practice, this can be defined by simulations or mathematically. For example, you could derive the long-run relative frequency of a fair coin landing on heads by simulating a coin flip in a program several iterations over, and seeing how many times the coin lands on heads. Mathematically, it's more simple. Since you know there are two outcomes to a coin flip, and they are equally likely, the relative frequency

is 1/2. Of course, that's just it's probability.

3.1.1 Three Conditions of Probability

1. Probability value must be non-negative
2. Sum of probabilities must = 1.0
3. For mutually exclusive events, their probabilities should sum to give the probability that either one or the other occurs.

3.2 Probability Distributions

A **probability distribution** is just a list of possible outcomes and their associated probabilities.

We often work with discrete probability distributions. For example, the probability distribution of rolling a die has 6 discrete outcomes. There can be continous probability distributions. For example, if you sample a group of x people's ages, the distribution will be technically continuous because there are x number of ages, going down into minutes and seconds. In practice, we can also discritize a continous distribution by blocking out buckets in which to fit data. In the age example, we might discretize age by years.

We refer to the probability of a discrete outcome as the **probability mass**, the proportion of the total possible outcomes that one specific outcome holds.

We can also think in terms of **probability density**, which is the probability of an outcome over its "width", how many "units" the outcome is spread over. For example, if you are looking at a large group of people's ages, let's say there is a 0.13 probability that someone in that group is between $55 - 60$ years old. The probability density would be the $\frac{0.13}{5.0}$ (0.026). This indicates that in this particular interval, there is a density of 0.026 probabilities per age. Note that probability densities can be greater than 1.0 (imagine a very large probabiltiy and a very small width).

3.3 Probability Density Functions

3.3.1 Deriving the function

Remember that for any distribution, the sum of all the probability masses should equal 1.0, as shown in equation (1).

$$\sum_i p([x_i, x_i + \Delta x]) = 1.0 \tag{1}$$

Where i is the index, Δx is the length of an interval, and $p([x_i, x_i + \Delta x])$ is the probability mass over an interval.

We can express this in terms of probability density by dividing and multiplying the main term by Δx, as seen in equation (2). We divide to give us the probability density and multiply to ensure the equation is the same.

$$\sum_i \Delta x \frac{p([x_i, x_i + \Delta x])}{\Delta x} = 1.0 \tag{2}$$

In the infinitessimal limit, we can approximate this as an integral, writing Δx as a dx, and $p([x_i, x_i + \Delta x])$ as $p(x)$, such that:

$$\int p(x)dx = 1.0 \tag{3}$$

Here, we define $p(x)$ as the probability density, but the same notation can be used to define probability mass. Ultimately, the context will be important in determining what is being used.

3.3.2 Normal Probability Density Function

Any function with non-negatives that integrates to 1.0 can be described as a probability function. The most famous is the normal distribution (Gaussian Function).

The Gaussian is governed by 2 paramaters, μ, the mean, and σ, the standard deviation. It is exrpessed as:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right) \tag{4}$$

In Figure 2, a plot of the normal probability density is shown. Note that probability density can exceed 1.0, but only over a very small interval. The area under the curve will still be = 1.0.

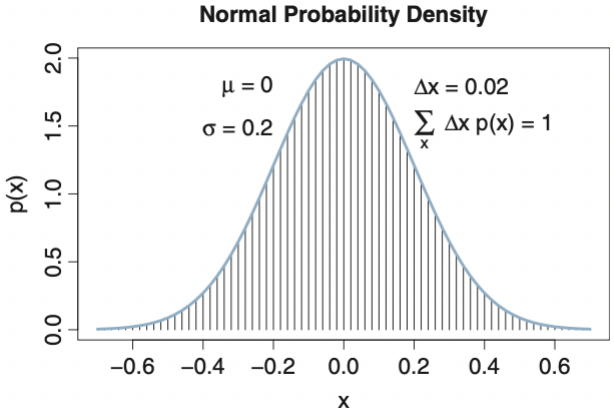

Figure 2: Normal probability density

3.4 Mean and Variance

The mean of a probability distribution can be thought of as the expected outcome if we repeatedly sampled x. A good example of this is rolling die. Each result has a probability of $\frac{1}{6}$, so the long-run average (a.k.a the expected value) is $1(\frac{1}{6})+2(\frac{1}{6})+...6(\frac{1}{6}) = 3.5$. So, 3.5 is the **mean** of the probability distribution.

We can formalize this example into our integral form from above by simply multiplying the probability density function(PDF) by x. In practice, this equation gives us our expected value, $E[x]$, or mean. Given the interval and a PDF $p(x)$, we would simply compute the integral over the interval to get the expected value.

$$E[x] = \int xp(x)dx \tag{5}$$

The **variance** is meant to encapsulate how far the data is from the mean, and is defined as follows:

$$var = \int p(x)(x - E[x])^2 dx \tag{6}$$

Where $(x = E[x])^2$ is the squared difference between x and the mean.

Note how the form of equation (6) is very similar to that of equation (5), just replacing x with $(x - E[x])^2$. This indicates that the variance is the average value of $x - E[x])^2$.

The square root of the variance is the standard deviation of the distribution. It is also known as the **root mean squared deviation (RMSD)**

3.5 Conditional Probability

Suppose you want to know the probability of someone having blond hair given that they have blue eyes. The probability that someone has blond hair, in this case, is conditional on them having blue eyes.

To compute this, you would find the probability that a person in the distribution has blue eyes; let's say it's 0.36, so that 36% of the distribution has blue eyes. We must also know the probability of someone having both blue eyes and blond hair, let's say 0.16.

Note that the probability of someone having both blue eyes and blond hair **IS NOT** the same as the probability of someone having blond hair given that they have blue eyes. In the later case, which is what we are interested in, the probability of the person having blodn hair is dependent on the person having blue eyes.

Finally, to compute the value, we would simply divide $\frac{0.16}{0.36}$, to find that 45% of the blue eyes population has blond hair.

How can we formalize this definition?

What we did here was define $p(h|e)$, the probability of a hair color, h, given an eye color, e. We defined it as $p(e, h)/p(e)$, the probability of someone having

blue eyes and blond hair divided by the probability of someone having blue eyes. This is essentially the definition of conditional probability.

We can extend this to non-specific parameters c and r, as shown below, where we look for the probability of c given r

$$p(c|r) = \frac{p(r,c)}{p(r)} \tag{7}$$

Kruschke simply describes this as "the probability of c given r is the probability that they happen together relative to the probability that r happens at all."

Note that we can define $p(r)$ as the sum of $p(r,c)$ over all values c. For example, $p(e)$ is the sum of all $p(e,h)$, so the probability of someone having blue eyes is the probability of someone having blue eyes and black hair plus the probability of someone having blue eyes and blond hair, so on for all hair colors.

We can define this sum as an integral of the interval c, as such:

$$p(c|r) = \frac{p(r,c)}{p(r)} = \frac{p(r,c)}{\int p(r,c)dc} \tag{8}$$

4 Bayes' Rule

4.1 Definition of Bayes' Rule

Bayes' rule is the mathematical interpretation of Bayesian inference that we discussed in section 2. Through some algebra on the conditional probability equation (7), we define Bayes' rule as follows:

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)} \tag{9}$$

We can also redefine p(r) as we did above, in terms of a sum rather than an integral:

$$p(c|r) = \frac{p(r|c)p(c)}{\sum_{c^*} p(r|c^*)p(c^*)} \tag{10}$$

where c^* takes in all possible c values, as opposed to c in the numerator, which corresponds to a specific parameter.

4.2 Intuition behind Bayes' Rule

First, we define **joint probability** as the probability that both r and c occur, not dependent on each other; $p(r,c)$.

Next, we define **marginal probability** as the probability that one thing will happen; in this case, $p(c)$ or $p(r)$.

Notice that the numerator of Bayes' rule is the joint probability, and the denominator is the marginal probability, $p(r)$.

By focusing on one row value, r, Bayes' rule determines the conditional probability $p(c|r)$ from the marginal probability $p(c)$. Bayes' rule, when focusing on one row, essentially normalizes the probabilities of that row (by dividing by the total probability). More practically, when holding for one variable, like eye color, Bayes' rule normalizes the probabilities of possible hair colors.

4.2.1 Disease Testing Example

Suppose you are testing a random person in a population for a disease. The characteristics of the test are as follows:

1. The test has a 99% hit rate, meaning it detects the disease when it is present 99% of the time
2. The test has a false alarm rate of 5%, meaning that 5% of the time, when the disease is not there, the test identifies it anyway.
3. the incidence of the disease in the population is $\frac{1}{1000}$, or 0.001

If we use θ to indicate true presence of the disease, with θ =Yay indicating its absence and θ = Nay indicating its presence, then $p(\theta$ =Yay$) = 0.999$ and $p(\theta$ =Nay$) = 0.001$. **These probabilities represent our prior beliefs**.

The test results are new information that adjusts our prior beliefs. If we use $T = +$ and $T = -$ to represent a positive and negative result, respectively, then the hit rate can be expressed as $p(T = +|\theta$ =Nay$) = 0.99$ and the false alarm can be expressed as $p(T = +|\theta$ =Yay$) = 0.05$.

When testing a patient looking for the disease, we are searching for the posterior probability $p(\theta$ =Nay$|T = +)$; what is the probability the patient has the disease given their result was positive?

The situation is encapuslated in Figure 3, a table from Kruschke. Note in the table that Yay =:) and Nay =: (.

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

	Disease		
Test result	$\theta =$ 😞 **(present)**	$\theta =$ 😊 **(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.

Figure 3: Disease Probability Table

Each quadrant in the figure show the four possible joint probabilities, and the marginal probabilities for the disease and test result are shown on the edges.

We can use Bayes' rule to determine the probability of someone having the disease if they test positive!

$$p(\text{Nay}|+) = \frac{p(+|\text{Nay})p(\text{Nay})}{\sum_\theta p(+|\theta)p(\theta)} \tag{11}$$

The probability of the patient having the disease given that they tested positive is the probability that they tested positive given that they have the disease (the hit rate) times the base rate of the disease, normalized by marginal probability of testing positive. The marginal probability of testing positive is the probability that they tested positive given they have the disease, $p(+|\text{Nay})$, times the base rate of the disease being present, $p(\text{Nay})$ plus the same factor but for the disease being absent (this is the summing term in the denominator).

We should note that actually computing the probability in equation 11 results in a 1.9% chance that the patient has a disease. That is relatively low given that the test had a 99% hit rate! So 98.1% of people who test positive will not end up having the disease.

Why is this so low? It's a result of the extremely low base rate of the disease and the false alarm rate.

The process of determining the probability of disease has been an exercise in Bayesian inference! The base rates of disease are the prior probabilities. The test's hit rate and false alarm rate were the conditional probabilities that pointed to which row(parameter) to focus on (+ or -). Once the analysis narrowed to that row, the conditional probabilities of disease could be computed for that given row via Bayes' rule. These conditional probabilities are the posterior probabilities that were re-allocated credibility given the new information we received.

4.3 Bayes', Parameters, and Data

In practice, Bayes' rule is most useful when the rows are data values and the columns are model parameters.

We define a **model** as the probability that we get a certain data value given the parameters, times the probability of the respective parameters. So,

$$p(\text{data values}|\text{parameter values}) * p(\text{parameter values}) \tag{12}$$

This equation can be used to find the real probability of interest, the probability that the parameter values are true given the data values, p(parameter values|data values)

The situation is clearly summarized in Figure 4, Table 5.5 from Kruschke:

Given this new notation with D representing data and θ representing parameters, we can restate Bayes' rule as follows:

$$p(\theta|D) = p(D|\theta)\frac{p(\theta)}{p(D)} \tag{13}$$

Where

1. $p(\theta|D)$ is the **posterior**
2. $p(D|\theta)$ is the **likelihood**
3. $p(\theta)$ is the **prior**
4. $p(D)$ is the **evidence** or **marginal likelihood**

And the denominator, $p(D)$, can be expressed as:

$$p(D) = \sum_{\theta^*} p(D|\theta^*)p(\theta^*) \tag{14}$$

Noting that θ^* is not a specific value of θ, like θ in the numerator.

Table 5.5 Applying Bayes' rule to data and parameters

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

Figure 4: Bayes' Rule Table

The prior represents the probability of the parameters without information from the data, D. The posterior is the probability of the parameters taking into account information from the data. The likelihood is the probability that the data D could come from a model with parameters θ. And the evidence, $p(D)$, is the overall probability of the data according to the model. The evidence is determined by averaging across all parameters weighed by the priors of those parameters (equation (14)).

We have been discussing Bayes' rule as it stands for discrete distributions, but can easily extend it to continuous distributions by expressing the evidence as an integral rather than a sum:

$$p(D) = \int p(D|\theta^*)p(\theta^*)d\theta^* \tag{15}$$

Again where θ^* is distinct from θ in that it is not a specific parameter.