Introduction to Bayesian Statistics

Egor Howell

What We Will Cover

- Understanding Probability: An Overview of Different Types
- Exploring Bayes' Theorem: Derivation and Significance
- Applying Bayes' Theorem: Practical Examples
- The Concept of Bayesian Updating: An Introduction
- Conjugate Priors in Bayesian Analysis

Intro

- Conditional probability and Bayes' theorem are fundamental ideas in statistics that even laymen have heard of. Bayes' theorem also gives rise to a separate branch of statistics, namely Bayesian inference.
- In Data Science we mainly deal and work in a Frequentist world and so we are, in my opinion, not fully aware of the Bayesian principles.
- However, I often find the word 'Bayes' being thrown around, but often is mis-represented to what it actually means.

Probability & Types

Marginal Probability

Marginal probability is what most people mean when they say/refer to as probability. It is just the probability of that certain event occurring.

For example, the marginal probability of flipping a head, P(H), on a coin is simply 0.5:

$$P(H) = 0.5$$

Joint Probability

- Let's take it one step further, what is the probability of flipping two heads? This is referred to as the joint probability as it is joining two events together.
- To solve this problem, we can just list the possible outcomes when flipping two coins: ***{H,H}, {H,T}, {T,H}, {T,T}***. Therefore, the probability of flipping two heads is *0.25*:

$$P(H \cap H) = 0.25$$

Commutativity

Another important property is that joint probabilities are commutative, which means:

$$P(A \cap B) = P(B \cap A)$$

Conditional Probability

- Conditional probability is when we determine the probability ‘given’ some condition/event occurring.
- What is the probability of picking the **3 diamonds** from a deck of hands given we have chosen a **red card**?

Conditional Probability: Example

- What is the probability of picking the **3 diamonds** from a deck of hands given we have chosen a **red card**?
- Well, the probability of picking a 3 of diamonds, ***P(3D)***, is:

$$P(3D) = \frac{1}{52}$$

Conditional Probability: Example

- What is the probability of picking the **3 diamonds** from a deck of hands given we have chosen a **red card**?
- And the probability of choosing a red card, ***P(R)***, is:

$$P(R) = \frac{26}{52} = \frac{1}{2}$$

Conditional Probability: Example

- What is the probability of picking the **3 diamonds** from a deck of hands given we have chosen a **red card**?
- Therefore, the probability of choosing the **3 diamonds** given we have a red card, **P(3D | R)**, is then:

$$P(3D \mid R) = \frac{\frac{1}{52}}{\frac{1}{2}} = \frac{1}{26}$$

Conditional Probability

The official mathematical definition for two events A and B is:

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

Bayes’ Theorem

Bayes Theorem

- Rearranging the conditional probability equation we get:

$$P(A \cap B) = P(B)P(A \mid B)$$

- Then subbing in the conditional formula again (remember joint distributions are commutative):

$$P(A)P(B \mid A) = P(B)P(A \mid B)$$

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

Bayes' Theorem

- *P(A)* is known as the prior, which is what we believe the probability to be before we observe our data. This is a marginal probability of this event.
- *P(B)* is the probability of observing the data/event on its own. This is a marginal probability of this event. It is sometimes referred to as the normalizing constant.
- *P(B | A)* is the probability given what we 'believe', which is known as the likelihood.
- *P(A | B)* is the posterior which is the probability of our 'belief' after we observe our data.

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

Law of Total Probability

- The final formula we will discuss is the Law of total probability:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

- One can think of this summation in two different ways:
  - The sum of all the overlapping regions A covers B.
  - The weighted average of A on B

Bayes' Theorem Example

- Let's say I have two decks of cards: One is a normal deck, ***D_1***, and the other is a deck with just the red cards (diamonds and hearts), ***D_2***.
- I randomly select a deck and pull out the 3 of diamonds (***3D***). What is the probability that this 3 of diamonds came from the normal deck (***D_1***)?

$$P(D_1) = \frac{1}{2}$$

$$P(D_2) = \frac{1}{2}$$

Bayes' Theorem Example

Let's start with stating the prior probabilities of randomly picking either deck 1, ***P(D_1)***, or deck 2, ***P(D_2)***. This is simply 50–50 because it is random:

$$P(D_1) = \frac{1}{2}$$

$$P(D_2) = \frac{1}{2}$$

Bayes' Theorem Example

Now let's calculate the likelihoods:

$$P(3D \mid D_1) = \frac{1}{52}$$

$$P(3D \mid D_2) = \frac{1}{26}$$

Bayes' Theorem Example

Then, we calculate the probability of observing the 3 of diamonds by using the Law of total probability:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

$$P(3D) = P(D_1)P(3D \mid D_1) + P(D_2)P(3D \mid D_1) = \frac{3}{104}$$

Bayes' Theorem Example

Combining this together using Bayes' theorem:

$$P(D_1 \mid 3D) = \frac{P(D_1)\, P(3D \mid D_1)}{P(3D)} = \frac{1}{3}$$

Bayesian Updating

Bayesian Updating

- We can use Bayes' theorem to update our hypothesis when new evidence comes to light.
- For example, given some data ***D*** which contains the one ***d_1*** data point, then our posterior is:

$$P(H \mid d_1) = \frac{P(H)P(d_1 \mid H)}{P(d_1)}$$

- Let's say we now acquire another data point ***d_2***, so we have more evidence to evaluate and *update* our belief (posterior) on. However, our prior now becomes our old posterior because this represents our new prior belief of our hypothesis:

$$P(H) = P(H \mid d_1)$$

Bayesian Updating

We normally omit the denominator ***P(D)*** as it is just a normalising constant to make the probabilities sum to 1. [There is great thread here from Stat Exchange that explains this very well.]

$$Posterior \propto Likelihood \times Prior$$

Bayesian Updating Example

Let’s say I have three different dice with three different number ranges:

- ***Dice 1: 1–4***
- ***Dice 2: 1–6***
- ***Dice 3: 1–8***

We randomly select a dice and do three subsequent rolls with that given dice. Using these rolls (data), we can compute how likely we picked up either dice 1, 2 or 3 after reach role (posterior).

Bayesian Updating Example: First Roll

In the first roll, we get the number 4. What is the probability that we selected dice 1, 2 or 3?

We can compute this using Bayes' theorem as follows:

$$P(Dice\ 1 \mid roll\ 4) = \frac{P(Dice\ 1)P(roll\ 4 \mid Dice\ 1)}{P(roll\ 4)} = \frac{0.33 \times 0.25}{0.18} \approx 0.46$$

$$P(Dice\ 2 \mid roll\ 4) = \frac{P(Dice\ 2)P(roll\ 4 \mid Dice\ 2)}{P(roll\ 4)} = \frac{0.33 \times 0.167}{0.18} \approx 0.31$$

$$P(Dice\ 3 \mid roll\ 4) = \frac{P(Dice\ 3)P(roll\ 4 \mid Dice\ 3)}{P(roll\ 4)} = \frac{0.33 \times 0.125}{0.18} \approx 0.23$$

Probability of the data (normalising value), P(roll 4), is just the sum of the likelihood and prior products.

Bayesian Updating Example: Second Roll

Using the same dice, we now roll a second time and get a 2. However, we have new priors, which are the calculated posteriors above where we rolled a 4.

$$P(Dice\ 1 \mid roll\ 2) = \frac{P(Dice\ 1 \mid roll\ 4)P(roll\ 2 \mid Dice\ 1)}{P(roll\ 2)} = \frac{0.46 \times 0.25}{0.196} \approx 0.59$$

$$P(Dice\ 2 \mid roll\ 2) = \frac{P(Dice\ 2 \mid roll\ 4)P(roll\ 2 \mid Dice\ 2)}{P(roll\ 2)} = \frac{0.31 \times 0.167}{0.196} \approx 0.26$$

$$P(Dice\ 3 \mid roll\ 2) = \frac{P(Dice\ 3 \mid roll\ 4)P(roll\ 2 \mid Dice\ 3)}{P(roll\ 2)} = \frac{0.23 \times 0.125}{0.196} \approx 0.15$$

Bayesian Updating Example: Third Roll

We roll a third time with our chosen dice and get a 5. Using our previous posterior as our new prior, the probabilities are now:

$$P(Dice\ 1 \mid roll\ 5) = \frac{P(Dice\ 1 \mid (roll\ 4, roll\ 2))\ P(roll\ 5 \mid Dice\ 1)}{P(roll\ 5)} = \frac{0.59 \times 0}{0.0622} = 0$$

$$P(Dice\ 2 \mid roll\ 5) = \frac{P(Dice\ 2 \mid (roll\ 4, roll\ 2))\ P(roll\ 5 \mid Dice\ 2)}{P(roll\ 5)} = \frac{0.26 \times 0.167}{0.0622} \approx 0.7$$

$$P(Dice\ 3 \mid roll\ 5) = \frac{P(Dice\ 3 \mid (roll\ 4, roll2))\ P(roll\ 5 \mid Dice\ 3)}{P(roll\ 5)} = \frac{0.15 \times 0.125}{0.0622} \approx 0.3$$

The value with the highest posterior is known as the Maximum a posteriori (MAP). This is analogous to the Maximum Likelihood but for Bayesian statistics and is the mode value of the posterior distribution.

Conjugate Priors

Bayesian Issues

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)}$$

$$P(D) = \int P(H)P(D|H)\, dH$$

This integral is often **intractable**. This basically means it is very computationally expensive or it doesn't have a **closed form solution**. I have linked **here** a StatExchange thread that explains why it is intractable.

Bayesian Conjugate Priors

Conjugate priors is one way of getting around the intractable integral issue in Bayesian inference. This is when both the prior and posterior are of the same distribution. This allows us to simplify the expression of calculating the posterior. In the next section we will show this phenomenon mathematically.

Binomial and Beta Contingency

One of the simplest and common conjugate distribution pair is the Beta (prior) and Binomial (likelihood).

Beta Distribution:

- Referred to as the distribution of probabilities because its domain is bounded between 0 and 1.
- Conveys the the most probable probabilities about the success of an event.

Its probability density function (PDF) is written as:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

Here *x* is bounded as $0 \leq x \leq 1$, so it can easily be interpreted as probability and $B(\alpha, \beta)$ is the Beta function.

Binomial and Beta Contingency

Binomial Distribution:

- Conveys the probability of a certain number of successes ***k*** from ***n*** trials where the probability of success is ***x***.

$$f(k) = \binom{n}{k} x^k (1-x)^{n-k}$$

Crucial Point

- The key difference between the Binomial and Beta distributions is that for the Beta distribution the probability, ***x***, is a random variable, however for the Binomial distribution the probability, ***x***, is a fixed parameter.

Relation To Bayes

We can rewrite Bayes' theorem using the probability of success, ***x***, for an event and the data, ***k***, which is the number of successes we observe:

$$P(x \mid k) = \frac{P(k \mid x)P(x)}{P(k)} = \frac{P(k \mid x)P(x)}{\int P(k \mid x)\ P(x)\ dx}$$

Our posterior is basically the probability distribution over all the possible probabilities of the success rate. In other-words, the posterior is a Beta distribution.

Relation To Bayes

We can express the above equation using the Binomial distribution as our likelihood and the Beta distribution as our prior:

$$P(x \mid k) = \frac{\binom{n}{k} x^k (1-x)^{n-k} \ \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}}{\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \ \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \, dx}$$

Yeah, doesn't look that nice. Nevertheless, we are now going to simplify it:

$$P(x \mid k) = \frac{\binom{n}{k} \frac{1}{B(\alpha,\beta)} \ x^{k+\alpha-1}(1-x)^{n-k+\beta-1}}{\binom{n}{k} \frac{1}{B(\alpha,\beta)} \ \int_0^1 \ x^{k+\alpha-1}(1-x)^{n-k+\beta-1} \, dx}$$

Relation To Bayes

Some of you may notice something special about that integral. It is the definition of the [Beta function](#)!

$$P(x \mid k) = \frac{\binom{n}{k} \frac{1}{B(\alpha,\beta)} \, x^{k+\alpha-1}(1-x)^{n-k+\beta-1}}{\binom{n}{k} \frac{1}{B(\alpha,\beta)} \int_0^1 x^{k+\alpha-1}(1-x)^{n-k+\beta-1} \, dx}$$

$$B(k+\alpha, n-k+\beta) = \int_0^1 x^{k+\alpha-1}(1-x)^{n-k+\beta-1} \, dx$$

Therefore, the final form of our posterior is:

$$P(x \mid k) = \frac{x^{k+\alpha-1}(1-x)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}$$

Why?

You may be scratching your head wondering why I have taken you through this awful derivation just to get another version of a Beta distribution?

What this beautiful result shows us, is that to do a Bayesian update we no longer need to compute the product of the likelihood and prior. This is computationally expensive and sometimes not feasible as I discussed earlier. We can now just use simple addition!

Real Life Example

Example: Baseball Batting Averages

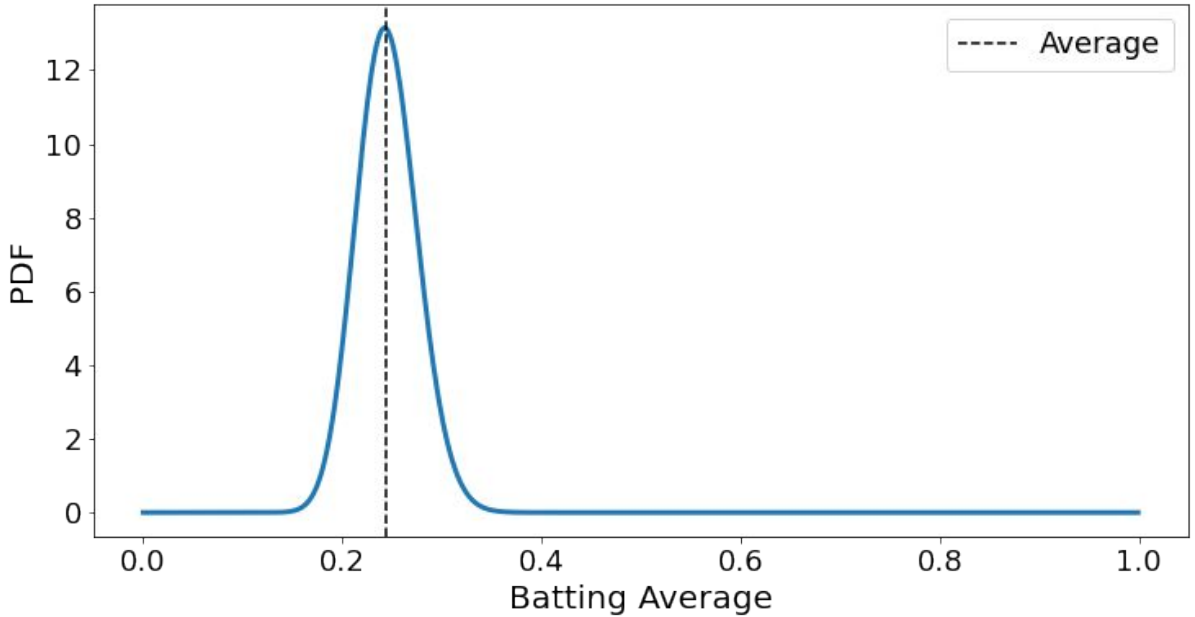
In Major League Baseball (MLB), the rate the batters hit the ball divided by the number of balls they are pitched is known as batting average. The batting average in 2021 in the MLB was 0.244 (24.4%).

A player starts the season very well and hits his first 3 balls. What would his batting average be? A frequentist would say it is 100%, however us Bayesians would come to a different conclusion.

Example: Baseball Batting Averages

Prior

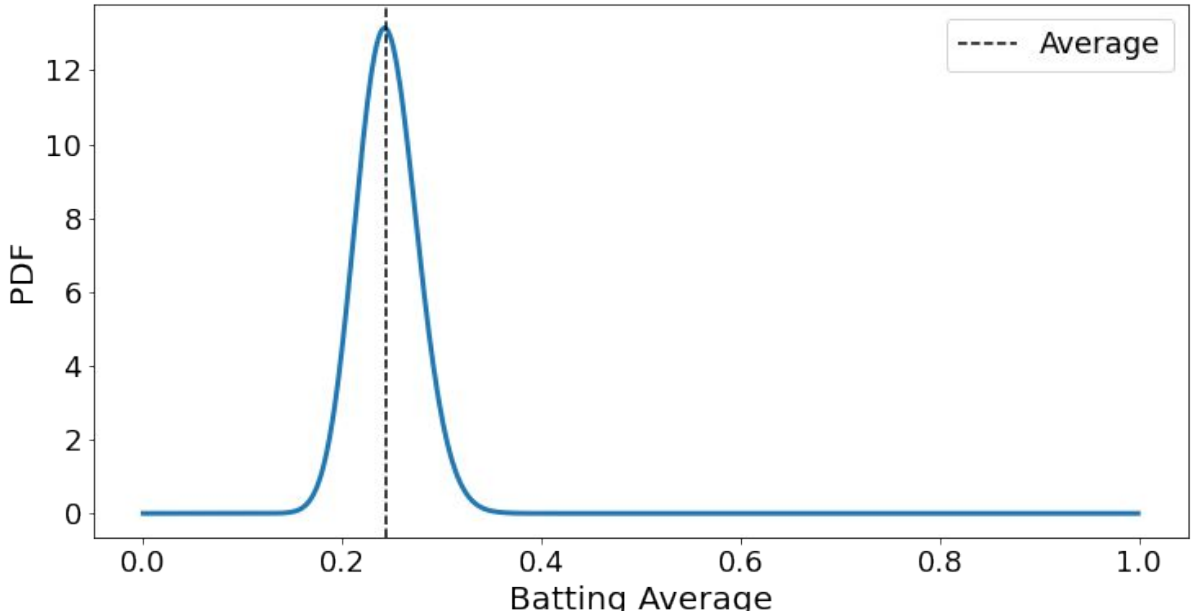
We know that the batting average is 0.244, but what about the possible range of values? A good average is considered to be around 0.3, which is the upper range and one below 0.2 is considered to be quite bad.



Example: Baseball Batting Averages

This looks reasonable as our range is pretty confined between 0.2 and 0.3. There was no particular reason why I chose the values of $\alpha=49$ and $\beta=151$, they just satisfy what we know about the prior distribution.

However, this is often the argument made against Bayesian statistics. As the prior is subjective, then so is the posterior. This means probability is no longer objective, but rather a personal belief.

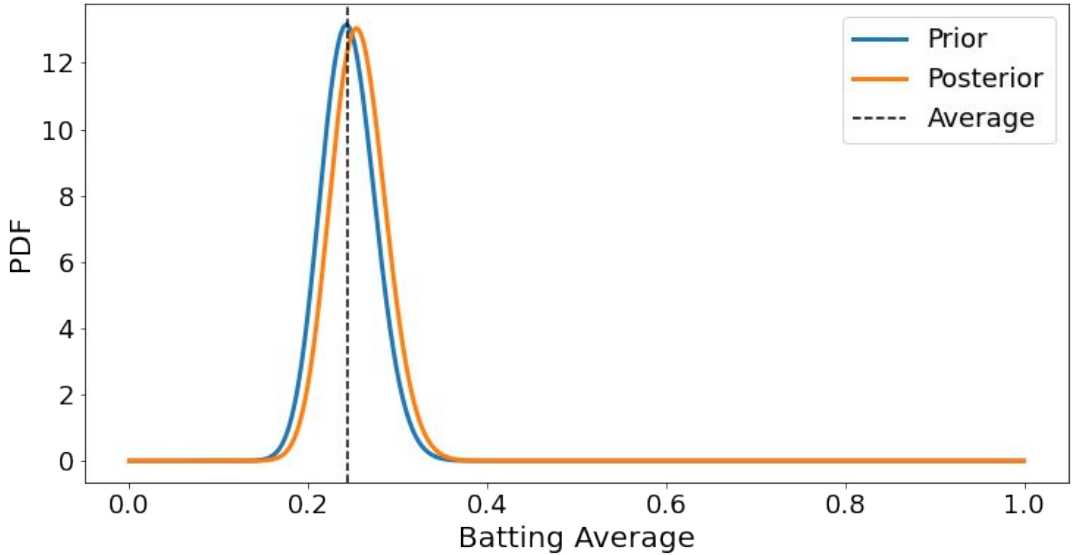


Example: Baseball Batting Averages

Likelihood and Posterior

The likelihood of the data is that the new player has hit 3 from 3, therefore they have have an extra 3 successes and 0 failures.

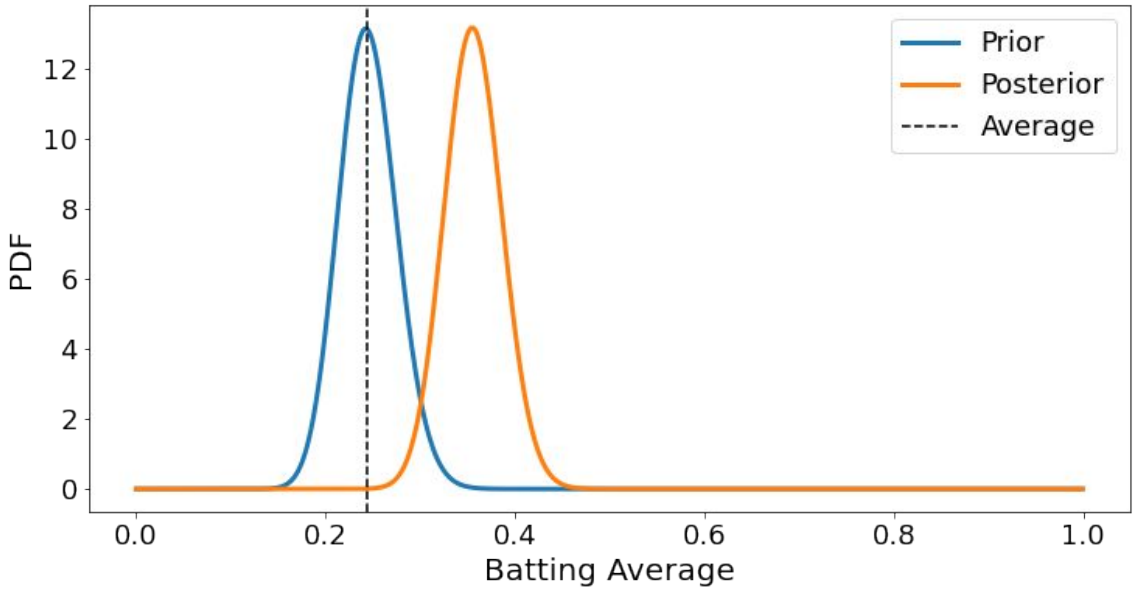
Using our knowledge of the conjugate prior, we can simply add an extra 3 to the value of *α* and 0 to *β*:



Example: Baseball Batting Averages

Likelihood and Posterior

It makes sense why the average has barely shifted as three balls is not that many. What if we now said the player hit **40** out of **50** balls, what would the posterior now look like?



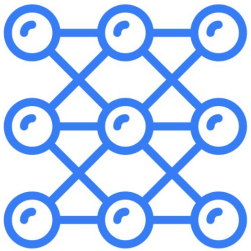
Summary

Key Points

- **Bayes' theorem is just a rearranged version of conditional probability. It's the way we apply is what gives rise to bayesian statistics.**
- **The main idea is to update our belief (probability) about an event in the light of new data.**
- **To apply Bayesian updating we need to start with an initial belief, called the prior. This is subjective and one of the criticism of Bayesian statistics.**
- **The trickiest part of Bayesian statistics is estimating the marginal probability of the event. This is where things like conjugate priors come in handy.**

@egorhowell

Newsletter



Dishing The Data