Published in Towards Data Science

Dr. Robert Kübler Follow

Dec 16, 2020 · 7 min read · ✦ · Listen



BAYESIAN STATISTICS

# A Gentle Introduction to Bayesian Inference

Learn about the difference between the frequentist and the Bayesian approach of reasoning

Photo by Sergi Viladesau on Unsplash

## A Short Story

The three friends Frequentist Frank, Stubborn Stu, and Bayesian Betty go to a funfair where a mysterious-looking tent catches their eyes. Inside, they meet Claire Voyant who claims to be a… fortune teller. The friends don't believe her, of course — *they need proof*. So they conduct a little experiment:

The friends take a normal deck of cards. Frank shuffles it and draws a card randomly, not showing it to Claire. He asks her to name the color of the card he has just drawn; red or black. *The chance of succeeding for a fortune teller should be 100%, for a normal person it is only 50%.*

**Claire answers correctly**. He puts the card back into the deck and hands it over to Stu. Stu repeats the same steps, drawing a card face down, and asking Clair to name the color. **She is right again**. And, it was bound to happen, **Clair has the right answer for the** third time when Betty did the experiment. The friends turn their backs on Claire and discuss. Which conclusion do they draw from this experiment?

#### Stu's Explanation

Stubborn Stu starts with a radical view:

> I don't care about the experiment. From my experience, it's highly unlikely that she has psychic powers. If I had to quantify my belief, I would say that there is a 0.1% chance that she has these kind of powers.

Stu has a prior belief in his head he thinks is the absolute truth. In mathematical terms, this is

$$p(\theta)$$

Stu's prior belief. Note that there is no 'data' term involved.

where θ is either 'fortune teller' with a probability of 0.1% and 'not a fortune teller' with a probability of 99.9%, i.e.

$$p(\theta = \text{fortune teller}) = 0.001 \text{ and}$$
$$p(\theta = \text{not a fortune teller}) = 0.999$$

#### Frank's Explanation

Stu has got a legitimate, yet very simple answer. But Frequentist Frank sees things differently:

> Data is everything. She got 3 out of 3 right, so I say she clearly has psychic powers.

He has come to this conclusion via a *Maximum Likelihood* approach, where he maximized the following *likelihood* formula with respect to θ:

$$p(\text{data} \mid \theta)$$

Frank tries to find the θ that maximizes this term: The probability of the observed data, given that Claire is or is not a fortune teller.

Here, the data is that **Claire scored right three times**, thus Frank can try out the two possible values for θ:

$$P(3 \text{ right} \mid \text{fortune teller}) = 1$$

The probability of getting 3 out of 3 right, given that she really is a fortune teller, is 100%.

$$P(3 \text{ right} \mid \text{not a fortune teller}) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

The probability of getting 3 right, given that she is only an ordinary human, is 12.5%.

Frank votes for psychic powers because the probability of observing this kind of outcome (getting 3 colors right) is higher for a fortune teller than for an ordinary person. **Note that he can't give a probability that she is not a fortune teller. The maximum likelihood approach does not allow for that.**

Betty thinks:

> Both explanations are extreme in their own ways. Stubborn Stu disregards the collected data, which is clearly stupid. However, Frank pays too much attention to the data, as if nothing else exists in this world. Maybe Claire was just lucky, especially with only 3 trials.

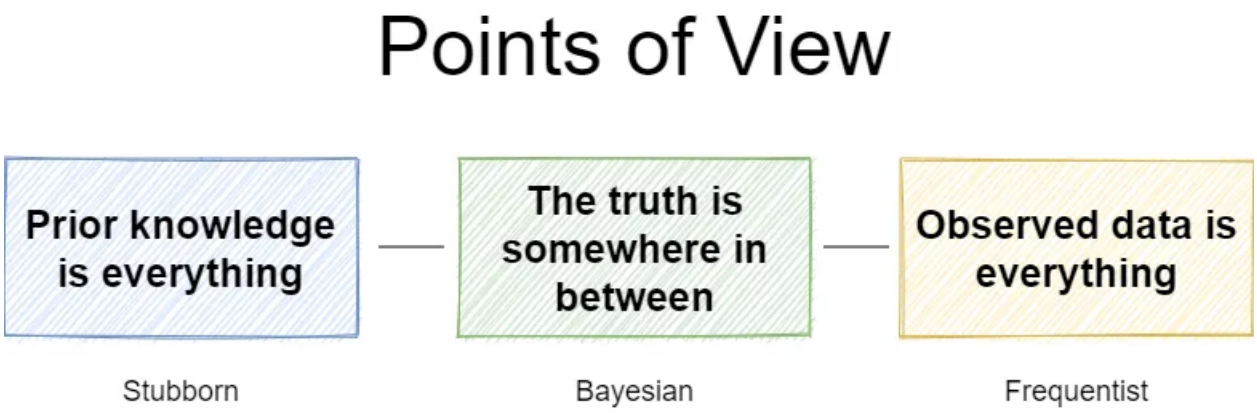
You can argue that Frank behaves unreasonably, too, **but he is merely doing what we Data Scientists often do.**

> *If we do a standard linear regression or train a neural network, for example, we behave like Frequentist Frank: we search for some real-valued parameters or weights of the model that best fit the observed data. The only difference is that Frank only has to determine one discrete parameter that can take the two values {'fortune teller', 'not a fortune*

> *teller'}. If he has to decide for one, 'fortune teller' is the better fit for the observed data (getting 3 answers right), according to the Maximum Likelihood approach. Frank's approach kind of* **overfits** *in the presence of very small data.*

#### Betty's Explanation

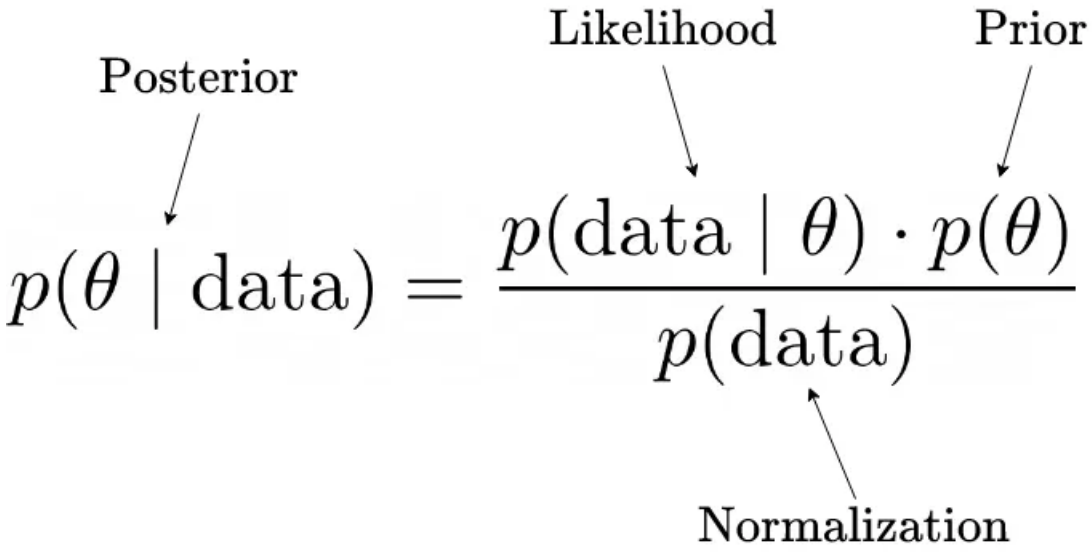
So, Bayesian Betty tries to find an explanation that somehow connects these extremes.




She reasons:

> I start off with a prior belief. Then, I look at the observed data and let each data point change my mind a little bit. The more data I observed, the further I can drift off my initial belief. This procedure results in my posterior belief.

The key to expressing this kind of reasoning is — you guessed it — the Bayes Formula, which contains some old friends:




The Bayes theorem. θ are the parameters of the model, data the observed data.

Here, you can see that the Bayes approach combines the frequentist's likelihood and the simple prior by multiplication. Because the resulting quantity *p(*data | *θ) p(θ)* is not a probability (or density) anymore in general, we have to scale it. That is the reason for having the *p(*data*)* around in the denominator.

But now we have another problem: What is *p(*data*)* anyway? How to compute it? The law of total probability says the following:

$$p(\text{data}) = \sum_{\theta} p(\text{data} \mid \theta) \cdot p(\theta)$$

and this formula only involves terms that we know already!

Bayesian Betty, with her great mental capacities, uses both formulas and plugs all the values in.

$$\begin{aligned} p(\text{fortune teller} \mid \text{3 right}) &= \frac{p(\text{3 right} \mid \text{fortune teller}) \cdot p(\text{fortune teller})}{p(\text{3 right} \mid \text{fortune teller}) \cdot p(\text{fortune teller}) + p(\text{3 right} \mid \text{not a fortune teller}) \cdot p(\text{not a fortune teller})} \\ &= \frac{1 \cdot 0.001}{1 \cdot 0.001 + \frac{1}{8} \cdot 0.999} \\ &\approx 0.008 \end{aligned}$$

In the same way, or even simpler, she computes

$$p(\text{not a fortune teller} \mid 3 \text{ right}) = 1 - p(\text{fortune teller} \mid 3 \text{ right}) \approx 0.992$$

After a longer break, she says:

> Considering Stu's prior and Frank's likelihoods, I think Claire merely a normal person. If I have to quantify my data-backed belief, I would say that there is a 0.8% chance that she has psychic powers.

Note how she has come to the same conclusion as Stu. The probability of Claire being a fortune teller is still low. However, Betty uses existing data in addition to a simple prior belief to boost the confidence of Claire being a fortune teller by **eight times** in comparison to Stu's 0.1%. And it makes sense: Even if they have repeated the experiment only 3 times, Claire undeniably has still gotten everything right. Intuitively, this **has to** increase the chance of her having psychic powers.

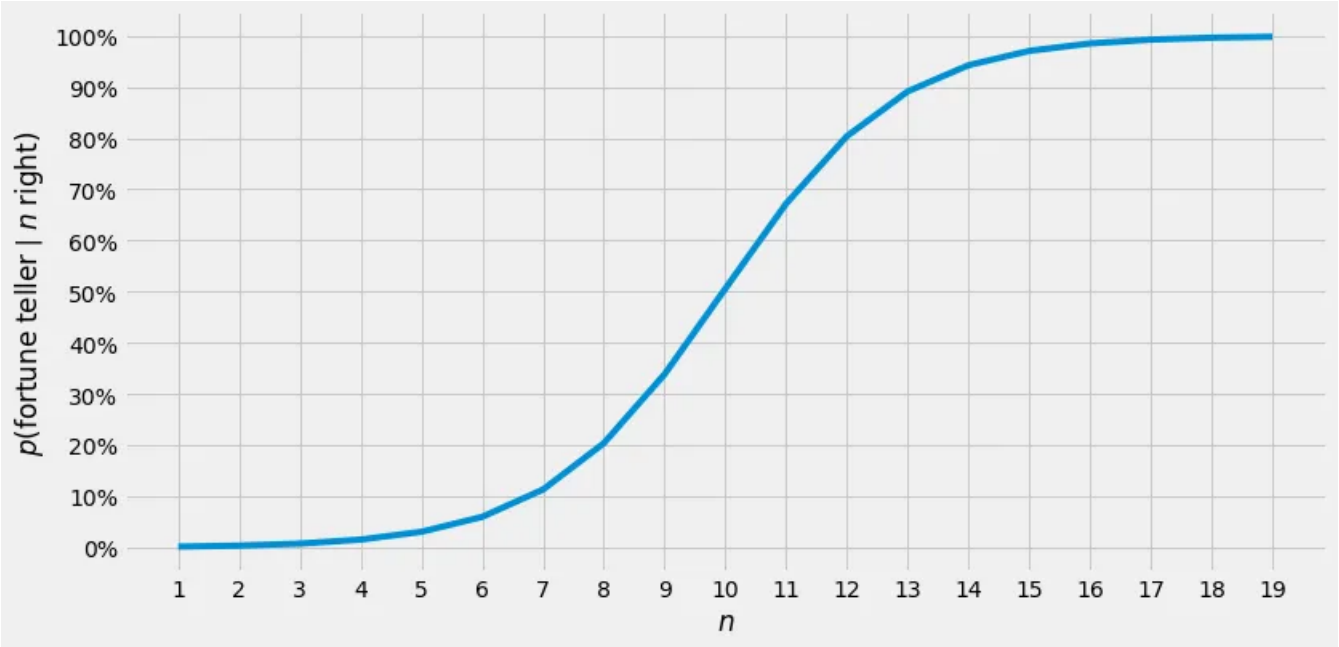
#### The Verdict

The friends discuss it again. Is Claire a fortune teller? They all agree with Betty's arguments and conclude:

> We don't know, but probably not.

The friends go home and call it a day.


Not satisfied? Well, with the power of ***M.A.T.H.S*** we can simulate different ways the story could have ended. What if they tried it with 10 cards instead of 3, and she got all of the colors right? Or even 20? Here's a picture of the answers:




We can see that for this special prior belief, 10 correct answers would have been enough to be quite uncertain about Claire's mystic powers already. For 13 correct answers, the probability of Claire being a fortune teller would be at a whopping **90%** already.

## Conclusion

In this article, we have seen the Bayesian approach in action with the help of a small example. It uses prior knowledge and updates it with observed data to create a posterior, exactly like humans intuitively do.


The Bayesian approach.

### Advantages

This approach is better than discarding the data and just proceeding with some prior, obviously. It is even more powerful than the maximum likelihood method: you can see this by choosing a *flat prior,* i.e. the prior gives the same probability (or density) to every possible value $\theta$ and is essentially a constant. By doing this the $\theta$ that maximizes the posterior distribution also maximizes the likelihood.

Furthermore, the Bayes method even gives you a distribution of the parameters, while the maximum likelihood method does not. This is a great advantage since you can always easily quantify your beliefs.

### Disadvantages

This approach is computationally more involved — for humans as well as computers. Computing the posterior is a difficult problem in general, and requires advanced algorithms such as Markov chain Monte Carlo. In our case, it was still easy enough, though. Betty could even do it in her head.

## Further Readings

Check out my hands-on articles about solving a slightly more difficult problem using Bayes.

**Beginner-friendly Bayesian Inference**

Let's do Bayesian inference hands-on with a classical coin example!

towardsdatascience.com

**Conducting Bayesian Inference in Python using PyMC3**

Revisiting the coin example and using PyMC3 to solve it computationally.

towardsdatascience.com

I hope that you learned something new, interesting, and useful today. Thanks for reading!

**As the last point, if you**

1. **want to support me in writing more about machine learning and**
2. **plan to get a Medium subscription anyway,**

**why not do it via this link? This would help me a lot!** 😊

*To be transparent, the price for you does not change, but about half of the subscription fees go directly to me.*

**Thanks a lot, if you consider supporting me!**

If you have any questions, write me on LinkedIn!

Bayesian Statistics Bayesian Machine Learning Maximum Likelihood Editors Pick Artificial Intelligence