Bayesian Inference 101 Notes

(iCapital)

Arnav Sheth
iCapital Network

November 6, 2023

Contents

1 Introduction

I will be talking about Bayesian Inference today.

Having talked with Weidong and Clinton about this, my main motivation here is to try and explain why I feel Bayesian statistics is the right way to approach the problems we are trying to solve.

It's going to be very high-level. If you guys are interested in learning more about the topic I'm happy to talk more about it.

And of course, hopefully, by the end of this talk you agree with me that Bayesian statistics is appropriate for this task!

Next Slide

Read the agenda.

2 Simple Example

It is generally assumed that a coin toss comes up heads half of the time. It has even become a standard metaphor for two events with equal probability.

Next Slide

But think about it—is it really 50-50? Suppose we always flip a coin starting with the heads side up. Could the outcome actually be biased toward either heads or tails? Think about it.

Next Slide

Let's start with a fair coin. That is, we assume the coin is physically symmetric. Mathematically, it's a relatively simple physical system: approximately a parabola.

Take any intro-level calculus class and you will learn that the outcome should be essentially determined by the initial conditions at the beginning of it's trajectory. It's pretty predictable. So, the randomness comes from the flipper.

Next Slide

Let's run an experiment: flip a coin $n = 10$ times, starting with heads up each time. (flip, flip, flip, ...).

Do we know anything more now than when we started? We got some data, so we should know more now. But probably we need more data! How much more? And how can we quantify our uncertainty about the answer?

Next Slide

As it turns out Diaconis et al. (2007) show mathematically in fact, the outcome is slightly biased and will come up the same way it started about 51% of the time!

This was confirmed by Bartos, et al in 2023 having tossed a coin 350,757 times.

Let's formalize the question using some mathematical notation. This will also help us answer the question.

3 Probability and Statistics Are Two Sides of the Same Coin

Let $X_1, \ldots, X_n$ be the outcomes of n coin flips, and suppose they are i.i.d. (independent and identically distributed), with the probability of heads equal to θ.

This defines a probabilistic model. In other words, if we knew θ, we could say all kinds of things about the distribution of $X_{1:n} = (X_1, \ldots, X_n)$. This is Probability.

Statistics, meanwhile, goes the other direction. Trying to obtain information about θ from $X_{1:n}$.

To see if the outcome is biased, based on the data $X_{1:n}$, perhaps the first thing that comes to mind is to maybe toss the coin a lot, look at the proportion of heads, and see if it's close to half. But on reflection, there are some issues with this:

- How close is "close"?
- How would we quantify our uncertainty about the correct answer?
- If n is very small, say 2 or 3, there is a good chance that the flips will all come up the same (all heads or all tails), in which case the proportion of heads would be 1 or 0.
  - But from experience, we know θ is unlikely to be close to 1 or 0.
  - Would it be better to take such prior knowledge into account?

4 History

Thomas Bayes answered these questions, but died before publishing them.

He was an ordained minister, accomplished mathematician and Fellow of the Royal Society.

His work was discovered by his friend Richard Price and published a few years later.

Pierre-Simon Laplace rediscovered and extended the idea, publishing it in 1774.

5 Frequentist Statistics

Traditional or frequentist statistics, (i.e., methods such as regression analysis, t-tests, ANOVA, etc.) stem from the understanding that the world is described by

parameters that are ***fixed*** and ***unknown***.

That there is some true value out there and we are trying to estimate it.

QUESTION *What are parameters?*
They can be all kinds of things — the rotation rate of the earth, the average life span of a naked mole rat, the average number of kittens in a litter of cats, or to use a finance example, the relative volatility on a stock (CAPM).

We can't directly measure the parameter, so we estimate parameters by taking random samples from the population, computing some statistic over the sample, and using that as our estimate of the population parameter.

It is rare that we can have access to the *entire* population of interest (e.g. all men aged 40-49, all private equity fund returns, the entire universe of stock returns).

Since these parameters are ***unknown***, we do not know their exact values. Since they are ***fixed***, however, we generally do not discuss them in probabilistic terms. By that I mean, we always assume that there is some true value out there. And that value is fixed.

EXAMPLE
The beta of MSFT in 2023 (so far) was 1.19. This value is fixed.

Probabilities, confidence intervals, p-values, etc. are only meaningful in the context of the outcome of multiple repeated random experiments drawn from the population.

6 Bayesian Statistics

The Bayesian says, "Who cares?". We can't really run the experiment several times for many reasons, e.g., it's expensive, there's limited data, and many other reasons.

Bayesian statistics applies probabilistic methods and reasoning directly to the parameters.

We can use probabilities not only to express the chance that something will occur, but we can also use them to express the extent to which we believe something. Mathematically, it somehow works.

So we can use the algebra of probabilities to quantify and describe how much we believe various propositions, such as “the beta for MSFT is 1.19.”

One of the fundamental differences, therefore, is that the frequentist can only apply probabilities to the act of repeating an experiment. The Bayesian can apply probabilities directly to their knowledge of the world.

6.1 Math

The idea is to assume a ***prior*** probability distribution for θ—that is, a distribution representing the ***plausibility*** of each possible value of θ before the data is observed.

Then, to make inferences about θ, one simply considers the conditional distribution of θ given the observed data.

This is referred to as the ***posterior*** distribution, since it represents the ***plausibility*** of each possible value of θ after seeing the data.

Note that I have not used the word ‘probability’ here, deliberately. We are going through every possible instance of whatever event we are looking at, and then computing the most plausible set of events, along with their relative plausibility.

Mathematically,

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)} \propto p(x \mid \theta) p(\theta)$$

Here, x is the observed data (for example, the coin flips, $x = x_{1:n}$).

More generally, the Bayesian approach—in a nutshell—is to assume a prior distribution on any unknowns, and then just follow the rules of probability to answer any questions of interest.

This provides a coherent framework for making inferences about unknown parameters θ as well as any future data or missing data, and for making rational decisions based on such inferences.

7 Some Advantages to Bayesian Statistics

The results are understandable. We can argue for hours about what *precisely* the p-value or a confidence interval means for a coefficient in a regression.

For example, it is quite difficult to understand exactly what a confidence interval is and how it arises; a 95% confidence interval does not mean that there is a 95% chance that the true parameter lies in the interval.

What it actually means is much more complex and is in terms of multiple repetitions of the experiment.

Bayesian probability distributions and posterior intervals, on the other hand, do exactly what they say — they directly express our belief about the parameter's value.

They are easy to understand and easy to explain. You can display the entire distribution and ascertain your own level of confidence in the results, based on those beliefs.

All parts of the model, including priors and other assumptions, are explicit and open to criticism. The only aspect of Bayesian inference taken for granted is Bayes' theorem. Everything else, in particular the prior assumptions, is made explicit, documented, and can be critiqued.

Frequentist statistics makes assumptions about how the world works, but these assumptions are often buried deep within the methods used and do not necessarily hold in the data to which they are applied.

For example, CAPM is a one-period model, but what is the period? If you calculate betas over a 5-year window, is that appropriate? Is the data stationary over that time? What about daily returns? Are they normally distributed?

There's all this uncertainty about the results but there's almost no way to quantify it. In fact, we can get false confidence in the results, which is worse!

In Bayesian inference, everything is explicitly articulated either in the priors or in the structure of the model itself. Some assumptions, such as independence, are still quite subtle, but they can still be explicitly dealt with.

The model is scrutable. Any tool makes assumptions about the data. Some of these assumptions, such as normal distribution of errors, can be checked; others, such as independence, are incredibly difficult to test.

In both cases, though, it is difficult as a working scientist to find good answers on how much violating a particular assumption affects the validity of the result.

Traditional statistical methods are largely black boxes; they are handed to us by the statisticians, but it is difficult for those of us without graduate degrees in statistics to peek inside, see how they work, and understand how they break.

When we build the model from scratch, we have to understand how it works, and we have the opportunity to trace various changes, assumptions, etc. through it and see how they affect the result.

Admissibility. This is the equivalent of the Pareto efficiency but in the context of decision theory.

An admissible decision rule is one that cannot be improved upon by any other decision rule. In other words, it's the "best" decision rule according to your loss function.

> A decision rule is inadmissible if there exists another decision rule that performs better (has lower expected loss) for at least one possible set of true conditions.
>
> A Bayesian estimator is admissible under certain conditions. You may refer to this text for more on this.

8 Examples

8.1 Coin Toss

So remember when I told you to toss a coin ten times? I did it.

The nice thing about a Bayesian model is that it updates with data. It actually learns as more data is fed into it. So I thought it might be interesting to see *how* the model updates.

We start with one set of plausibilities. These are the prior plausibilities. In this case, I've actually assumed nothing, i.e., a uniform prior. I've basically assumed equal plausibility for all probabilities of heads from zero to one. I'll show you what I mean by that a few slides down.

Meanwhile, once I run the model, the model "learns," and updates itself, , with each datapoint, i.e., with each toss.

For the first toss, we get heads. So with only one datapoint, it assumes a probability of one for heads.

For our second toss, we get tails. So now, it assumes that $p(H) = 0.5$.

For our third toss, we get tails again. And it updates accordingly.

And so on.

The important thing to see here is that these lines are just ***collections of plausibilities***.

The plausibility of $p(H)$ being some value is what each point on the line represents.

By the time we get to our 10th toss, for me I had a total of 5 heads and 5 tails. So the model updated appropriately.

QUESTION *N*ow what if I had not used a flat prior? What if I had used some different priors? Let's see what happens there.

Next slide

The first slide here is the flat prior itself.

The second slide is a step prior which does not allow for any values below 0.5. So the plausibility of getting anything below 0.5 is zero.

And finally, we have a peaked prior.

For the last slide, I have actually taken the distribution generated by my 10 tosses, sampled 10,000 values from it, and then drawn a histogram of those values.

So you can actually ***generate data*** from a set of posterior plausibilities.

8.2 Vampirism

There are two ways to approach this, one more intuitive than the other.

1. We are told the following:
    - Pr(positive test result | vampire) = 0.95
    - Pr(positive test result | mortal) = 0.01

- Pr(vampire) = 0.001

We can now use Bayes' theorem to invert the probability to get Pr(vampire | positive):

$$\begin{aligned}\mathbb{P}(\text{vampire} \mid \text{positive}) &= \frac{\mathbb{P}(\text{vampire} \mid \text{positive})\mathbb{P}(\text{vampire})}{\mathbb{P}(\text{positive})}\\ \mathbb{P}(\text{positive}) &= \mathbb{P}(\text{positive} \mid \text{vampire})\mathbb{P}(vampire) + \\ &\quad \mathbb{P}(\text{positive} \mid \text{mortal})\left(1 - \mathbb{P}(\text{vampire})\right)\\ &= (0.95)(0.001) + (0.01)(0.999) = 0.0868\end{aligned}$$

Or an 8.7% chance that the suspect is actually a vampire.

2. Not convinced? Let's try it a different way:

   (a) In a population of 100,000 people, 100 of them are vampires.

   (b) Of the 100 who are vampires, 95 of them will test positive for vampirism.

   (c) Of the 99,900 mortals, 999 will test positive for vampirism.

   So if we test all 100,000 people, what proportion of those who test positive for vampirism are actually vampires?

   There are 95 + 999 = 1094 people who test positive. Of the 1094, 95 are really vampires. So: $\mathbb{P}(\text{vampire} \mid \text{positive} = \frac{95}{1094} \approx 0.087$!