An Introduction to Bayesian Statistics

Manuele Leonelli

2021-01-22

Contents

Preface

These lecture notes are used to support the university course *Bayesian Statistics* given to 3rd year students of the Bachelor in Data & Business Administration at IE University, IE University, Madrid.

Chapter 1

Introduction

Bayesian statistics is the name given to a whole branch of statistics which differs from the traditional approach taught in introductory statistics classes.

In order to understand what this difference is, let's first review the basic traditional approach, which is often referred to as *frequentist* (we will see later on where this word comes from). In general, we are interested about an unknown characteristic of a population, usually called a population *parameter*. For instance, we may be interested in the mean annual salary of an individual living in the city of Madrid. Let's call this quantity μ. Suppose it is impossible to exactly compute this quantity by accessing the information about the salary of everyone living in Madrid.

The next step would be to collect a sample, let's call it y, with information about the salary of some inhabitants of Madrid. We would then use this sample to come up with a best guess, or an estimate, of μ. Due to multiple reasons which are not of interest here, we know that computing the sample mean $\bar{y}$ is an optimal way to estimate μ and we denote such a value $\widehat{\mu}$.

In most cases, having just a single value is not satisfactory enough, and we instead want an interval of values which would likely contain the true value μ. So a follow-up step would be to construct a so-called *confidence interval*. You may recall that a confidence interval for a mean can be computed as

$$\widehat{\mu} \pm z_{\alpha} \sqrt{\frac{\widehat{\sigma^2}}{n}},$$

where

- n is the sample size;
- $\widehat{\sigma}^2$ is an estimate of the population variance, that is a guess of how much variability there is around the mean;

- z_α is some value that comes either from a Normal or T-Student distribution which depends on how confident we want to be that the true μ lies within the interval.

In the so-called frequentist approach we only used the data to guide the estimation of the unknown parameter and we assumed that we had no *prior* information about what plausible values of μ might be. Suppose on the other hand that in advance we had some guessing that the true value of μ may lie within 20k and 50k. In the frequentist approach there is no way to embed this information within the inferential process. Only the data can be used to guide the estimation of the parameters. The *Bayesian* approach on the other hand is designed to formally account prior information about the unknown parameter in the inferential process.

1.1 A more technical discussion

In order to understand how the Bayesian approach actually works, let's discuss slightly more formally how the frequentist estimation process works. Let's more generally call θ the unknown population parameter that we want to estimate and y the sample. Depending on the type of data we are working with, we start choosing a statistical model, or a data-generating process, $p(y|\theta)$. To make this clearer, consider the following examples:

- suppose that we observe coin tosses and we are interested in the probability of head: then the standard choice for $p(y|\theta)$ would be the probability mass function of a Bernoulli distribution.
- suppose we collect information about the number of goals scored in football matches: then the standard choice for $p(y|\theta)$ would be the probability mass function of a Poisson distribution;
- often we let $p(y|\theta)$ be the probability density function of a Normal distribution. Such a choice is in general due to the *Central Limit Theorem* which tells us that if our sample size is large enough then any distribution is well approximated by a Normal.

The quantity $p(y|\theta)$ is also often referred to as the *likelihood* and written as $L(\theta|y)$. Notice that the order of θ and y is reversed: whilst $p(y|\theta)$ is seen as the distribution of y given the parameter θ, $L(\theta|y)$ is seen as a measure of how likely is that the parameter is θ given that we observed the sample y.

No matter what the interpretation is, we view as random only the process that generated the data which is assumed to behave according to some distribution $p(y|\theta)$. The parameter θ is itself assumed to be fixed and unknown: it is not random, it is just that we do not know its value. We use $p(y|\theta)$ to come up with

an estimate $\hat{\theta}$ of the unknwon θ. For instance in maximum likelihood estimation $\hat{\theta}$ is found as

$$\hat{\theta} = \arg\max_{\theta} p(y|\theta)$$

The Bayesian approach differs from the frequentist approach since it consider the unknown parameter θ to also be a random variable and not simply fixed. Therefore in Bayesian statistics the unknown parameter is also assigned a distribution, $p(\theta)$, which is referred to as *prior* distribution. Furthermore, the main building block of Bayesian statistics is a joint distribution for the data-generating process and the unknown parameter, since both are random. Such a distribution is

$$p(y, \theta),$$

which by using basic rules of conditional probabilities can be written as

$$p(y, \theta) = p(y|\theta)p(\theta).$$

The above expression can be seen as the product of the data-generating distribution, or likelihood, with the prior distribution.

The prior distribution encodes our beliefs about the unknown parameter of interest before observing any data. Now suppose we collect the sample y: we would like to use it to revise or update our beliefs about the unknown parameter θ. This is done using Bayes theorem, hence the name Bayesian statistics. Most likely, you are already familiar with the version of Bayes theorem for events namely:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Similarly we can write

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.1}$$

and $p(\theta|y)$ is usually called the *posterior distribution*: it encodes our beliefs about θ after having observed the sample y. Equation (1.1) is the backbone of Bayesian statistics and the main task in a Bayesian analysis is to compute such a posterior distribution.

There are two important observations to make now. First, in Equation (1.1) the terms $p(y|\theta)$ and $p(\theta)$ are chosen by the modeler, whilst $p(y)$ can be computed from these two as:

$$p(y) = \int p(y, \theta)d\theta = \int p(y|\theta)p(\theta)d\theta$$

and it is called *marginal likelihood*. So Equation (1.1) can also be written as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}.$$

Second, our object of study is the parameter θ and $p(\theta|y)$ is a function of θ only. So the term $p(y)$ at the denominator of Equation (1.1) is actually only a normalization constant to make sure that $p(\theta|y)$ integrates to 1. So Equation (1.1) is often presented as

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{1.2}$$

Equation (1.2) is actually the one that most often we will work with since $p(y)$ is not actually necessary and it is often very hard to compute.

One question you might have is: why did we combine the prior distribution and data-generating distribution using Bayes Theorem? We could have used different rules to come up with a so-called posterior distribution. We will not enter into the details of this, but if beliefs are encoded as probability distributions, then it can proved that Bayes theorem is the optimal way to update them in the light of data.

1.2 A first example

Suppose we are interested in estimating the prevalence of COVID-19 cases in the city of Madrid. For this purpose we select a sample of 100 individuals living in the city. Let's assume that each individual has the same probability of having the disease and that each individual has the disease independently of others.

Then we could model the fact that each individual has the disease as a Bernoulli distribution, where the value 1 denotes having the disease. Suppose the result of COVID testing over these 100 individuals shows that 10 of them are positive. Then we would estimate the parameter θ of the Bernoulli distribution as $\widehat{\theta} = 10/100 = 0.1$, which would in turn be our estimate for the prevalence of COVID cases in Madrid. Furthermore, we could construct a 95% confidence interval which, if you recall from previous courses, can be computed as

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.1 \pm 1.96\sqrt{\frac{0.1 \cdot 0.9}{100}} = (0.041, 0.159).$$

Let's take a Bayesian approach instead. For the data-generating process it is still of course reasonable to believe that a Bernoulli distribution with unknown parameter θ is appropriate. Now we also need to define a prior distribution. Suppose that from data related to other European cities, we believe that such a prevalence number is between 0.05 and 0.25 with an average around 0.15. We will later learn ways to choose prior distributions, but suppose that such a prior information can be represented by the distribution colored in blue in Figure 1.1. This distribution represents our beliefs about the prevalence of COVID in the city of Madrid.

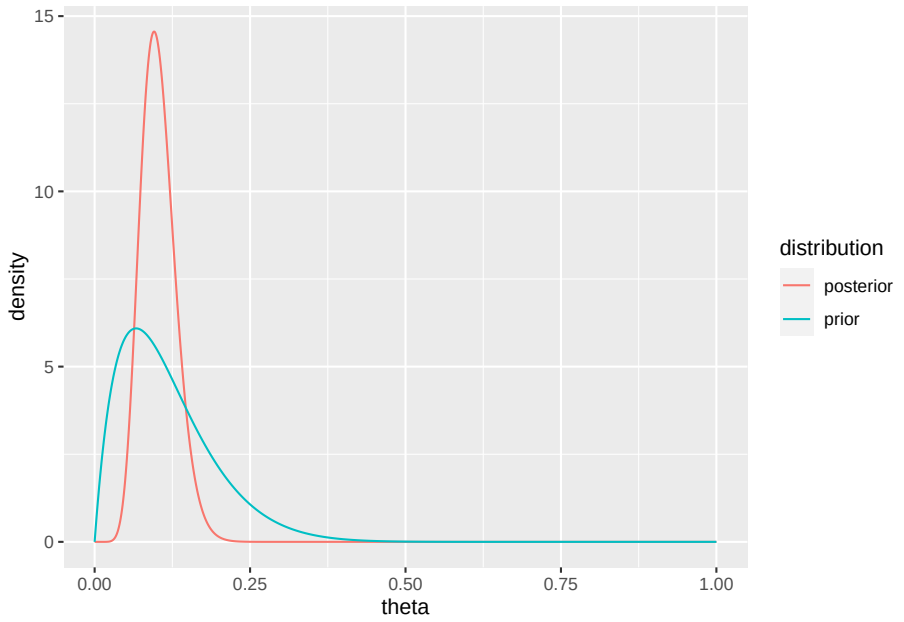

Figure 1.1: Prior and posterior distribution for the COVID example.

Suppose we collect the same sample as before of 10 positive out of 100 and compute the posterior distribution using Bayes theorem. This is reported in Figure 1.1 by the red distribution. So differently from the frequentist case where we have a single point estimate of θ or a confidence interval of plausible values, we now have a full distribution for the variable θ in the light of data and prior beliefs. Given such a distribution we can do multiple things:

- we can come up with a single point estimate for θ, for instance the mean or the mode of the distribution;
- we can identify a plausible region of values of θ in the same spirit of a confidence interval.

It is important to notice that our beliefs about the prevalence of COVID has now changed in the light of data. Our prior distribution was quite spread around values between 0.05 and 0.25, whilst the posterior is a lot more concentrated around 0.1 which is actually the sample proportion of COVID.

1.3 Why Bayesian statistics

The previous example showed an example of a simple Bayesian analysis and how it differs from a frequentist one. One might wonder what is the real advantage of taking a Bayesian approach with respect to a frequentist one: we saw that in the end the conclusion from both approaches was pretty much the same.

In general we can notice the below advantages of a Bayesian approach:

- it allows to more flexibly construct complicated models. This is a concept we will see in later chapters when we will discuss hierarchical models;
- it allows to easily embed in inference other type of information which is not only in the form of data: for instance, expert judgment or results from other studies;
- it allows to use data in a sequential manner. Suppose we carried out our analysis about COVID prevalence and once finished we are actually given the result of tests on new individuals. In the Bayesian framework, we could then use our posterior from the previous step as our new prior and use the same machinery to come up with a new posterior. In a frequentist setting, we would need to use the full dataset again to come up with an estimate of θ. Of course this is trivial in the COVID example, but for much more complicated models, it may be very costly to repeat the analysis with the full dataset.
- it leads to an intuitive interpretation of confidence intervals. Standard confidence intervals are probability statements about $\widehat{\theta}$ and not θ itself as often erraneosly thought. The correct interpretation of a 95% confidence interval is that if we were to collect many many times samples under the same conditions and each time construct a confidence interval, then 95% of the times the interval would include the true value θ. However, most often people interpret confidence intervals as with 95% probability the true unknown parameter lies in the interval, which is not correct. However, this is the interpretation of confidence intervals in the Bayesian setup;
- it can deal more easily with rare situations. Let's consider the COVID example again and suppose that on the other hand we observed zero positive tests. Then our point estimate of the prevalance would be zero and confidence intervals at any level of confidence (even 99.9999%) would simply be the point zero, which we would in general not believe unless the sample size is extremely large! Using the same prior as before, in the case of zero positive cases our posterior would be the one in Figure 1.2. Although most of the distribution is around zero, we still have some probability that the prevalence is some small number close to zero. The more and more only negative tests we would collect, the more the distribution would be concentrated in zero!

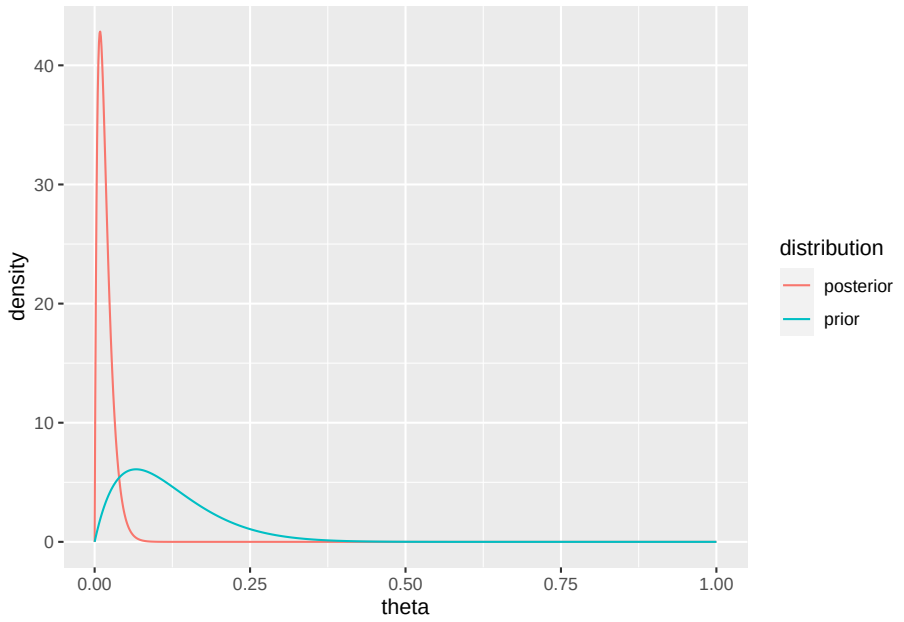

Figure 1.2: Prior and posterior distribution for the COVID example.

Of course there are also drawbacks of the Bayesian approach. The main critic regards the prior distribution and the addition of a "subjective" element in the analysis. We will discuss a lot more this issue in the next chapter. The second drawback is that it is computationally much more expensive and most often Bayesian methods require more computational power and computational time. As we will see in the next section, this was actually one of the reasons Bayesian methods were not used for many years.

1.4 A bit of history

The term Bayesian statistics comes from the fact that inference is based upon a sequential use of Bayes theorem. Reverend Thomas Bayes was an English minister whose 1763 posthumosly paper "An Essay Towards Solving a Problem in the Doctrine of Chances" gives the first account of what we now call Bayes theorem. As a matter of fact, his account of the result is not in the form you are familiar with. It was Pierre Simon Laplace, an 18th century French scientist, who introduced Bayes theorem in a form much more similar to the one we use today in his essay "Memoire sur la Probabilite des Causes par les Evenements".

Laplace was actually studying the probability of a "success" in a Binomial experiment given that data was observed. In the terminology we introduced he was

characterizing the posterior distribution of θ. The method of deriving a probability distribution for an unknown parameter given data was overall called the "inverse probability" problem and became the gold standard throughout the 19th century.

Of course there were many scientists that critiqued such a method, including Boole and Venn, due to the non-objectivity of the method. However, since no alternative was available at the time, the inverse probability method continued to be the gold-standard.

It was only at the beginning of the 20th century with the work of Ronald Alymer Fisher, Jerzy Neyman and Egon Pearson, that the frequentist approach became to emerge. Therefore, the approach of statistics that is most frequently taught is actually less recent than Bayesian ideas.

Although frequentist approached dominated statistics, Bayesian ideas were still being developed in the first half of the 20th century through the work on subjective probabilities of Harold Jeffreys, Bruno de Finetti and Leonard Savage.

In the second half of the century there was a resurgence of Bayesian methods. Two of the main reasons were:

- the work on expected utility of von Neumann and Morgenstern where a subjective view of probability can be more easily accepted became very popular;
- computational power increased at a speed never seen before and a lot of complex problems could be at last approached with a Bayesian approach.

Nowadays Bayesian statistics is as popular as frequentist statistics and research is carried out almost equally in the two frameworks. There are indeed problems that can be more easily tackled with a Bayesian approach (and the other way around too, of course!).

1.5 What's next

This introduction should have given you a feeling of what a Bayesian analysis entails and the various steps required. In the next chapters we will dig deeper into the various components of a Bayesian analysis, namely:

- we will first discuss various interpretations of what probability is which will in a way motivate or justify the use of Bayesian methods in a variety of settings;
- we will learn how to compute posterior distributions for a variety of simple models;

- we will discuss how to summarize a posterior distribution to come up with point estimates and confidence intervals;
- we will investigate various types of prior distributions and their effect to the posterior distribution;
- we will learn how to construct more complex models within a Bayesian framework, which are usually called hierarchical or multilevel models.