An Introduction to Bayesian Statistics

Manuele Leonelli

2021-02-11

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.

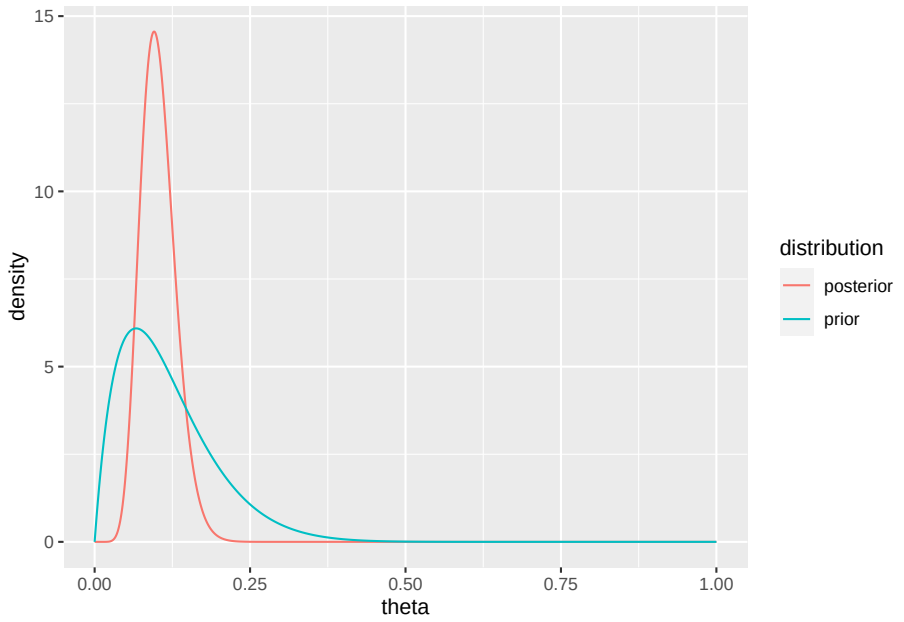

Figure 1.1: Prior and posterior distribution for the COVID example.

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!

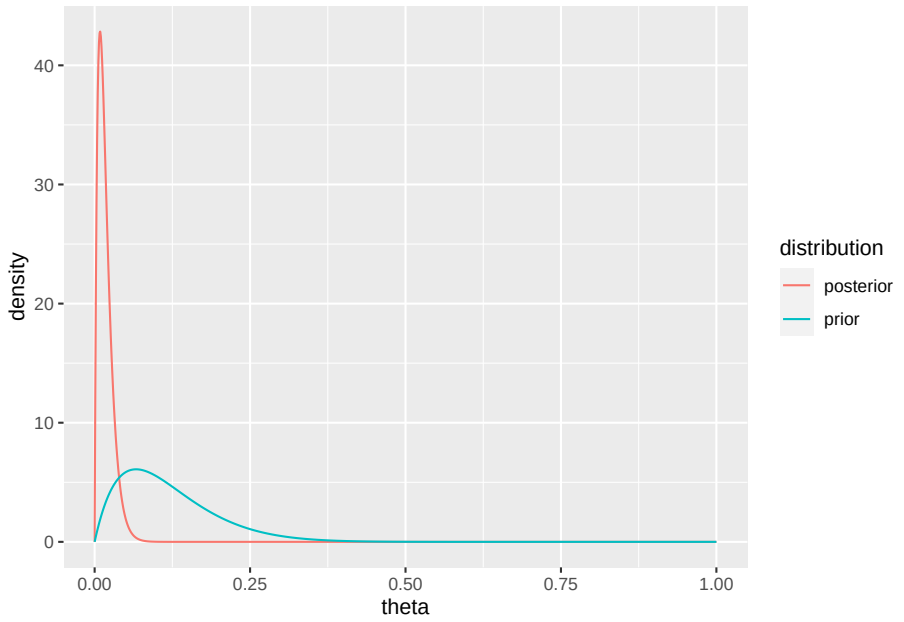

Figure 1.2: Prior and posterior distribution for the COVID example.

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 Interpretations of probability

The field of Bayesian statistics is deeply connected to a different interpretation of probability then the ones you are probably familiar with.

The most common and basic interpretation of probability is due to Pierre Simon Laplace and is based on the notion of symmetry of elementary outcomes or, to put it differently, on the notion of equiprobability. In Laplace's definition the probability of an event is defined as the number of favourable cases over the number of total cases. Let's consider the throw of a simple dice and the event that the number shown is even. There are three favourable cases (2,4 and 6) and six total cases (the numbers from 1 to 6). The probability of this event is therefore one half as one would expect. Of course this definition applies only to

cases where elementary outcomes are equiprobable: in the example if the dice is not biased.

The second most common interpretation is the so-called *frequentist* one. Probability is defined as the relative frequency in a long sequence of identical independent trials. It is assumed that there is an underlying generating process which every times generates an independent instance. Let's consider again the event of an even number in a dice throw. The probability of this event is defined as the frequency of even numbers if we were to repeat the throw of a dice an infinite number of times. Figure 1.3 illustrates how the relative frequency of even numbers evolves in 500 throws. The red line is the relative frequency: for the first 100 tries it is quite far away from the value 0.5 and then it slowly stabilizes at the true value.

```
set.seed(2021)
x <- sample(1:6,500,TRUE)%% 2 == 0
plot(cumsum(x)/1:500,type="l", ylim = c(0,1), col = "red", ylab="Probability")
abline(h=0.5)
```

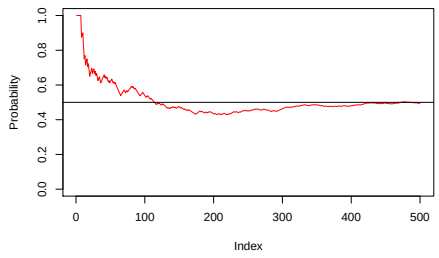

Figure 1.3: Relative frequency of even numbers in 500 dice throws

Now let's think at some other types of events:

- Spain winning the next world cup;
- the next expedition to Mars fails;
- it rains tomorrow.

It becomes challenging to define probability as relative frequency since we cannot think of a sequence of repetitions of such processes that are exactly equal. Such events are in general *non-reproducible*. However, we can think of defining a probability that these events happen, which in a way encodes our *degree of belief* that the event will happen. Notice that such a degree of belief is personal and may vary from individual to individual (a Spaniard may give a higher

probability of Spain winning a World Cup than a foreigner). For this reason, this interpretation of probability is usually referred to as *subjective*.

A classical way to define probabilities in the subjective framework is in terms of betting odds. Consider an event A. Then $P(A)$, the probability of A, is the amount you are willing to pay for a lottery ticket that pays one Euro if A happens. Notice the *you* in the previous sentence. It is your probability and depends on the information available to *you*. We will not focus on this, but it has been proved that if one specifies probabilities in this way, and assuming you choose the numbers in order to win money, then subjective probabilities respect the usual axioms probabilities (probabilities are numerical quantities, defined on a set of 'outcomes,' that are non-negative, additive over mutually exclusive outcomes, and sum to 1 over all possible mutually exclusive outcomes).

The following quote from De Finetti is provocative.

> My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:
>
> PROBABILITY DOES NOT EXIST
>
> The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time, . . . , orFairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs.

This interpretation of probability is the one underlying most Bayesian analyses. In Bayesian statistics unknown quantities (for instance parameters of interest) are random variables and probabilities must be assigned to them by the modeler.

Another approach sees probability as a *logic of plausibility*. This approach was put forward by statisticians as Jeffreys, Cox and Jaynes. According to this interpretation, the rules of probability extend ordinary ("Boolean") logic, where statements are known to be either true or false, to inductive logic, where statements are true or false, but we don't know which. Probability is then a scale used to describe how strongly, based on specific information, we believe a statement to be true. It is a more objective approach in the sense that anyone with the same knowledge should assign the same probabilities. Again we will not focus on this, but then one can prove the probabilities so defined obey the usual axioms.

The following quote from Maxwell gives a justification of this approach.

> They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the

> actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

The above discussion is brief and not at all comprehensive but it should have given you at least an idea of possible different interpretations of probability.

1.6 A review of probability

1.6.1 Random variables

A random variable is defined as an unknown numerical quantity about which we make probability statements.

1.6.1.1 Discrete random variables

Let Y be a random variable and $\mathbb{Y}$ be the set of all possible values of Y. Usually, $\mathbb{Y}$ is called the *sample space*. We say that Y is discrete if $\mathbb{Y}$ is countable, meaning that $\mathbb{Y}$ can be expressed as $\mathbb{Y} = \{y_1, y_2, \dots\}$.

For each $y \in \mathbb{Y}$, we define the *probability density function* (pdf) $p(y) = P(Y = y)$ which must obey the following conditions:

- $p(y) \geq 0$ for all $y \in \mathbb{Y}$;
- $\sum_{y \in \mathbb{Y}} p(y) = 1$.

For any $A \subseteq \mathbb{Y}$, the probability that $Y \in A$ can be computed via summation as

$$P(Y \in A) = \sum_{y \in A} p(y).$$

1.6.1.2 Continuous random variables

If the sample space $\mathbb{Y}$ is an interval, not necessarily finite (for instance it could be the set of all real numbers $\mathbb{R}$), we say that Y is a continuous random variable.

The pdf of a continuous random variable Y is now defined as the function p for which:

$$P(a \leq Y \leq b) = \int_a^b p(y)dy$$

We cannot define $P(a \leq Y \leq b)$ as equal to $\sum_{a \leq y \leq b} p(y)$ because the sum does not make sense (the set of real numbers between a and b is "uncountable"). So pdfs are defined indirectly as the function such that its integral is the probability of the corresponding event.

The pdf of a continuous random variable must again obey two conditions:

- $p(y) \geq 0$, for all $y \in \mathbb{Y}$;
- $\int_{y \in \mathbb{R}} p(y)dy = 1$.

We can see that integration for continuous distributions behaves similarly to summation for discrete distributions. In fact, integration can be thought of as a generalization of summation for situations in which the sample space is not countable. However, unlike a pdf in the discrete case, the pdf for a continuous random variable is:

- Not necessarily less than 1;
- $p(y)$ is not "the probability that $Y = y$";
- such that $P(Y = y) = 0$ for any $y \in \mathbb{Y}$.

However, if $p(y_1) > p(y_2)$ we will sometimes informally say that y_1 "has a higher probability" than y_2.

1.6.2 Joint distributions

1.6.2.1 Discrete case

Let Y_1 and Y_2 be two discrete random variables with $\mathbb{Y}_1$ and $\mathbb{Y}_2$ as sample spaces, respectively. We are often interested in joint beliefs about two variables. For instance we may want to know

$$P(Y_1 = y_1, Y_2 = y_2), \qquad \text{for } y_1 \in \mathbb{Y}_1 \text{ and } y_2 \in \mathbb{Y}_2$$

The (joint) pdf is then $p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$.

The *marginal density* of Y_1 can be computed from the joint as:

$$p(y_1) = P(Y_1 = y_1) = \sum_{y_2 \in \mathbb{Y}_2} P(Y_1 = y_1, Y_2 = y_2) = \sum_{y_2 \in \mathbb{Y}_2} p(y_1, y_2)$$

The *conditional density* of Y_2 given $Y_1 = y_1$ can be computed from the joint and the marginal as:

$$p(y_2|Y_1 = y_1) = \frac{p(y_1, y_2)}{p(y_1)}$$

Re-arranging the above equation we can also see that

$$p(y_1, y_2) = p(y_2|Y_1 = y_1)p(y_1)$$

Let's consider an example. Suppose we have a random variable Y_1 which is the outcome of a COVID-19 test (either positive or negative) and Y_2 which is whether an individual has COVID or not (sick or healthy). The joint probability distribution is defined as

	sick	healthy
positive	0.10	0.09
negative	0.01	0.80

The above table is a joint pdf since all numbers are positive and sum to one. The marginal probability of Y_1 (test result) is the row sums: 19% of tests are positive and 81% of tests are negative. The marginal probability of Y_2 is the column sums: 11% of individuals are sick and 89% are healthy. Given these marginals we can also compute conditional probabilities:

- $p(\text{positive}|\text{sick}) = \frac{p(\text{positive,sick})}{p(\text{sick})} = \frac{0.10}{0.11} = 0.91$
- therefore $p(\text{negative}|\text{sick}) = 0.09$.

Other conditional probabilities can be derived similarly.

1.6.2.2 Continuous case

In the continuous case the definitions are analogous but summations are substituted with integrals. For instance,

$$p(y_1) = \int_{y_2 \in \mathbb{Y}_2} p(y_1, y_2) dy_2$$

1.6.3 Expectation and variance

The mean or expectation of a random variable Y is

- $E(Y) = \sum_{y \in \mathbb{Y}} yp(y)$ in the discrete case;
- $E(Y) = \int_{y \in \mathbb{Y}} yp(y)dy$ in the continuous case.

The mean is the center of mass of the distribution.

In addition to the location of a distribution we are often interested in how spread out it is. The most popular measure of spread is the variance of a distribution:

$$V(Y) = E((Y - E(Y))^2) = E(Y^2) - E(Y)^2$$

The variance is the average squared distance between values of Y and its mean $E(Y)$. The standard deviation is the square root of the variance, and is on the same scale as Y.

Above we have defined conditional densities of the form $p(y_2|Y_1 = y_1)$. Since formally they are pdfs they have expectation and variance which are usually denoted as $E(Y_2|Y_1 = y_1)$ and $V(Y_2|Y_1 = y_1)$.

Furthermore, it is often useful to express the mean and variance of a random variable Y_2 in terms of the conditional mean and variance given some related quantity Y_1. The mean of Y_2 can be obtained by averaging the conditional mean over the marginal distribution of Y_1:

$$E(Y_2) = E(E(Y_2|Y_1 = y_1)).$$

Let's see why this is true:

$$\begin{aligned}
E(Y_2) &= \sum_{y_2 \in \mathbb{Y}_2} y_2 p(y_2) \\
&= \sum_{y_2 \in \mathbb{Y}_2} y_2 \sum_{y_1 \in \mathbb{Y}_1} p(y_1, y_2) \\
&= \sum_{y_2 \in \mathbb{Y}_2} \sum_{y_1 \in \mathbb{Y}_1} y_2 p(y_2|Y_1 = y_1) p(y_1) \\
&= \sum_{y_1 \in \mathbb{Y}_1} p(y_1) \sum_{y_2 \in \mathbb{Y}_2} y_2 p(y_2|Y_1 = y_1) \\
&= \sum_{y_1 \in \mathbb{Y}_1} p(y_1) E(Y_2|Y_1 = y_1) \\
&= E(E(Y_2|Y_1 = y_1))
\end{aligned}$$

The corresponding result for the variance includes two terms, the mean of the conditional variance and the variance of the conditional mean:

$$V(Y_2) = E(V(Y_2|Y_1 = y_1)) + V(E(Y_2|Y_1 = y_1))$$

1.6.4 Independence

We say that two random variables are independent if

$$p(y_1, y_2) = p(y_1)p(y_2)$$

that is if the joint distribution can be written as the product of the marginal distributions.

In Bayesian statistics we usually make the assumption that we have independent random variables $Y_1, Y_2, \ldots, Y_N$ which all depend on a parameter θ, which is also believed to be random variable. We say that $Y_1, \ldots, Y_N$ are conditionally independent given θ if

$$p(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} p(y_i|\theta).$$

Another way to say this is that $Y_1, \ldots, Y_N$ are conditionally independent and identically distributed (iid).

Notice that in standard statistical practice when we analyze a sample we start with the assumption that the data are realizations of independent and identically distributed random variables (not conditionally, since the parameter is not considered random).

1.7 Exchangeability

In many situations with several random variables, we would believe that the specific order of observation of these random variables is not important. For example, consider a random sample of 3 participants from an infinite population which may or may not have a property (1 or 0) It makes sense that

$$p(1, 0, 0) = p(0, 1, 0) = p(0, 0, 1).$$

Such a property is called *exchangeability*.

Let $Y_1, \ldots, Y_N$ be random variables. If $p(y_1, \ldots, y_N) = p(y_{\pi_1}, \ldots, y_{\pi_N})$ for all permutations π of $\{1, \ldots, N\}$, then $Y_1, \ldots, Y_N$ are exchangeable.

Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels convey no information about the outcomes.

The following result (often called De Finetti's Theorem) tells us that it is sufficient to assume exchangeability for random variables to be (conditionally) iid.

$$Y_1, \ldots, Y_N \text{ are iid } \iff Y_1, \ldots, Y_n \text{ are exchangeable for all N's}$$

So notice that in Bayesian statistics we are actually starting from a weaker assumption about the data-generating process.

1.8 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 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.

Chapter 2

The Binomial Model

We now start looking at actual Bayesian inference for a variety of data types and models, starting from the simplest case of binary outcomes.

Consider the following motivating example. A survey conducted in 2020 asked 2000 Spaniards whether they were happy or not. 920 of the respondents said they were indeed happy. Such a situation could be modeled using a *Binomial* model if we were to believe the following two assumptions:

- each respondent has a same unknown probability θ of replying yes;
- each respondent is independent of all others.

Under these assumptions, let Y be the random variable denoting the number of respondents that said they were happy. Then

$$P(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

by recalling the form of the Binomial probabilities. In general we say that Y follows a Binomial distribution with parameters n and θ if its pdf can be written as above. Then we also have that $E(Y) = n\theta$ and $V(Y) = n\theta(1-\theta)$.

Notice that we explicitly write $P(Y = y|\theta)$, meaning that these probabilities are conditional on an unknown parameter θ denoting the probability that an individual is happy. The number n is not considered random - the number of people interviewed. This is indeed usually fixed by design choices.

Assume that we fixed $\theta = 0.5$, an individual is equally likely to be happy/unhappy. Then the probability of observing 920 happy individuals out of 2000 is:

$$P(Y = 920|\theta = 0.5) = \binom{2000}{920}0.5^{920}(1-0.5)^{1080}.$$

Using R this number can be computed as

```
dbinom(920, size = 2000, prob = 0.5)
```

```
## [1] 2.953299e-05
```

The overall distribution of a Binomial with parameters $n = 2000$ and $\theta = 0.5$ is reported in Figure 2.1.

```
x <- 0:2000
qplot(x, ymax = dbinom(x, size = 2000, prob = 0.5), ymin = 0,
      geom = "linerange",  xlab = "number of successes", ylab = "probability") +
  theme_bw()
```

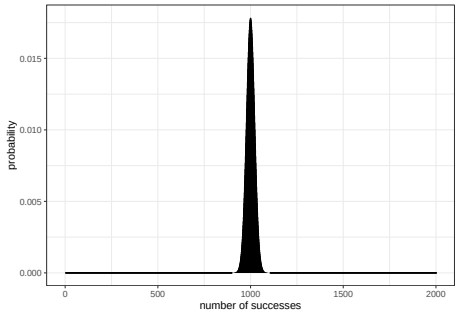

Figure 2.1: Probability density function of a Binomial with parameters 2000 and 0.5

Our aim in this chapter is to develop methods to answer the following question: given a sample $y_1, \dots, y_N$ of independent and identically distributed binary outcomes (just as in our happiness survey) and some prior distribution $p(\theta)$ for the parameter of success θ, what are our posterior beliefs $p(\theta|y_1, \dots, y_n)$ and how can we summarize them?

2.1 Inference Using a Uniform Prior

Let's consider again our happiness survey. We collected a sample $y_1, \dots, y_{2000}$ of binary outcomes happy/unhappy (1/0) of which 920 replied they were happy. The likelihood of this data is therefore

$$p(y|\theta) = \binom{n}{\sum_{i=1}^n y_i} \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n-\sum_{i=1}^n y_i}$$

In order to complete our model definition we also need to define a prior distribution for the parameter θ, which takes values between zero and one.

Let's start choosing a uniform prior distribution between zero and one, that is:

$$p(\theta) = \begin{cases} 1, & 0 \leq \theta \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

and reported in Figure 2.2.

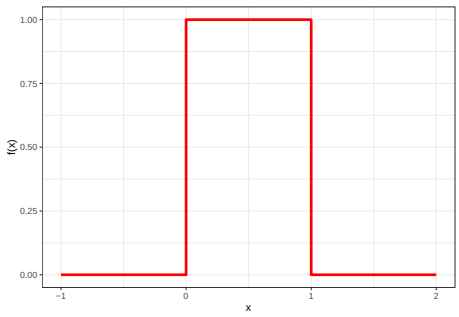

Figure 2.2: Probability density function of the Uniform between zero and one

The assumption of a Uniform distribution implies that the same probability is given to any subinterval of $[0, 1]$ of the same length. This is the simplest prior distribution we can choose.

Given these prior and data-generating process our posterior is:

$$\begin{aligned} p(\theta|y_1, \dots, y_{2000}) &= \frac{p(y_1, \dots, y_{2000}|\theta)p(\theta)}{p(y_1, \dots, y_{2000})} \\ &= \frac{p(y_1, \dots, y_{2000}|\theta) \cdot 1}{p(y_1, \dots, y_{2000})} \\ &\propto p(y_1, \dots, y_{2000}|\theta) \\ &= \binom{n}{\sum_{i=1}^n y_i} \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \\ &\propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \\ &= \theta^{920}(1-\theta)^{1080} \end{aligned}$$

The expression $\theta^{920}(1-\theta)^{1080}$ is proportional to the posterior distribution $p(\theta|y_1, \dots, y_{2000})$, meaning that it does not integrate to one. Using results from calculus it can be indeed shown that

$$\int_0^1 \theta^{920}(1-\theta)^{1080} d\theta = \frac{\Gamma(921)\Gamma(1081)}{\Gamma(921+1081)}$$

where $\Gamma(\cdot)$ is the so-called Gamma function (its value for any $x > 0$ can be computed in `R` using `gamma(x)`).

How is the above integral result useful to derive the form of the posterior? Recall that the posterior is

$$p(\theta|y_1, \dots, y_{2000}) = \theta^{920}(1-\theta)^{1080}\frac{1}{p(y_1, \dots, y_{2000})},$$

and it must be such that

$$\int_0^1 p(\theta|y_1, \dots, y_{2000})d\theta = 1.$$

Using everything that we have learned so far we can then deduce that

$$\begin{aligned} 1 &= \int_0^1 p(\theta|y_1, \dots, y_{2000})d\theta \\ &= \int_0^1 \theta^{920}(1-\theta)^{1080}\frac{1}{p(y_1, \dots, y_{2000})}d\theta \\ &= \frac{1}{p(y_1, \dots, y_{2000})}\int_0^1 \theta^{920}(1-\theta)^{1080}d\theta \\ &= \frac{1}{p(y_1, \dots, y_{2000})}\frac{\Gamma(921)\Gamma(1081)}{\Gamma(921+1081)} \end{aligned}$$

This implies that what we called the marginal likelihood

$$p(y_1, \dots, y_{2000}) = \frac{\Gamma(921)\Gamma(1081)}{\Gamma(921+1081)}$$

and therefore the posterior is

$$p(\theta|y_1, \dots, y_{2000}) = \frac{\Gamma(921+1081)}{\Gamma(921)\Gamma(1081)}\theta^{920}(1-\theta)^{1080}.$$

Although you probably have never seen this expression, the above is the density of the so-called *Beta* distribution, which will be formally introduced next.

Before this, let's consider Figure 2.3. The prior distribution is reported by the blue line and is the flat uniform. Given that we observed 920 individuals who are happy our posterior distribution now reflects the information in the sample and it peaks around the sample proportion 920/2000=0.46. Furthermore, the variance has decreased and the density is concentrated around the sample proportion.

```
ggplot(data.frame(x=c(0,1)), aes(x)) +
  stat_function(fun= function(x) dbeta(x, 921, 1081),aes(colour="Posterior"))+
   stat_function(fun= function(x) dunif(x),aes(colour="Prior"))+
  theme_bw() + ylab("p(theta)") + xlab("theta")
```

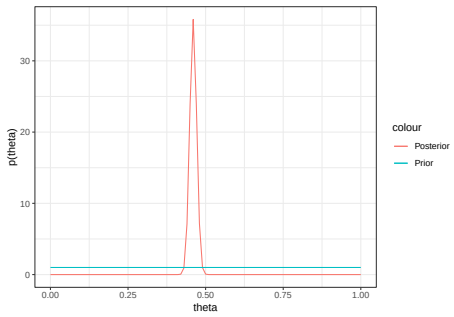

Figure 2.3: Prior and posterior distribution for the happiness survey

2.2 The Beta Distribution

A random variable θ is said to follow the Beta distributions with parameters a and b if its pdf is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

for $\theta \in [0,1]$. The pdf for various choices of a and b is reported in Figure 2.4. Importantly, we can see that the Uniform distribution is a special case of the Beta distribution when parameters are fixed to $a = 1$ and $b = 1$. Indeed,

$$p(\theta) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^0(1-\theta)^0 = \frac{1}{1\cdot 1}1\cdot 1 = 1,$$

since $\Gamma(x) = x!$ if $x > 1$ and integer, and $\Gamma(1) = 1$.

```
ggplot(data.frame(x=seq(0,1,0.01)), aes(x)) +
  stat_function(fun= function(x) dbeta(x,0.5,0.5),aes(colour="a = 0.5, b = 0.5"))+
     stat_function(fun= function(x) dbeta(x,1,1),aes(colour="a = 1, b = 1"))+
    stat_function(fun= function(x) dbeta(x,1,3),aes(colour="a = 1, b = 3"))+
    stat_function(fun= function(x) dbeta(x,5,2),aes(colour="a = 5, b = 2"))+
  theme_bw() + ylab("p(theta)") + xlab("theta") + labs(colour = "Parameters")
```

If θ follows a Beta random variable with parameters a and b, one can prove that

- $E(\theta) = \frac{a}{a+b}$
- $mode(\theta) = \frac{a-1}{(a-1)+(b-1)}$ if $a > 1$ and $b > 1$
- $V(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$

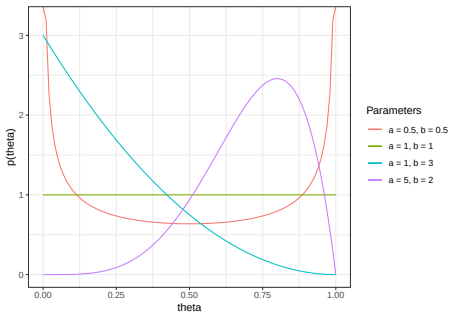

Figure 2.4: Density of the Beta distribution for various choices of parameters.

2.3 Inference using a Beta Prior

Now we turn our attention to cases where the parameter θ is given a Beta prior distribution. First let's revisit the example of the prior uniform distribution.

2.3.1 Uniform as Beta

We have a sample $y_1, \dots, y_n$ of independent and identically distributed (same probability of success) binary outcomes, so that $p(y_1, \dots, y_n|\theta)$ is Binomial with parameters n and θ. The prior distribution for θ is uniform, which is equivalent to a Beta distribution with parameters $a = 1$ and $b = 1$.

Then the posterior is such that

$$p(\theta|y_1, \dots, y_n) \propto \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} = \theta^y(1-\theta)^{n-y} \tag{2.1}$$

where for simplicity we called $\sum_{i=1}^n y_i = y$, the number of successes. The above expression is a function of θ and we can spot that it has the elements of a Beta distribution: it is only missing the ratio $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. However, all terms involving θ of the pdf of a Beta are in the expression. In particular, the above expression must be proportional to the density of a Beta with parameter $a = y + 1$ and $b = n - y + 1$. Notice in particular that the parameter a of the posterior is the number of successes plus the parameter $a = 1$ of the uniform prior and the parameter b of the posterior is the number of failures plus the parameter $b = 1$ of the prior.

Let's consider again the happiness survey data, where 920 individuals said that they were happy out of 2000. Using the properties of the Beta distribution, we then have that

- $E(\theta|y_1, \dots, y_n) = \frac{921}{2002} = 0.46004$;
- $mode(\theta|y_1, \dots, y_n) = \frac{920}{2000} = 0.46$;

- $V(\theta|y_1, \ldots, y_n) = 0.00012$

So we actually derived the posterior distribution, which is Beta with parameters $y+1$ and $n-y+1$, by looking at expression (2.1) and recognizing that it must be proportional to the known Beta distribution. Furthermore, we also recognized the value of the parameters by comparing the expressions. This is a trick we will use multiple times which allows us to derive the posterior by only looking at

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

instead of

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

which requires the computation of $p(y)$ often involving complex integration.

2.3.2 A Generic Beta Prior

Let's now take a more general approach and let's suppose that θ is given a prior distribution $p(\theta)$ which is Beta with some parameters a and b. What is the form of the posterior?

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\theta)p(\theta) \\ &= \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{n-y+b-1} \end{aligned}$$

Using the same trick as before, we notice that the above expression is proportional to the density of a Beta distribution with parameters $y+a$ and $n-y+b$. So we started with a prior Beta distribution and our posterior is again Beta. Such a property of a prior distribution $p(\theta)$ and a data-generating process $p(y|\theta)$ is usually referred to as *conjugacy*.

A class of prior distributions $\mathcal{P}$ for θ is said to be **conjugate** for a data-generating process $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

So the Beta distribution is the conjugate prior to the Binomial: a Beta prior combined to a Binomial likelihood gives a posterior distribution which is again Beta.

Using the properties of the Beta distribution we can then derive that:

- $E(\theta|y_1,\dots,y_n) = \frac{a+y}{a+b+n}$
- $mode(\theta|y_1,\dots,y_n) = \frac{a+y-1}{a+b+n-2}$
- $V(\theta|y_1,\dots,y_n) = \frac{(a+y)(n-y+b)}{(n+a+b+1)(n+a+b)^2}$

Let's look into the posterior mean more carefully. We can rewrite it as

$$\begin{aligned} E(\theta|y_1,\dots,y_n) &= \frac{a+y}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ &= \frac{a+b}{a+b+n}E(\theta) + \frac{n}{a+b+n}\bar{y}_n, \end{aligned}$$

where $\bar{y}_n$ is the sample mean.

So the posterior mean is a weighted average between the prior mean $E(\theta)$ and the sample mean $\bar{y}_n$. By looking at the weights we can also see that if $n >> a+b$ then $\frac{a+b}{a+b+n} \approx 0$ and $\frac{n}{a+b+n} \approx 1$ and therefore the posterior mean is equal to the sample mean. This means that if the sample size is very large, then our posterior beliefs are mostly driven by the data.

The parameters of the prior Beta distribution can be interpreted as follows:

- $a+b$ is a prior sample size: the larger this number the more emphasis we want to put on the prior;
- a is the prior number of successes;
- b is the prior number of failures;

Figure 2.5 illustrates the effect of various parameter choices a and b for the prior distribution.

2.4 Predictive Distribution

An important feature of Bayesian inference is the existence of a *predictive distribution* which gives the probability of observing a specific new observation given that we have already observed a sample $y_1,\dots,y_n$, that is $p(\tilde{y}|y_1,\dots,y_n)$ where $\tilde{y}$ is a possible value of the random variable of interest. In the specific case of a binary outcome, we can compute $p(\tilde{y}=1|y_1,\dots,y_n)$ the probability of observing a success given that we observed the sample.

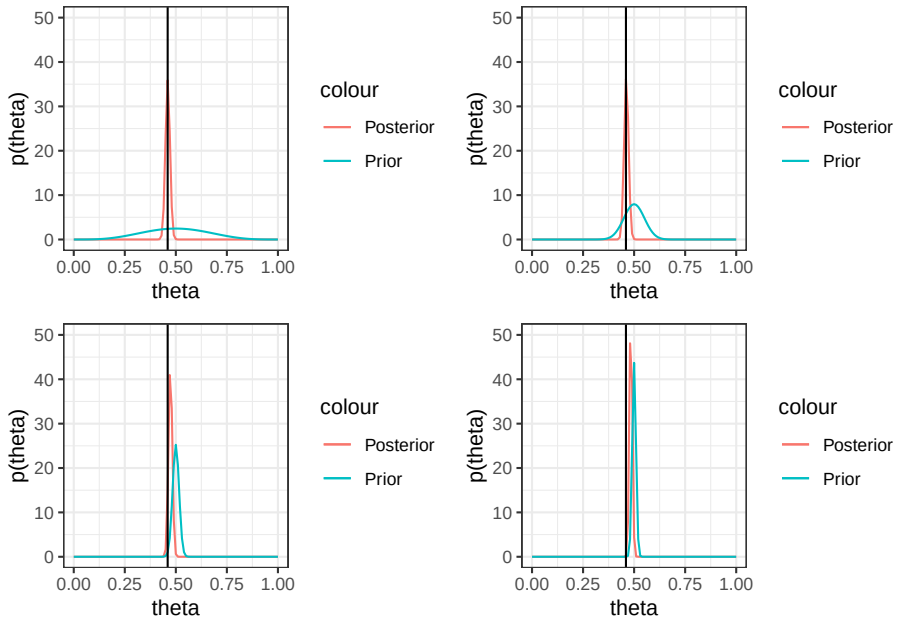

Figure 2.5: Posterior distribution for different prior distributions. Sample proportion - black vertical line. Top left: a = b = 5; Top right: a = b = 50; Bottom left: a = b = 500; Bottom right: a = b = 1500.

Let's compute this probability.

$$\begin{aligned} p(\tilde{y} = 1|y_1, \ldots, y_n) &= \int_0^1 p(\tilde{y} = 1, \theta|y_1, \ldots, y_n)d\theta \\ &= \int_0^1 p(\tilde{y} = 1|\theta, y_1, \ldots, y_n)p(\theta|y_1, \ldots, y_n)d\theta \\ &= \int_0^1 \theta p(\theta|y_1, \ldots, y_n)d\theta \\ &= E(\theta|y_1, \ldots, y_n) \end{aligned}$$

So the predictive probability of a success is equal to the posterior mean in the case of binary outcomes. We will see that in other cases it is not as easy to derive this distribution.