The Basics of Bayesian Computation

So let me go back to the basics in some sense. So Bayes theorem, which is outlined there in equation one gives me a way of computing the probability of observing event A given that I observed already event B. This relationship, equation one is going to be present throughout all the exercises on all the points we'll cover today, although not necessarily in that way, but applying this theorem to random variables rather than to events because, as you know, as Andre mentioned in previous video Bayesian, treat parameters, regression coefficients, missing data, for instance, as random, random variables that have a certain distribution.

So in the context of Bayesian statistical inference, we're going to start from a parameter from which we want to draw inferences. This parameter is a random variable, and thus has a certain distribution, which is denoted by fΘ. And in addition to that parameter, and intuitive distribution, we're going to have data and we're gonna have the node by f(data|Θ) sampling distribution, or roughly the likelihood of it, right. And in the context of Bayesian statistical inference, we're going to work with equation two, which is bayes theorem applied to random variables, and in particular, applied to a parameter of interest. So in equation two, in the numerator, we have fΘ, which is what Bayesians called the prior distribution of the parameters, which in turn says this can be understood as the probability distribution of the parameter before we observe any data, it is obviously a gratification but suppose we start with some prior beliefs about how we became summarise in the probability distribution fΘ we then have f of data given data, which as I said, is the probability distribution of the data which is something you work with. And you are familiar if you have worked with maximum likelihood estimation. And, and what we want to compute is what will be known as the posterior distribution of the posterior distribution of the parameter, f(data|Θ) given the data. So this is the version of the bayes, or the bayes theorem that we're going to work with when we to Bayesian inference.

And from equation two, we have given that f(Θ)as its stated, there is independent of my random variable, which is my parameter, we can simplify this and to work with equation three, which essentially is what I would call a call in the slides, the Bayesian management, which is, the posterior distribution of the parameters of interest is proportional to the likelihood. The data times the prior distribution, we're going to use this relationship throughout the slides throughout the exercises included in this with the slide and throw your Bayesian life, right. The posterior distribution parameter is proportional to the likelihood and the prior. So in some sense, or very informally, when we do basic analysis, we start from a prior distribution, what we believe about the behaviour of a parameter. Before observing the data, we then observe the data and we update our prior beliefs of that parameter once we see the data and arrived at a posterior distribution of the parameter

And the whole purpose of Bayesian inference is exactly deriving the posterior distribution of data starting from the prior distribution. And again, very informally, my beliefs about the parameter for instance, and the likelihood distribution of the data, right. So essentially, I would say that Bayesian inference, the whole big world of Bayesian inference, can be summarised as a process that consists of

four steps. First, we start from a likely explanation of the data and x's and y's, the dependent and independent variables in our models, we then need to specify a prior distribution for the parameters of that model I'm examining, which in this case, I'm summarising all these parameters or denoting all these parameters as data. Once I have the data, I have the prior, and we get step three is perhaps the key step which is deriving the posterior distribution, which is, in some sense, the most difficult step. And finally, once I have once I have derived the posterior distribution, I can summarise it computing posterior mean variances, reaching posterior what we know as posterior summaries.

So let me walk you through a simple example of how these four steps are applied in practice. Suppose you have two candidates, A and B competing in an election. And you're gonna have to excuse me for using a political science example, but I am a political science. So we have a and b two candidates running for office. And suppose we have an opinion poll, based on a representative sample conducted before the election, a poll with 1067 potential voters 556 of which stated they will vote for A 511 declaring they will vote for B. So based on this data, candidate A highers a Bayesian researcher to so that this researcher estimates the probability that A wins.

Okay, so let's go over the four steps of Bayesian analysis in this instance. Step one, segment one is specifying a model for the data. This case, and suppose I am ignoring the representative abstention. I have two possible responses for each individual who participated in this opinion poll, we have two potential answers to the question, Who would you vote for? Vote for A vote for B. So we can think of this? Remember, we have an opinion poll with 1067 responders can think of this as 1067 independent Bernoulli trials, where the success is voting for A voting for candidate A. So we're going to denote by P, the probability of success and by Y to be the choice that individual I makes, and I can take two r values. One, if I will vote for A, or zero, if I will go for B.

So we can think of the sampling distribution of the data as what we observe in equation 4for each individual in the sample has a probability P of success, and 1-p of failure. Again, success being voting for A, failure being voting for B because a Bayesian was hired by A. So, based on the responses to the public opinion poll, we arrived at these distributions later, which we see in equation five. Now, as I said, again, going back to the Bayesian mantra, the posterior distribution of the parameters, in this case, the parameter is going to be P, the probability of success is a simple example with the Bernoulli.

So the parameter of a Bernoulli distribution is p, the probability of success. So going back to the Bayesian mantra, posterior is proportional to the likelihood times the prior. What we have is that is that good exterior distribution of p, even the data is going to be proportional to the sampling distribution times the prior, which I will denote by f (p). And this takes us, this takes us to the second step. Remember I said any base analysis has four steps, step one, specifying the distribution data. Step two, choosing the prior distribution for p.

So we need to choose a prior for p and turns out that a typical prior distribution for parameters and present proportions rate estimated distribution. Specifically, the beta is what we know as a conduit prior distribution. A conduit prior is a prior such that the posterior follows the same distribution of the prior which is, as we will see, in this several examples, is very convenient. So what I'm saying here is that if I assume that beta, that the posterior still don’t know, how it looked like, but I know it will be Beta.

Just a side note, why beta? well, because beta, as i said this use a sub prior for proportions of rights, because it has a very nice property of essentially bounding parameter between zero and one, which is what I want from and you can play with the code there are included with this materials called Beta Distribution.R which allows you to come to different values obtain different beta distributions, depending on what I will call here, the hyper parameters to distribution. So alpha and beta essentially control how the shape of the prior distribution, okay, so here we have a graph of betas with different combinations of parameters. Alpha and Beta, which control the form of the prior are known as hyper parameters. And these hyper parameters will influence how much weight the prior distribution has in the posterior distribution, essentially, is hyper parameters. And this applies to the beta distribution, but to any distribution. In general, the hyper parameters of a prior distribution, in some sense, control how much influence the prior distribution is going to have relative to the data in the determination of the posterior distribution.

Let me give you an example. Suppose I have no prior information at all about alpha and beta, or about the parameter P, and the parameter p being the probability of voting for A, suppose I have no idea, I have no information at all. But one way of incorporating this lack of information into the beta distribution is choosing alpha and beta equal to one. Why? Because as you can see here, if when I choose alpha and beta equal to one, I end up with a beta distribution, that the generated one or another way of putting this is that when I combine my data, the likelihood for the data will beta, prior parameters one and one, it turns out that the prior essentially incorporates no information at all another way of saying is that the posterior is going to be completely determined by the data. This type of or mainly determined by the. So these type of prior distributions that incorporate very little information to the data, essentially, this type of priors in which the posterior form of the posterior is essentially determined by later another prior or called vague, or weakly informative prior. And you can see here, the posterior distribution for P is essentially proportional to the likelihood and the prior is sometimes vanished.

Now Let's go to the opposite extreme. Suppose that in addition to the poll I am currently analysing. There are several other polls that were conducted last month, two months ago, three months ago. And suppose to make things very simple, that I want to incorporate all that information. I think I want to use the result from previous polls and incorporate them in the prior and use that thing to inform, to inform our inferences and essentially, in the derivation in the posterior distribution of the parameters.

Well, one way of doing that is using same the same beta distribution, which I said is a conjugate prior. So having a beta prior leads to a beta posterior, which is nice. I know the form of a beta distribution. But now instead of having parameters alpha and beta equal to one I'm using, I'm specifying the alphas and betas based on or incorporating all information from the previous polls. I will collect connected before dealing with a polling question. So I use those priors. And the posterior now is going to be quite different. It turns out, you're going to be heavily influenced by the prior.

So we are already in step three of Bayesian inference procedure, what we have is that starting from a likelihood for later and choose having chosen the prior distribution for the parameters in this case, p, we arrived at the posterior distribution for the parameter and this posterior distribution is going to be a [inaudible] going to depend on the data but also on the priors. Specifically in this particular example,

whether I use vague priors, which added very little information to the parameter, or I use more informative priors, for instance, incorporating information from previous posts.

Finally, I said the last step of elevation analysis, step one, defining the likelihood, step two, finding the prior to the parameters, step three, deriving the posterior and finally summarising the posterior distribution, because and this is, again, the key difference between Bayesian inference and classical inferences, statistics. I don't get an estimate for B, I don't get a point estimate, as one would typically expect in Bayesian setting, when again, the result of my Bayesian analysis, is this a posterior distribution, which is great. But the client who hired the Bayesian statistician doesn't really understand what it means to posterior distribution. what the client wants, wants to know is what was the probability that I will win the election. So I need to summarise this posterior distribution in some way. I can compute then the posterior mean, I can compute the posterior variance, I can compute credibility interval, right.

So for instance, the posterior mean, if I knew, because I know that the posterior distribution of beta because the prior of beta and beta is a conjugate prior, meaning the posterior is going to have the same beta. I also have a distribution with different parameters, beta distribution, you can apply the formula for the expected value of the distribution to compute the expected value of P. I can compute also credibility intervals. I can compute things like oh, what's the probability that my client A wins? Well, that's a probability p exceeds point five. What's the probability that my client A loses as well as probability that B relative success is less or equal than point five can from the posterior distribution, I can compute all those quantities.

And essentially, that's a first exercise in Bayesian computation. I just want to show you that, obviously, how close the posterior is the prior. Or the other way of saying how much importance the prior with a bit of data has on the posterior depends on my choice of priors. So remember that, in one case, I use informative priors, meaning I incorporated the information from several previous polls into analysis in this case, turns out that the posterior is very close to the prime. However, when I use vague or weakly informative priors, alpha and beta equal to one, my example, turns out that the posterior is closer to the data way to say this is if we use vague priors or weakly informative priors, the form of a posterior is going to be primarily influenced by the data.