Example Bayes inference

A Bayesian statistical analysis has the following set of steps:

- First in the dataset, identify the outcome measure which is of your interest. This will be the response. Then some of the other covariates could be considered as predictors. You may not have any predictor in your model which will then be an unsupervised model.
- Identify the distribution suitable for that outcome measure. The "distribution" part is not always required for frequentist scheme of inference, but an absolute requirement for Bayes.
- The choice of distribution could be guided by a common scheme, educated assumption, etc. For example, when the outcome measure is continuous, normal distribution is a common choice. For count data, Poisson distribution is a common choice.
- This definitely put unnecessary assumptions to the dataset. But it is the statistician's job to minimize that by introducing more structure. For example, in our case, we have added a predictor. That way, it has more structure than assuming the model to be $M_t \sim$Poisson(τ). Adding predictors adds more structure to the model, as it explicitly assumes a relationship between the response and the predictor(s).
- Next, write down the likelihood using the assumed distribution.
- It is good to simplify the Likelihood expression first and then go to the prior distribution selection phase as we see in the example. The expression of the likelihood will guide us appropriate choice of the prior distribution.
- Then choose a prior distribution for the parameter to follow a Bayesian route for inference. So, a 'prior' is essentially a distribution too (in most cases; we will see some contradictions later).
- The next step is computation of the posterior.
- If the posterior distribution belongs to a standard family of distribution, for example from the distribution.pdf, life is good. We can directly apply the properties of that known distribution such as formula for mean, median, mode, variance, etc. Otherwise, we need to generate samples from the posterior distribution to 'approximately' learn the distribution. Then compute mean, median, variance etc from the samples.

In our exercise: Response is the male death and the associated observations are $M_1, \ldots, M_T$.

The predictor is the total death and the associated observations are $N_1, \ldots, N_T$. (Note that $T = 72$)

The inference question is "what proportion of total deaths is male?" The parameter is τ.

Since, M_t's ($t = 1, \ldots, T$) are count valued, I suggested to fit a Poisson model for this data and to appropriately describe the inference question, I suggested the (mean) parameter of Poisson to be τN_t. Then we have under the Poisson model $\mathbb{E}(M_t) = \tau N_t$.

Now the next part is writing down the likelihood which is the joint probability of the (response) data given the parameters i.e. for the problem specifically $P(M_1, \ldots, M_T \mid \tau) = L(\tau \mid M_1, \ldots, M_T) = \prod_{t=1}^{T} P(M_t \mid \tau)$, due to the assumed independence of the data. Now we use the Poisson likelihood to write down $\prod_{t=1}^{T} P(M_t \mid \tau)$.

And we also need to simplify the product expression little bit to understand it's nature $\prod_{t=1}^{T} P(M_t \mid \tau) = \prod_{t=1}^{T} \frac{e^{-\tau N_t}(\tau N_t)^{M_t}}{M_t!} = e^{-\tau \sum_{t=1}^{T} N_t} \tau^{\sum_{t=1}^{T} M_t} \prod_{t=1}^{T} \frac{1}{M_t!} \mathbf{1}_{0 \le \tau \le 1}$ (verify!)

If we now only consider the terms involving the parameter/parameters (τ in this case), those terms are $e^{-\tau \sum_{t=1}^{T} N_t} \tau^{\sum_{t=1}^{T} M_t} \mathbf{1}_{0 \le \tau \le 1}$.

This above simplified likelihood will now guide us the selection of prior step. Note that in any problem, after simplifying the likelihood, the expression, you get may or may not resemble some known class of density. If it does, the favorable choice is to pick that distribution as your prior distribution. If it doesn't, the prior selection may be a bit tricky. Here, 'favorable' means computationally favorable.

In our case, $e^{-\tau \sum_{t=1}^{T} N_t} \tau^{\sum_{t=1}^{T} M_t}$-expression looks like a Gamma distribution as the density function of Gamma$(\alpha, \beta)$is $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$. If we replace x with τ, this above expression will look very similar to $e^{-\tau \sum_{t=1}^{T} N_t} \tau^{\sum_{t=1}^{T} M_t}$ with $\alpha = \sum_{t=1}^{T} M_t + 1$ and $\beta = \sum_{t=1}^{T} N_t$.

[The second parameter of the Gamma distribution can be interpreted as a rate parameter or a scale parameter. Depending on its interpretation, the expression for the probability density function will change. In this answer as well as in the class, I always refer to the second parameter as rate parameter. The above mentioned $f(x)$ corresponds to this definition.]

Hence, Gamma(α_0, β_0) should be a good prior choice. Now, the question requires $\mathbb{E}(\tau) = 0.5$ and $V(\tau) = 0.25$ (these requirements are for the prior). If $\tau \sim$Gamma(α_0, β_0), we then need $\mathbb{E}(\tau)\alpha_0/\beta_0 = 0.5$ and $V(\tau)\alpha_0/\beta_0^2 = 0.25$. Solving these two equations, $\alpha_0 = 1$ and $\beta_0 = 2$ is a reasonable choice.

With this prior our posterior

$$P(\tau \mid M_1, \dots, M_T) \propto \prod_{t=1}^{T} P(M_t \mid \tau)\tau^{\alpha_0 - 1}e^{-\tau\beta_0} \propto \tau^{\alpha_0 + \sum_{t=1}^{T} M_t - 1}e^{-\tau(\beta_0 + \sum_{t=1}^{T} N_t)}\mathbf{1}_{0 \leq \tau \leq 1},$$

here I am periodically ignoring the terms not involving τ. If there is confusion while following these steps, you should write down the whole expressions and then drop the terms not involving τ and check if you arrive at the same expression as above.

We thus arrive at the above expression which looks like a Gamma distribution's probability density function (p.d.f.) only without the $\mathbf{1}_{0 \leq \tau \leq 1}$ part. Hence, our posterior is a Truncated-Gamma distribution with parameters $\bar{\alpha} = \alpha_0 + \sum_{t=1}^{T} M_t$ and $\bar{\beta} = \beta_0 + \sum_{t=1}^{T} N_t$. The 'Truncated' part is introduced due to the part $\mathbf{1}_{0 \leq \tau \leq 1}$ which forces the distribution to be truncated between 0 and 1.

```
ldeaths
fdeaths
mdeaths

##Parameters of the posterior
alphabar <- 1+sum(mdeaths)
betabar  <- 2+sum(ldeaths)

##Generate from posterior
posteriorsample <- rgamma(1000, alphabar, betabar)

##We now only consider the samples that satisfies the truncation condition
posteriorsample <- posteriorsample[which(posteriorsample<=1)]

##Posterior mean
mean(posteriorsample)

##Equal tail 95% credible interval
quantile(posteriorsample, prob=c(0.025, 0.975))
```

*We can approximately learn any theoretical quantity from samples. In the above example, we could have mathematically derived the posterior mean which is $\mathbb{E}(\tau \mid M_1, \dots, M_T) = \int \tau P(\tau \mid M_1, \dots, M_T)$. But this computation could be time-consuming and tedious. Instead, we compute this from samples, i.e. we obtain a 'sample' estimate of the true theoretical mean.

Even while computing the credible intervals (a, b), actually they are the solutions of 1) $P(\tau \leq a \mid M_1, \dots, M_T) = \alpha/2 = 0.05/2 = 0.025$ and 2) $P(\tau \leq b \mid M_1, \dots, M_T) = 1 - \alpha/2 = 1 - 0.05/2 = 0.975$ (as we are looking at 95% interval, thus $\alpha = 1 - 0.95 = 0.05$ by definition.) Solving 1) and 2) can be hard and difficult, thus we can again get sample estimates of a and b. This is done using the `quantile` function in R.

Even in a frequentist setting these days, bootstrap approaches are widely popular. Bootstrap approaches also rely on sampling, but of the observations themselves. So it's called a resampling approach. The end idea is kind of similar.