Introduction to Bayesian Inference

Chase Joyner

Mathematical Sciences

June 3, 2014

1 Introduction

Bayesian inference is a method in which Baye's rule is primarily used in order to obtain a posterior distribution, and this distribution can provide all information on unknown parameters of interest. The benefit of using Bayesian methods rather than Frequentist methods is that instead of obtaining just point estimates and confidence intervals like the Frequentist, a Bayesian obtains an entire distribution for the parameter we wish to gain inference on. For example, suppose you flip a fair coin 100 times and record 64 heads and 36 tails. Would you begin to consider the coin to be bias? As you can see, using a Frequentist approach requires larger sample sizes to obtain a long-run frequency. This leads us towards a Bayesian approach, where we can include some prior knowledge of the coin to assess its fairness.

2 Bayesian Inference

So how do we gain inference on unknown parameter(s) of interest? The goal is to obtain a posterior distribution. Bayesian techniques use prior information on the parameter to specify a prior distribution, such as $p(\theta)$, and a likelihood function specified by the data, such as $p(\mathbf{y}|\theta)$. After obtaining these requirements, we can apply Baye's rule to formulate a posterior distribution as follows [1, p. 2]

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

$$\propto p(\mathbf{y}|\theta)p(\theta)$$

Generally, we can simplify the work by finding the proportional distribution to the posterior, and then integrating over the support while setting equal to 1 to find the normalizing constant. A lot of times the posterior distribution $p(\theta|\mathbf{y})$ is not of a known form, so it is difficult to gain inference on the parameters using this unknown density; however, if we specify a conjugate prior distribution to the likelihood function, it will ensure a recognizable posterior distribution.

Conjugate Priors

In Bayesian analysis, a conjugate prior for the likelihood function is when the prior distribution and the posterior distribution are both of the same family [1, p. 38]. This result is ideal as it guarantees a known posterior form.

Beta-Binomial

Suppose we have independent and identically distributed (iid) data that follows a $Bin(1, \theta)$, this is our likelihood function. If we use a Beta prior, then we will obtain a Beta posterior distribution.

$$\text{prior: } p(\theta) = \frac{\gamma(a+b)}{\gamma(a)\gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$
$$\text{likelihood: } p(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1} \end{aligned}$$

Notice the posterior distribution is a Beta, indicating $\theta|y \sim Beta(a+y, b+n-y)$. Therefore, the conjugate prior for the Binomial likelihood is a Beta distribution.

Normal-Normal

Assume we have n iid samples from a $N(\theta, \sigma^2)$, where σ^2 is known. Let the prior distribution for $\theta \sim N(\mu, \sigma_0^2)$.

$$\begin{aligned}
p(\theta) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\{-\frac{1}{2\sigma_0^2}(\theta - \mu)^2\} \\
p(\mathbf{y}|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\} \\
&\propto exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\}
\end{aligned}$$

$$\begin{aligned}
p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta)p(\theta) \\
&\propto exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\}exp\{-\frac{1}{2\sigma_0^2}(\theta - \mu)^2\} \\
&\propto exp\{-\frac{1}{2\sigma^2\sigma_0^2}[\theta^2(2\sigma_0^2 n + 2\sigma^2) + \theta(-2\sigma_0^2 2\sum_{i=1}^{n} y_i - 2\sigma^2 2\mu)]\} \\
&= exp\left\{ - \frac{1}{2\frac{1}{\frac{1}{\sigma_o^2} + \frac{n}{\sigma^2}}} \left(\theta - \frac{\frac{\mu}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \right)^2 \right\}
\end{aligned}$$

Therefore, the posterior $\theta|\mathbf{y} \sim N\left(\frac{\frac{\mu}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_o^2} + \frac{n}{\sigma^2}} \right)$. You can refer to the appendix for a complete derivation.

Gamma-Poisson

Another example for conjugate priors is the Gamma prior distribution for the Poisson likelihood function. If our data follows a Poisson distribution and we specify a Gamma prior distribution, then our posterior distribution will also be Gamma. This is shown below:

$$\begin{aligned}
\text{prior: } p(\theta) &= \frac{\beta^\alpha}{\gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\
\text{likelihood: } p(\mathbf{y}|\theta) &\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}
\end{aligned}$$

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta)p(\theta) \\ &\propto \theta^{\sum\limits_{i=1}^{n} y_i} e^{-n\theta}\theta^{\alpha-1}e^{-\beta\theta} \\ &= \theta^{(\alpha+\sum\limits_{i=1}^{n} y_i)-1} e^{-(n+\beta)\theta} \end{aligned}$$

Here we see the posterior is Gamma$(\alpha + \sum\limits_{i=1}^{n} y_i, n + \beta)$.

3 Two-Parameter Models

So far we have looked at Bayesian inference in one-parameter models, but now let us look at how to conduct inference in two-parameter models. If $\boldsymbol{\phi}$ is a vector of parameters, then we simply apply the same technique as in the one-parameter case:

$$\begin{aligned} p(\boldsymbol{\phi}|\mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p(\mathbf{y})} \\ &\propto p(\mathbf{y}|\boldsymbol{\phi})p(\boldsymbol{\phi}) \end{aligned}$$

Now us look at how to gain joint inference on the mean and the variance in a Normal model. Note, conjugate priors are also applicable in two-parameter models.

Joint Inference in a Normal Model

Suppose we are interested in joint inference on the mean θ and the variance σ^2. Similar to an example shown in [1, p. 74], the priors and the likelihood function are identified as:

$$\begin{aligned} y_i &\sim N(\theta, \sigma^2) \\ \theta|\sigma^2 &\sim N(\mu_0, \tfrac{\sigma^2}{\kappa_0}) \\ \sigma^2 &\sim IG(\tfrac{\nu_0}{2}, \tfrac{\nu_0}{2}\sigma_0^2) \end{aligned}$$

Note that we have two priors instead of a single, joint prior. This is simply because of the fact that $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$. However, with these priors and this likelihood, the underlying joint posterior distribution is not of a known form and renders sampling difficult. This leads us towards an approach called Gibbs Sampling, and with this we can approximate the posterior distribution empirically.

Gibbs Sampling

The idea behind Gibbs Sampling is to generate a sequence of samples of the unknown parameters, using the full conditional posterior distributions of each parameter of interest. To calculate the full conditional posterior distributions, we simply apply the same technique as in the one-parameter models [1, p. 93]

$$p(\theta|\sigma^2, \mathbf{y}) \propto p(\mathbf{y}|\theta, \sigma^2)p(\theta|\sigma^2) \tag{1}$$

$$p(\sigma^2|\theta, \mathbf{y}) \propto p(\mathbf{y}|\theta, \sigma^2)p(\sigma^2|\theta) \tag{2}$$

These posterior distributions are considered full conditional posterior distributions for θ and σ^2, respectively. This is because each distribution is for the parameter of interest, given everything else. With these, we are able to estimate the joint posterior distribution $p(\theta, \sigma^2|\mathbf{y})$ by generating a dependent sequence of parameters. Given a starting point such as $\sigma^{2(0)}$, sample as follows [1, p. 94]

$$\theta^{(1)} \sim p(\theta|\sigma^{2(0)}, \mathbf{y})$$
$$\sigma^{2(1)} \sim p(\sigma^2|\theta^{(1)}, \mathbf{y})$$

Generally, this process takes the form of:

$$\theta^{(t+1)} \sim p(\theta|\sigma^{2(t)}, \mathbf{y})$$
$$\sigma^{2(t+1)} \sim p(\sigma^2|\theta^{(t+1)}, \mathbf{y})$$

And let us update our parameter vector $\boldsymbol{\phi}$ at each iteration, where $\phi_{t+1} = (\theta^{(t+1)}, \sigma^{2(t+1)})$ such that we obtain:

$$\boldsymbol{\phi} = \{\phi_1, \phi_2, ..., \phi_n\}$$

Gibbs sampling is extremely useful as it allows an approximation for a finite sample n, and in fact as $n \to \infty$, these samples form a joint sampling distribution that approaches the joint posterior distribution of interest.

Furthermore, according to the law of large numbers, with these samples we are able to induce properties such as [1, p. 54]

$$\mathrm{E}[\theta|\mathbf{y}] = \tfrac{1}{n}\sum_{i=1}^{n}\theta^{(i)} \text{ as } n \rightarrow \infty$$
$$\mathrm{E}[\sigma^2|\mathbf{y}] = \tfrac{1}{n}\sum_{i=1}^{n}\sigma^{2(i)} \text{ as } n \rightarrow \infty$$

But what if we also cannot obtain the full conditionals proposed in equations (1) and (2) above? If this is the case, then Gibbs Sampling cannot be used and we must take another approach.

Metropolis-Hastings

Suppose that we have a starting position of s initial values for our parameters, such as $\{\theta^{(1)}, ..., \theta^{(s)}\}$ and $\{\sigma^{2(1)}, ..., \sigma^{2(s)}\}$. If we can obtain a new value $\theta^{(*)}$ and $\sigma^{2(*)}$, then by intuition these new values should be included in our set if the densities are greater than or equal to the densities of $\theta^{(s)}$ and $\sigma^{2(s)}$. However, if the densities are not greater than or equal to, then we should accept $\theta^{(*)}$ and $\sigma^{(*)}$ with some probability. Using this basic idea, we compute the acceptance ratio [1, p. 174]

$$r = \frac{p(\theta^{(*)}|\mathbf{y})}{p(\theta^{(s)}|\mathbf{y})} = \frac{p(\mathbf{y}|\theta^{(*)})p(\theta^{(*)})}{p(\mathbf{y})}\frac{p(\mathbf{y})}{p(\mathbf{y}|\theta^{(s)})p(\theta^{(s)})} = \frac{p(\mathbf{y}|\theta^{(*)})p(\theta^{(*)})}{p(\mathbf{y}|\theta^{(s)})p(\theta^{(s)})}$$

Here, $p(\cdot)$ refers to the density proposed by the distribution. After computing r, set

$$\theta^{(s+1)} = \begin{cases} \theta^{(*)} & \text{if } r \geq 1 \\ \\ \theta^{(*)} \text{ or } \theta^{(s)} & \text{with probability } r \text{ and } 1-r \text{ respectively, } r < 1 \end{cases}$$

The second line of the piecewise function above can be achieved by simply sampling $u \sim$ Ber(r) and setting $\theta^{(s+1)} = \theta^{(*)}$ if $u = 1$, $\theta^{(s+1)} = \theta^{(s)}$ otherwise.

Now how do we compute the new value $\theta^{(*)}$? Metropolis proposed to sample the new value from a symmetric distribution around the previous value, namely $J(\theta|\theta^{(s)})$. Some simple examples are

$$\theta^{(*)} \sim \text{uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta) = J(\theta|\theta^{(s)})$$
$$\theta^{(*)} \sim \text{normal}(\theta^{(s)}, \delta^2) = J(\theta|\theta^{(s)})$$

where the choice of δ determines how efficient the algorithm runs. Refer to [1, p. 179] for an explanation on how to choose an efficient δ. As a result from this Metropolis method came the Metropolis-Hastings algorithm, which proposed that the sampling distribution for $\theta^{(*)}$ can be a symmetric distribution as described above, the full conditional distributions (Gibbs Sampling), or some other distribution.

4 Conclusion

Bayesian techniques are extremely useful and have benefits over other methods. Bayesian methods allows the inclusion of prior beliefs, and these prior beliefs may vary from person to person. If you have this prior belief that the coin is fair, then you should include this. As in the example during the introduction, even though it appears that the mean number of heads is 64 out of 100, we still want to use our belief that the coin is fair. Finally, unlike Frequentists who obtain only point estimates, Bayesian inference results in an entire distribution for the parameters of interest.

Appendix A

$$\begin{aligned}
p(\theta) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\{-\frac{1}{2\sigma_0^2}(\theta-\mu)^2\} \\
p(\mathbf{y}|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\} \\
&\propto exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\}
\end{aligned}$$

$$\begin{aligned}
p(\theta|\mathbf{y}) &\propto p(\mathbf{y}|\theta)p(\theta) \\
&\propto exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\} exp\{-\frac{1}{2\sigma_0^2}(\theta-\mu)^2\} \\
&= exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2 - \frac{1}{2\sigma_0^2}(\theta-\mu)^2\} \\
&= exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2) - \frac{1}{2\sigma_0^2}(\theta^2 - 2\theta\mu + \mu^2)\} \\
&\propto exp\{-\frac{1}{2\sigma^2}(n\theta^2 - 2\theta\sum_{i=1}^{n} y_i) - \frac{1}{2\sigma_0^2}(\theta^2 - 2\theta\mu)\} \\
&= exp\{-\frac{1}{2\sigma^2\sigma_0^2}[2\sigma_0^2(n\theta^2 - 2\theta\sum_{i=1}^{n} y_i) + 2\sigma^2(\theta^2 - 2\theta\mu)]\} \\
&= exp\{-\frac{1}{2\sigma^2\sigma_0^2}[\theta^2(2\sigma_0^2 n + 2\sigma^2) + \theta(-2\sigma_0^2 2\sum_{i=1}^{n} y_i - 2\sigma^2 2\mu)]\} \\
&= exp\Bigg\{ - \frac{2\sigma_0^2 n + 2\sigma^2}{2\sigma^2 2\sigma_0^2}\Bigg[\theta^2 - 2\theta\Bigg(\frac{2\sigma_0^2\sum_{i=1}^{n} y_i + 2\sigma^2\mu}{2\sigma_0^2 n + 2\sigma^2}\Bigg)\Bigg]\Bigg\} \\
&\propto exp\Bigg\{ - \frac{2\sigma_0^2 n + 2\sigma^2}{2\sigma^2 2\sigma_0^2}\Bigg(\theta - \frac{2\sigma_0^2\sum_{i=1}^{n} y_i + 2\sigma^2\mu}{2\sigma_0^2 n + 2\sigma^2}\Bigg)^2\Bigg\} \\
&= exp\Bigg\{ - \frac{\sigma_0^2 n + \sigma^2}{2\sigma^2\sigma_0^2}\Bigg(\theta - \frac{\sigma_0^2\sum_{i=1}^{n} y_i + \sigma^2\mu}{\sigma_0^2 n + \sigma^2}\Bigg)^2\Bigg\} \\
&= exp\Bigg\{ - \frac{1}{2\frac{1}{\frac{1}{\sigma_o^2}+\frac{n}{\sigma^2}}}\Bigg(\theta - \frac{\frac{\mu}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\Bigg)^2\Bigg\}
\end{aligned}$$

References

[1] Hoff, Peter D. *A First Course in Bayesian Statistical Methods*. New York: Springer, 2010. Print.