Bayesian Methods of Inference

Tyler Paul
Matt Hruskocy

1 Introduction

There are two schools of thought when it comes to Statistics; the Frequentist school and the Bayesian school. One of the main goals of statistics is to draw inferences about the parameter of a population and both of these two schools use different approaches for doing so. Historically, the Frequentist approach has been the most common approach to drawing inferences. The Frequentist statistician treat the population parameter as unknown (yet fixed) and constructs confidence intervals and point estimators to gain knowledge about the parameter based on a sample from the population. Alternatively, the Bayesian statistician treats the population parameter as a random variable and constructs a probability function which reflects his/her knowledge of the population parameter before any sample is even taken. This distribution function and the sample are then used to draw inferences about the parameter using point estimators and credible intervals. We will now explore the details of Bayesian inference.

2 Bayesian Formalism

Let $Y_1, Y_2, \dots, Y_n$ be independent random variables of an underlying population associated with a sample of size n. Now suppose we take a sample and observe that $Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n$. Suppose we seek to determine the value of a parameter θ of the underlying population. The idea behind the Bayesian approach to statistics is to use our observed values $y_1, \dots, y_n$ along with the knowledge we had regarding θ before a sample was taken, to draw infererences about θ. To do so we treat θ as a random variable and formulate what is called a *prior distribution* of the parameter. This distribution, which we shall denote $g(\theta)$, encompasses our uncertainty of θ. If we are rather certain of the value of θ then we may choose a prior distribution with a mean near our proposed value of θ and with a small variance. If we are rather uncertain then we would need to choose a larger variance of the prior distribution [1].

As stated above, we seek to determine the value of θ by knowing $y_1, y_2, \dots y_n$. To aid in determining this value we seek to determine what is called the *posterior distribution* of θ given that we observed $y_1, y_2, \dots, y_n$. We denote this distribution as $g^*(\theta|y_1, y_2, \dots, y_n)$. Note that the posterior distribution is simply a particular conditional distribution.

We now obtain a formula for determining this distribution. Note that by the definition of conditional probability,

$$g^*(\theta|y_1, y_2, \dots, y_n) = \frac{f(y_1, y_2, \dots, y_n, \theta)}{m(y_1, y_2, \dots, y_n)}$$

where f is the joint distribution of the random variables $Y_1, \dots, Y_n, \theta$ and m is the marginal density of the random variables $Y_1, \dots, Y_n$. However, another application of the definition of conditional probability yields that $f(y_1, y_2, \dots, y_n, \theta) = L(y_1, y_2, \dots, y_n|\theta)g(\theta)$ where L is the likelihood function. Also note that the marginal distribution m, can be obtained simply by $\int_{-\infty}^{\infty} f(y_1, y_2, \dots, y_n, \theta)\mathrm{d}\theta$. In summary, we now have the tools to compute the posterior distribution g^* knowing the prior distribution g and the likelihood function L of the sample $y_1, \dots, y_n$ given a value of θ. Thus

$$g^*(\theta|y_1, y_2, \dots, y_n) = \frac{L(y_1, y_2, \dots, y_n|\theta)g(\theta)}{\int_{-\infty}^{\infty} L(y_1, y_2, \dots, y_n|\theta)g(\theta)\mathrm{d}\theta}.$$

We now consider some examples and explore them using R.

3 Derivation of an Example Posterior Distribution

Suppose that we conduct independent Bernoulli trials with success probability p and record Y, the number of the trial on which the first success occurs. The random variable Y can be shown to have a geometric distribution with parameter p. This fact is easy to see, since the probability that the first success occurs when $Y = y$ means that a failure occured for all positive integers less than y; that is a failure occured $y - 1$ times where each failure had the probability $1 - p$ of occuring. Thus

$$P(Y = y) = (1 - p)^{y-1}p$$

which is precisely the PMF of the geometric distribution. In this example we seek to use Bayesian Statistics to draw inferences about the parameter p. Thus we must first treat p as a random variable and choose a prior distribution for p. Due to the fact that $0 \leq p \leq 1$ then the beta distribution seems to be an adequate prior distribution since the beta distribution has the interval $[0, 1]$ as a support. Thus the prior distribution is given by the beta distribution with parameters α and β:

$$g(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1} \qquad \forall p \in [0, 1].$$

[We can also use the fact that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ where B is the beta function to write the distribution function more compactly. We will make use of this later.]

Note that we can choose α and β in such a way that reflects our subjective knowledge of the actual value of p if we are considering a particular experiment (essentially choosing out prior distribution), but for now we treat α and β as arbitrary. Now that we chose a prior distribution we can work towards determining the posterior distribution. Suppose now that we take a sample of size 1. Thus when we record an observed value of Y we will have obtained a particular value y. We use this to determine the posterior distribution. Note that

$$L(y|p) = (1-p)^{y-1}p.$$

Thus

$$f(y,p) = L(y|p)g(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}(1-p)^{y-1}p = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha}(1-p)^{\beta+y-2}$$

and

$$m(y) = \int_0^1 f(y,p)\,\mathrm{d}p = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{(\alpha+1)-1}(1-p)^{(\beta+y-1)-1}\,\mathrm{d}p = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}B(\alpha+1,\beta+y-1).$$

Notice that we used the fact that the interval in the above line is the definition of the beta distribution with parameters $\alpha+1$ and $\beta+y-1$ (the notation B indicates the beta function). Therefore if we recall our earlier definition of the poster distribution,

$$g^*(p|y) = \frac{f(y,p)}{m(y)} = \frac{p^{\alpha}(1-p)^{\beta+y-2}}{B(\alpha+1,\beta+y-1)} = \frac{p^{(\alpha+1)-1}(1-p)^{(\beta+y-1)-1}}{B(\alpha+1,\beta+y-1)}.$$

Thus g^* has the beta distribution with parameters $\alpha+1$ and $\beta+y-1$.

4 Estimators

Once we have obtained a prior distribution of a parameter θ, we want to draw inferences about the actual value of the parameter. This can be accomplished by point estimates (often simply called estimators), or interval estimates (called credible intervals). We first discuss estimators. The estimator of θ, which we shall denoted $\hat{\theta}$, is given by

$$\hat{\theta} = E(\theta|y_1, y_2, \ldots, y_n)$$

where E is the expected value function. In integral form,

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta g^*(\theta|y_1, y_2, \ldots, y_n)\,\mathrm{d}\theta.$$

5 Credible Intervals

As stated above, an interval estimate of a parameter θ can also be obtained. In Bayesian statistics, we say (a,b) is a credible interval if

$$P(a \leq \theta \leq b) = \int_a^b g^*(\theta)\,\mathrm{d}\theta.$$

Furthermore, if $P(a \leq \theta \leq b) = 1-\alpha$ then (a,b) is called a $100(1-\alpha)\%$ credible interval for θ.

A credible interval is analogous to a confidence interval in some ways, but they are also different. In Frequentist statistics an interval $(\hat{\theta}_L, \hat{\theta}_R)$ is a $100(1-\alpha)\%$ confidence interval if

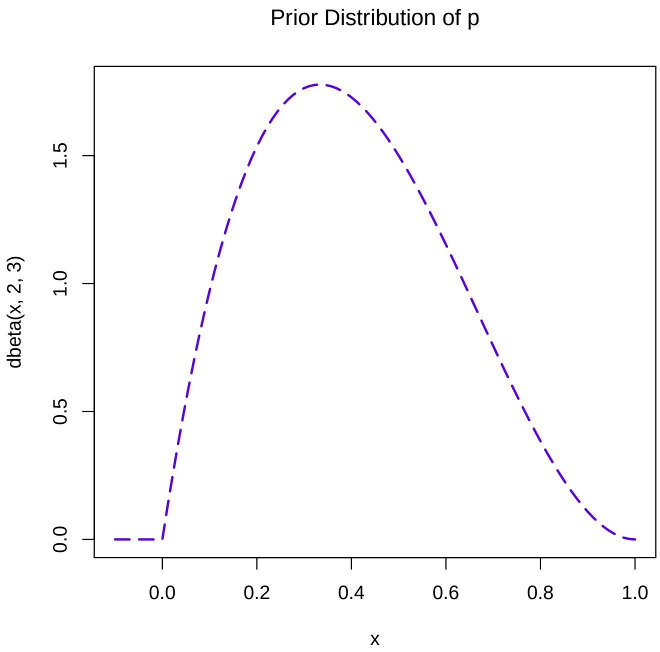
$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R) = 1 - \alpha.$$

However, $\hat{\theta}_L$ and $\hat{\theta}_R$ are random variables. The observed value of these random variables depend on the sample. So the endpoints of the confidence interval will change from sample to sample. If the confidence interval is say a 90% confidence interval then the observed confidence interval for a given sample will contain the true population parameter 90% of the time across different samples. It is easy to mistake the definition of a confidence interval as meaning that there is a $100(1-\alpha)\%$ chance that an observed confidence interval contains the true population mean, after all a definition such as this is more intuitive and even more useful. However, this definition is not the definition of a confidence interval but is precisely the definition of a credible interval. Due to the fact that the posterior distribution of the population parameter is known then we are actually able to state the probability that the parameter is in a fixed interval (the credible interval).

6 Determining Parameter Estimates of an Example

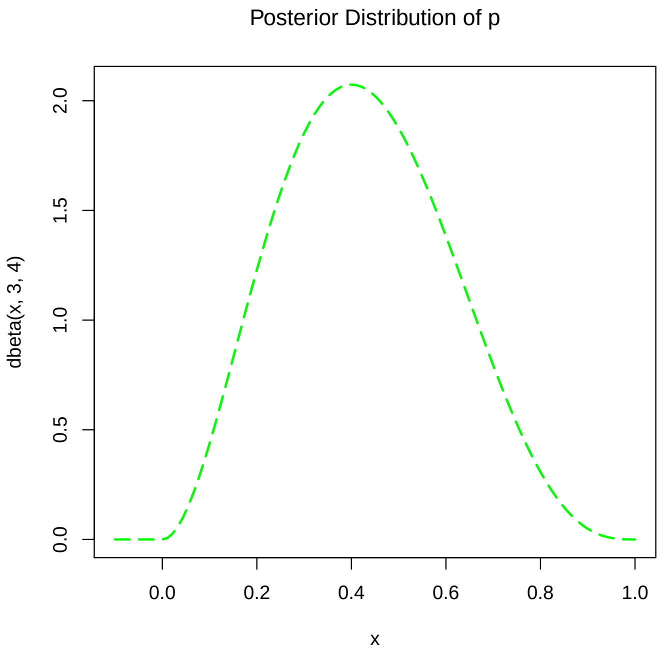
We now consider an application of the posterior distribution we obtained in our earlier example. Suppose that we are attempting to determine if a particular coin is weighted. Suppose we flip the coin until it comes up heads. Let Y be the random variable the corresponds to the first trial that yields a head. Thus Y has a geometric distribution with success parameter p as explained earlier. Suppose we conduct an experiment and observe that $Y = 2$. We can now use Bayesian Statistics to draw inferences about p. As stated above, we can choose a beta prior distrubution since $0 \leq p \leq 1$. We must now choose appropriate values for the parameters α and β of the beta prior distribution. Suppose that we have reason to consider that p is relatively near 0.4. We will choose α and β such that the mean of the prior distribution is 0.4 and the standard deviation is relatively wide. It can be shown that the mean of the beta distribution is given by $\frac{\alpha}{\alpha+\beta}$ and the standard deviation by $\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$. Thus if we set $\frac{\alpha}{\alpha+\beta} = 0.4$ then we obtain the relation $\alpha = \frac{2}{3}\beta$. Thus we may choose $\alpha = 2$ and $\beta = 3$. This results in a standard deviation of 0.2 which we deem to accurately represent our uncertainty of the actual value of p. We can plot our prior in R using the code:

```
curve(dbeta(x,2,3),-0.1,1,n=100,col="blue",lty=5,lwd=2,
+ main=expression("Prior Distribution of p"))
```



As shown earlier the posterior distribution is a beta distribution with parameters $\alpha^* = \alpha + 1$ and $\beta^* = \beta + y - 1$. Thus $\alpha^* = 3$ and $\beta^* = 4$. We now plot the posterior in R:

```
curve(dbeta(x,3,4),-0.1,1,n=100,col="green",lty=5,lwd=2,
+ main=expression("Posterior Distribution of p"))
```



Note that the peak of the posterior distribution appears to have shifted a little to the right compared to the prior. Intuitively this indicates that the actual value of p is a little greater than our prior assumption of being around 0.4. We now use estimators to solidify this hypothesis. The point estimator is given by:

$$\hat{p} = E(p|y) = \int_{-\infty}^{\infty} pg^*(p|y)\,\mathrm{d}p$$

However, this is simply the mean of the beta distribution (the formula for the mean was show earlier). So

$$\hat{p} = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{3}{3+4} \approx 0.429.$$

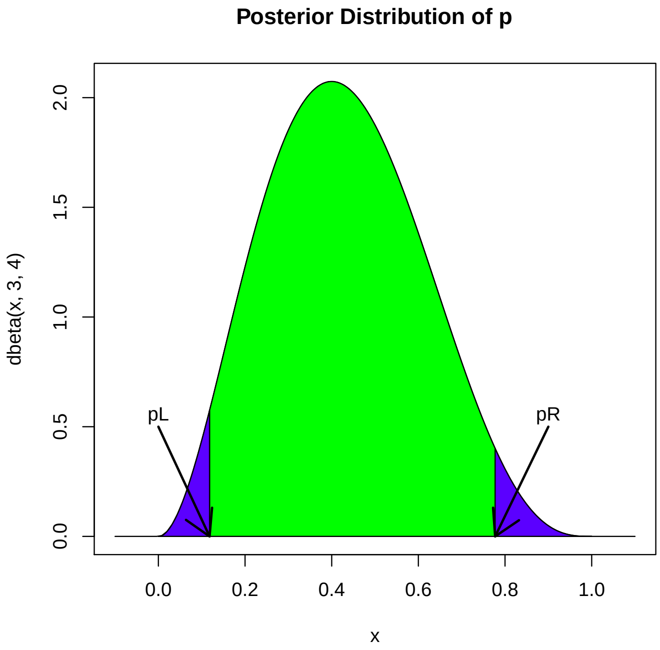
Furthermore, we can find an interval estimator of p. To determine a 95% credible interval, we must determine values p_L and p_R such that the posterior density curve has 0.025 area to the left of p_L and 0.025 area to the right (or equivalently 0.0975 area to the left) of p_R. This is a simple task using R. We use the code:

```
pL=qbeta(0.025,3,4)
pR=qbeta(0.975,3,4)
round(pL,3)
round(pR,3)
```

This yields that $p_L = 0.118$ and $p_R = 0.777$. So the 95% credible interval is given by

$$(p_L, p_R) = (0.118, 0.777).$$

This is certainly a wide interval, but it has given us more information than we had previously known about p. The R plot below shows the location of p_L and p_R in the posterior distribution. The area of the green region is 0.95 and the support of the distribution restricted to this region is the credible interval.



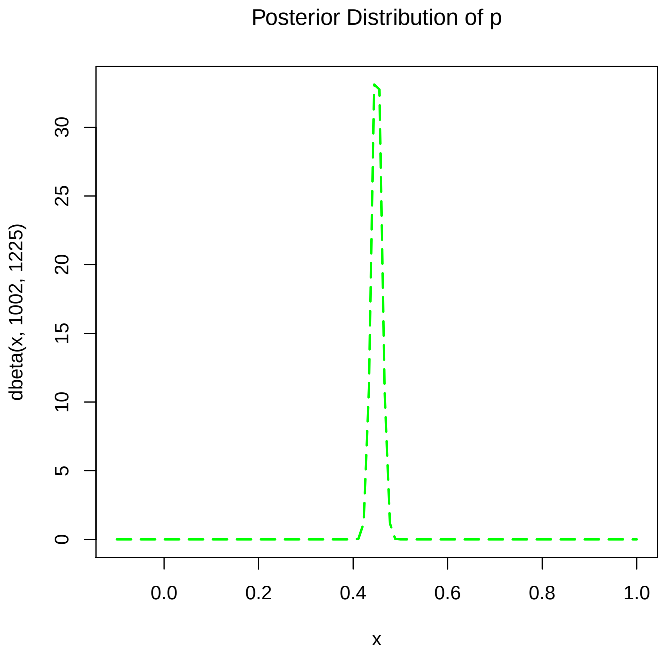
We can improve our estimates by taking a sample of larger size. We had previously only considered a sample of size 1 in our example involving the geometric distribution, but we can also consider a sample of size n. In this case since we assume $Y_1, \dots, Y_n$ are independent in our sample then

$$L(y_1, y_2, \dots, y_n|p) = L(y_1|p)L(y_2|p)\cdots L(y_n|p) = (1-p)^{y_1-1}p\cdots(1-p)^{y_n-1} = (1-p)^{\sum y_i - n}p^n.$$

It follows that

$$g^*(\theta|y_1, y_2, \dots, y_n) = \frac{L(y_1, y_2, \dots, y_n|\theta)g(\theta)}{\int_{-\infty}^{\infty} L(y_1, y_2, \dots, y_n|\theta)g(\theta)\mathrm{d}\theta} = \frac{p^{(\alpha+n)-1}(1-p)^{(\beta+\sum y_i - n)-1}}{B(\alpha+n, \beta+\sum y_i - n)}.$$

Thus the posterior has a beta distribution with parameters $\alpha^* = \alpha+n$ and $\beta^* = \beta+\sum y_i - n$. The posterior distribution will approach a spike at the true value of the population parameter upon increasing the sample size. For example, suppose that we take a sample of size 1000 and observe that $\sum_{i=1}^{1000} y_i = 2222$. Then the posterior distribution is a beta distribution with $\alpha^* = \alpha + n = 2 + 1000 = 1002$ and $\beta^* = \beta + \sum y_i - n = 3 + 2222 - 1000 = 1225$. The distribution is shown below.



Note that due to the increased sample size we now have narrowed in on the true value of the population parameter to a much larger extent. The point estimate of p is now given by

$$\hat{p} = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{1002}{1002 + 1225} \approx 0.45$$

and the 95% credible interval is given by $(0.429, 0.471)$ which is a much less wide than the credible interval obtained from the sample of size 1. Thus we have evidence that the coin is slightly weighted in our theoretical example.

7 Conclusion

We are now familiar with methodology of Bayesian inferences. We treat a population parameter as a random variable and construct a prior distrubution. Then we take a sample and use this sample along with the prior distribution to obtain a posterior distribution. Finally, we use this posterior distribution to draw inferences using point estimators or credible intervals. The biggest drawback to Bayesian inference seems to be the subjective nature of the prior distribution. The inferences that can be drawn are highly dependent on the prior distribution chosen so a proper prior should be carefully chosen. Otherwise two different people may obtain vastly different results if they do not agree on the prior distribution. However, the use of a prior distribution also has advantages. If a statistician has already has an idea regarding the nature of the population parameter then it would be advantageous to include this information in their analysis. Essentially we can eliminate values of the parameter that we know will not occur. Additionally, credible intervals in Bayesian inference seem to have more meaning than confidence intervals do in Frequentist inference as we explained eariler. Overall, there are advantages and disadvantages to drawing interences using both the Frequentist approach and the Bayesian approach. A choice of an approach to inference making should be chosen based on the context of the analysis.

References

[1] Wackerly, Dennis. *Mathematical Statistics with Applications 7th Edition.* California: Brooks/Cole, 2008. Print.