Inquiry into Bayesian Methods

Jared Murphy

Department of Mathematics and Statistics, University of North Florida, USA

Abstract

Bayesian methods of statistics are becoming increasingly relevant as the technological revolution unfolds. With ever increasing applications, a basic outline for Bayesian inference is much needed. The theoretical foundations for Bayesian methods are fairly straightforward, the core consideration being that a parameter is now considered to be a random rather than a fixed quantity. This simple shift from classical Frequentist methodology has a wide range of theoretical implications, allowing for a more direct and coherent interpretation of statistical results. The major concern with Bayesian methods is that they are too subjective and that establishing prior distributions at the experimenter's discretion will lead to unscientific results. However, there are ways to remain objective, as is explained in this paper.

1. Introduction

Throughout the majority of the 20th century, Frequentist methodology has been the major statistical approach used in data analysis. Under Frequentist methods of estimation, a parameter of interest, such as the population mean, is considered to be a fixed quantity that cannot necessarily be described probabilistically. In other words, when making an interval estimate of the population mean, one cannot explain the accuracy of the estimate in terms of probability because the parameter of interest is not considered to be a random quantity. Instead, the accuracy of the estimate is given as a level of confidence, meaning

that a certain percentage of the random intervals generated in repeated sampling will theoretically contain the parameter of interest. The confidence interval is one example of statistical theory that arose from the underlying assumptions of frequentist methodology, but what happens if some of those assumptions are challenged? For instance what theory of estimation would come about if a parameter of interest were considered to be a random instead of a fixed quantity? What if the data itself were to be considered fixed? Would this be a more accurate representation of natural systems, thus more useful in everyday life situations? Bayesian Statistics answers these questions.

Based off the concepts first presented by mathematician Thomas Bayes in 1763, Bayesian Statistics takes Bayes Theorem, a simple and well-known theory describing conditional probability, and expounds upon it, creating a complex and robust mathematical framework that can be used to answer a host of statistical questions (Bayesian-Inference, 2014). In the past, statisticians and mathematicians alike had shied away from this method. The level of analysis required for Bayesian methods is quite advanced and tedious when compared to that of the Frequentist approach. Because it was widely believed that the results of the two methods differed only slightly, many thought Bayesian methods were an impractical waste of time. Recently, however, interest surrounding Bayesian Statistics has reached an all time high, mostly due to the computing revolution. With the power of computers, the tedious and sometimes nearly impossible integration necessary to make Bayesian estimations can be carried out. With the rise of statistical computing, it is being shown that the differences in results produced by the two methods are much larger than previously thought, thus creating the debate, which method is best.

This paper seeks to discuss the foundational concepts of Bayesian Statistics. We will start from the ground up, first explaining the concepts of prior and posterior distributions for a parameter of interest. We will then use the posterior distribution, along with several other concepts, to make both point and interval parameter estimates. Hypothesis testing will then be discussed, followed by a more in depth look at prior distributions. By the end of this paper, the reader should have a clear understanding of the most fundamental concepts in Bayesian methods.

2.1 Posterior Distributions

Suppose that we have a random variable X whose distribution is dependent on a given value of θ, and that θ is an experimental value of random variable Θ whose distribution is defined by the experimenter as the prior distribution. We clarify with the notation,

$$X|\theta \sim f(x|\theta)$$

$$\Theta \sim h(\theta)$$

Now suppose that $X_1, X_2, \dots, X_n$ constitutes a random sample from the conditional distribution of X given $\Theta = \theta$. Vector notation is most convenient when dealing with these concepts. Let $\mathbf{X}' = (X_1, X_2, \dots, X_n)$ and $\mathbf{x}' = (x_1, x_2, \dots, x_n)$. Thus we can write the Likelihood function, which is the conditional pdf of $\mathbf{X}$ given $\Theta = \theta$, as

$$L(\boldsymbol{x}|\theta) = f(x_1|\theta)f(x_2|\theta) \dots f(x_n|\theta)$$

Thus the joint pdf of $\mathbf{X}$ and Θ is

$$g(\mathbf{x}, \theta) = L(\boldsymbol{x}|\theta)h(\theta)$$

If Θ is a continuous random variable, the joint marginal pdf of $\mathbf{X}$, is given by

$$g_1(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x}, \theta)\, d\theta = \int_{-\infty}^{\infty} L(\boldsymbol{x}|\theta)h(\theta)d\theta$$

If Θ is of the discrete type, just substitute summation for integration. Continuing with the continuous type, the conditional pdf of Θ given the random sample **X** is,

$$k(\theta|\mathbf{x}) = \frac{g(\mathbf{x}, \theta)}{g_1(\mathbf{x})} = \frac{L(\boldsymbol{x}|\theta)h(\theta)}{g_1(\mathbf{x})} = \frac{L(\boldsymbol{x}|\theta)h(\theta)}{\int_{-\infty}^{\infty} L(\boldsymbol{x}|\theta)h(\theta)d\theta}$$

$k(\theta|\mathbf{x})$ is called the posterior distribution. While the prior distribution $h(\theta)$ reflects the experimenter's subjective belief in the distribution of Θ *before* the random sample is drawn, the posterior distribution is the conditional distribution of Θ *after* the random sample has been drawn. For discrete cases, we replace integration with summation and density functions with mass functions.

Example 2.1.1 (obtained from Hogg, McKean, and Craig, 2013)

Consider the instance in which we are given

$$X|\theta \sim Poisson\ (\theta)$$

$$\Theta \sim \Gamma(\alpha, \beta)\ where\ \alpha, \beta\ are\ known$$

A random sample is then drawn from the Poisson distribution, resulting in a likelihood function,

$$L(\boldsymbol{x}|\theta) = \frac{\theta^{x_1}e^{-\theta}}{x_1!} \ldots \frac{\theta^{x_n}e^{-\theta}}{x_n!}\ , x_i = 0, 1, 2, \ldots and\ i = 1, 2, \ldots, n$$

with prior pdf

$$h(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}\ , \theta > 0.$$

Therefore the joint mixed discrete-continuous density function is

$$g(\mathbf{x}, \theta) = L(\boldsymbol{x}|\theta)h(\theta) = \frac{\theta^{n+\alpha-1}e^{-(n+1)\theta/\beta}}{x_1! \ldots x_n!\, \Gamma(\alpha)\beta^{\alpha}}\ , \qquad x_i = 0, 1, 2, \ldots and\ i = 1, 2, \ldots, n$$

$$\theta > 0$$

and the marginal distribution of the random sample is

$$g_1(\mathbf{x}) = \int_0^\infty g(\mathbf{x}, \theta)\, d\theta = \int_0^\infty L(\boldsymbol{x}|\theta)h(\theta)d\theta = \frac{\Gamma(\sum x_i + \alpha)}{x_1! \dots x_n!\, \Gamma(\alpha)\beta^\alpha (n + 1/\beta)^{\sum x_i + \alpha}}.$$

Thus, the posterior pdf of Θ is

$$k(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)h(\theta)}{g_1(\mathbf{x})} = \frac{1}{\Gamma(\sum x_i + \alpha)\left[\frac{\beta}{(n\beta + 1)}\right]^{\sum x_i + \alpha}} * \ \theta^{\sum x_i + \alpha - 1} e^{\frac{-\theta(n\beta+1)}{\beta}}$$

$$\theta > 0,\ x_i = 0, 1, 2, \dots and\ i = 1, 2, \dots, n$$

Hence,

$$k(\theta|\mathbf{x}) \sim \Gamma(\alpha^*, \beta^*)$$

$$where\ \alpha^* = \sum x_i + \alpha\ \ and\ \beta^* = \frac{\beta}{(n\beta + 1)}\ \blacksquare$$

The reason for separating $k(\theta|\mathbf{x})$ into two parts is to highlight that $\frac{1}{\Gamma(\sum x_i + \alpha)\left[\frac{\beta}{(n\beta+1)}\right]^{\sum x_i + \alpha}}$ is the "constant" that makes the posterior distribution a valid density function. In other words, we can simplify our answer to,

$$k(\theta|\mathbf{x}) \ \propto\ L(\mathbf{x}|\theta)h(\theta) = \theta^{\sum x_i + \alpha - 1} e^{\frac{-\theta(n\beta+1)}{\beta}}.$$

Before continuing, let us make a remark on sufficient statistics, and how this relates to posterior distributions. Suppose you have a sufficient statistic $Y = u(\mathbf{x})$ such that,

$$L(\mathbf{x}|\theta) = g[u(\mathbf{x})|\theta]H(\mathbf{x})$$

where $g(y|\theta)$ is the density function of *Y* given a value of Θ. It follows that,

$$k(\theta|\mathbf{x}) \ \propto\ g[u(\mathbf{x})|\theta]h(\theta)$$

Where $H(\mathbf{x})$ can be dropped because it can be considered part of the "constant" that that makes $k(\theta|\mathbf{x})$ a valid density function. Thus, if a sufficient statistic Y for the parameter θ exists, and $g(y|\theta)$ is known, we can obtain the posterior,

$$k(\theta|y) \propto g(y|\theta)h(\theta).$$

2.2 Point Estimation

Suppose we want to make an estimate of θ, a future value of the random variable Θ. To do this, we use a decision function $\delta(\mathbf{x})$ that minimizes the conditional expectation

$$E(\mathcal{L}[\theta, \delta(\mathbf{x})]|\mathbf{x}) = \int_{-\infty}^{\infty} \mathcal{L}[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

where $k(\theta|\mathbf{x})$ is the posterior distribution of Θ and $\mathcal{L}[\theta, \delta(\mathbf{x})]$ is a specified loss function. A loss function is a way to track the accuracy of our estimate. Choosing a decision function that minimizes the average loss ensures that our estimates are reliable, and the decision function that does this is called the Bayes Estimator.

One of the most common loss functions used in Bayesian estimation is the squared-error loss function defined as,

$$\mathcal{L}[\theta, \delta(\mathbf{x})] = [\theta - \delta(\mathbf{x})]^2$$

Because $E[(T-b)^2]$ is at a minimum when $b = E(T)$, where T is a random variable and b is a number, the Bayes Estimator for this loss function is $E(\theta|\mathbf{x})$.

The other loss function most commonly used is the absolute-error loss function defined as,

$$\mathcal{L}[\theta, \delta(\mathbf{x})] = |\theta - \delta(\mathbf{x})|$$

In this case the Bayes estimator is the median of the posterior distribution, because $E|T-b|$ is at a minimum when b is equal to the median of T.

Example 2.2.1 (obtained from Bayesian Estimation Theory, 2012)

A radioactive source emits n radioactive particles, and an imperfect Geiger counter records k ≤ n of them. Our problem is to estimate n from the measurements k. We assume that n is drawn from a Poisson distribution with known parameter λ:

$$P(n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad n \geq 0$$

The Poisson distribution characterizes the rate of emission of a process in a given interval of time (or space).

The number of recorded counts follows a binomial distribution, i.e. the number of successful recordings k for every n radioactive particles emitted. Assume that p is known. The distribution is described as,

$$p(k|n) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n$$

To proceed with the Bayesian analysis we must find the posterior distribution of the random variable n given k. It is simple to find the joint distribution of n and k to be,

$$P(n,k) = P(n)P(k|n) = e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n$$

And with some algebraic manipulation we find the marginal distribution of k,

$$P(k) = \sum_{n-k=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!}\binom{n}{k} p^k (1-p)^{n-k} = \sum_{n-k=0}^{\infty} e^{-\lambda}\frac{\lambda^{n-k}}{k!\,(n-k)!}(\lambda p)^k (1-p)^{n-k}$$

$$= \frac{e^{-\lambda}(\lambda p)^k}{k!}\sum_{n-k=0}^{\infty}\frac{(\lambda(1-p)^{n-k}}{(n-k)!} = e^{-\lambda p}\frac{(\lambda p)^k}{k!}, \quad k \geq 0.$$

Therefore the posterior pmf of n given k is,

$$P(n|k) = \frac{P(n,k)}{P(k)} = \frac{P(n)P(k|n)}{P(k)} = \frac{1}{(n-k)!}(\lambda(1-p))^{n-k}e^{-\lambda(1-p)}, \;\; n \geq k$$

Which is a Poisson distribution with displacement k. Seeking to make a point estimate of a future experimental value of n given value k we use a squared-error loss function $\mathcal{L}[n, \delta(\mathrm{k})] = [n - \delta(\mathrm{k})]^2$ which is minimized, on average, when $\delta(\mathrm{k}) = E(n|\mathrm{k})$. The conditional expectation of n given k will be our Bayes Estimator. We obtain this indirectly using the method of moment-generating functions. Let $z = \lambda(1-p)$

$$M_{n|k}(t) = E(e^{nt}|k) = \sum_{n=k}^{\infty} \frac{e^{nt}}{(n-k)!} z^{n-k}e^{-z} = z^{-k}e^{-z}\sum_{n=k}^{\infty}\frac{(ze^t)^n}{(n-k)!}$$

$$= z^{-k}e^{-z}(ze^t)^k \sum_{n=k}^{\infty}\frac{(ze^t)^{n-k}}{(n-k)!} = e^{-z}e^{tk}e^{e^t z} = e^{z(e^t-1)+tk}$$

Therefore,

$$\frac{dM_{n|k}(t)}{dt}|_{t=0} = \; E(n|\mathrm{k}) = z + k = \lambda(1-p) + k \; \blacksquare$$

2.3 Interval Estimation

When compared with the Frequentist approach, Bayesian interval estimation is straightforward. Instead of constructing an abstract notion of "confidence intervals", we can make probabilistic statements about the parameter directly because it is considered to be a random variable. Hence,

$$P[u(\mathbf{x}) \leq \Theta \leq v(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \int_{u(\mathbf{x})}^{v(\mathbf{x})} k(\theta|\mathbf{x})d\theta.$$

Instead of a confidence interval this is known as a credible or probability interval.

Example 2.3.1 (obtained from Hogg, McKean, and Craig, 2013 and D'Agostini, 2003)

Suppose we wish to make interval estimates on an unknown mean of a normal population. We obtain a random sample of size n, each observation distributed as,

$$X_i|\mu \sim N(\mu, \sigma^2) \; where \; \sigma \; is \; known.$$

And then establish a prior distribution for μ to be

$$\mu \sim N(\mu_0, {\sigma_0}^2) \; where \; \mu_0, {\sigma_0}^2 \; are \; known.$$

Using the sufficient statistic $Y = \bar{X}$, we can use its distribution

$$Y|\mu \sim N(\mu, \frac{\sigma^2}{n})$$

to find the posterior $k(\theta|y)$. It follows that,

$$k(\mu|y) \; \propto \; g(y|\mu)h(\mu) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma_0} exp\left[-\frac{(y-\mu)^2}{2(\sigma^2/n)} - \frac{(\mu-\mu_0)^2}{2{\sigma_0}^2}\right]$$

After eliminating all constant factors, including those consisting solely of y, and completing the square, we obtain,

$$k(\mu|y) \; \propto exp\left[-\frac{\left[\mu - \frac{y{\sigma_0}^2 + \mu_0(\sigma^2/n)}{{\sigma_0}^2 + (\sigma^2/n)}\right]^2}{\frac{2(\sigma^2/n){\sigma_0}^2}{{\sigma_0}^2 + (\sigma^2/n)}}\right]$$

Which implies that $k(\mu|y)$ is a normal distribution with mean $\frac{y{\sigma_0}^2 + \mu_0\left(\frac{\sigma^2}{n}\right)}{{\sigma_0}^2 + \left(\frac{\sigma^2}{n}\right)}$ and variance $\frac{\left(\frac{\sigma^2}{n}\right){\sigma_0}^2}{{\sigma_0}^2 + \left(\frac{\sigma^2}{n}\right)}$.

Because the posterior distribution of μ given $Y = y$ is normal, interval estimation of μ is straightforward. Let $\mu_1 = \frac{y{\sigma_0}^2 + \mu_0\left(\frac{\sigma^2}{n}\right)}{{\sigma_0}^2 + \left(\frac{\sigma^2}{n}\right)}$ and $\sigma_1 = \sqrt{\frac{\left(\frac{\sigma^2}{n}\right){\sigma_0}^2}{{\sigma_0}^2 + \left(\frac{\sigma^2}{n}\right)}}$. Under the empirical rule of normal probabilities it follows that,

$$68.3\% \; credible \; interval: \mu_1 \pm \sigma_1$$

$$90\% \; credible \; interval: \mu_1 \pm 1.65\sigma_1$$

$$95\%\ credible\ interval:\ \mu_1 \pm 1.96\sigma_1$$

$$99\%\ credible\ interval: \mu_1 \pm 2.58\sigma_1$$

$$99.73\%\ credible\ interval: \mu_1 \pm 3\sigma_1 \blacksquare$$

As it turns out, example 2.3.1 can be generalized for any posterior obtained from a normal prior and likelihood, or conditional distribution of a sufficient statistic. Therefore if

$$X|\mu \sim N(\mu, \sigma^2)\ where\ \sigma^2\ is\ known$$

$$\mu \sim N(\mu_0, {\sigma_0}^2)\ where\ \mu_0, {\sigma_0}^2\ are\ known$$

Then,

$$\mu|x \sim N(\mu_1, {\sigma_1}^2)\ where$$

$$\mu_1 = \frac{x{\sigma_0}^2 + \mu_0\sigma^2}{{\sigma_0}^2 + \sigma^2} \quad and \quad {\sigma_1}^2 = \frac{\sigma^2{\sigma_0}^2}{{\sigma_0}^2 + \sigma^2}.$$

2.4 Hypothesis Testing

Much like the previous sections, the Bayesian method for testing a hypothesis is primarily a matter of finding the posterior distribution. Suppose we wish to test,

$$H_0: \theta \in \omega_0\ \ vs.\ H_a: \theta \in \omega_1$$

where $\theta \in \Omega,\ \omega_0 \cup \omega_1 = \Omega$ and $\omega_0 \cap \omega_1 = \emptyset$. Assuming we have a likelihood function $L(\mathbf{x}|\theta)$ and a prior $h(\theta)$, we find the posterior distribution $k(\theta|\mathbf{x})$ to obtain,

$$P(\Theta \in \omega_0|\mathbf{x})\ and\ P(\Theta \in \omega_1|\mathbf{x}).$$

These probabilities represent the respective truth-values of H_0 and H_a.

Several rejection rules can be used for making a decision, the simplest of which is to,

$$Reject\ H_0\ if\ P(\Theta \in \omega_1|\mathbf{x}) \geq P(\Theta \in \omega_0|\mathbf{x}),$$

but in most instances this will be too loose a criterion. Another option for evaluation is to use what is outlined in *Bayesian Statistics: An Introduction* (2004) as the Bayes Factor, which is

$$B = \frac{\frac{P(\Theta \in \omega_0|\mathbf{x})}{P(\Theta \in \omega_1|\mathbf{x})}}{\frac{P(\Theta \in \omega_0)}{P(\Theta \in \omega_1)}} = \frac{\frac{P(\Theta \in \omega_0|\mathbf{x})}{1 - P(\Theta \in \omega_0|\mathbf{x})}}{\frac{P(\Theta \in \omega_0)}{1 - P(\Theta \in \omega_0)}} = \frac{posterior\ odds\ for\ H_0}{prior\ odds\ for\ H_0}.$$

This odds ratio reflects how much the data collected changes our beliefs about the null hypothesis. When the ratio is close to one, the data has not revealed any significant reason for null rejection.

If one were looking for something analogous to the Frequentist approach, a probability limit α could be established. So for instance, if we tested

$$H_0: \theta \geq \theta_0 \ \ vs.\ H_a: \theta < \theta_0$$

with a specified α, we could reject the null if

$$Reject\ H_0\ if\ P(\theta < \theta_0|\mathbf{x}) \geq 1 - \alpha$$

or equivalently,

$$Reject\ H_0\ if\ \int_{-\infty}^{\theta_0} k\,(\theta|\mathbf{x})d\theta \geq 1 - \alpha.$$

It is important to not mistake α for the Type I error in Frequentist theory. Because we are now treating θ as a random quantity, there is no instance when the null can be assumed "true". Therefore, the statement $\alpha = P(Reject\ H_0\ when\ H_0\ is\ true)$ does not make any sense in this context.

Example 2.4.1 (obtained from Hogg, McKean, and Craig, 2013)

Referring back to example 2.1.1, in which a random sample was drawn from a Poisson distribution with unknown parameter θ, suppose we are interested in conducting some hypothesis tests of the form,

$$H_0: \theta \leq 10 \ \ vs.\ H_a: \theta > 10$$

where we reject the null if $P(\theta \leq 10|\mathbf{x}) \leq .05$. Further, suppose that we think θ is around 12. Hence, we establish our prior distribution to be $\Gamma(10,1.2)$ which has a mean 12 and variance of 14.4. After drawing a random sample of $n = 20$ from the Poisson distribution, we find that

$$\sum_{i=1}^{\mathbf{20}} x_i = 177.$$

Using the posterior distribution calculated in example 2.1.1, we obtain

$$k(\theta|\mathbf{x}) \sim \Gamma(\alpha^*, \beta^*) = \Gamma(177 + 10, [20(1.2) + 1]) = \Gamma(187,0.048)$$

Which has a mean 8.976 and a variance 0.431. Using a statistical computing package, the probability of the null is found to be

$$P(\theta \leq 10|\mathbf{x}) = P[\Gamma(187,0.048) \leq 10] = 0.9368$$

Therefore, according to our rejection criterion, we cannot reject H_0.■

3. Prior Distributions

One of the main points of contention in Bayesian theory is the use of prior distributions and the methods by which they are established. The main concern is "…that as scientists we should be concerned with objective knowledge rather than subjective belief…" and that, no matter what the case, there is no clear way to effectively evaluate subjective beliefs (Gelman, 2008). But, on the other hand, it can also be argued that "as we collect more data, the influence of the prior generally goes to zero, provided the number of parameters does not grow with the sample size" (Bayesian Methodology in Biostatistics). We now investigate prior distributions more intensively as to give the reader an opportunity to settle the issue for themselves.

3.1 Conjugate Priors

In general, "a prior is conjugate for a family of distributions if the prior and the posterior are of the same family" (Bayesian Methodology in Biostatistics). In virtually every example presented in this paper, the prior has indeed been conjugate. A gamma prior combined with Poisson likelihood produced a gamma posterior. A normal prior combined with normal likelihood produced a normal posterior. The following table outlines several pairs of conjugate priors and their respective families of distributions.

Figure 3.1.1 (obtained from Bayesian Methodology in Biostatistics)

Family	*Conjugate Prior for θ*
Binomial (n, θ)	$\theta \sim$ *Beta (α,β)*
Poisson (θ)	$\theta \sim$ *Gamma (α,β)*
Normal (θ, σ^2); σ^2 known	$\theta \sim$ *Normal (θ_0, σ_0^2)*
Gamma (α, θ); α known	$\theta \sim$ *Gamma (δ_0, γ_0)*
Normal (α, θ); α known	$\theta^{-1} \sim$ *Gamma (δ_0, γ_0)*

3.2 Proper vs. Improper Priors

All the examples of prior distributions given up to this point have been proper, i.e. for a density $h(\theta)$

$$\int_{-\infty}^{\infty} h(\theta)d\theta = k, \;\; k \in \mathbb{R}^+.$$

To be more specific, all prior densities presented have integrated to unity, making each example of $h(\theta)$ that was given a pdf in and of itself. However, this does not have to be the case. A prior is said to be improper if

$$h(\theta) \geq 0, \quad \int_{-\infty}^{\infty} h(\theta)d\theta = \infty$$

and, given a likelihood function $L(\mathbf{x}|\theta)$, the posterior $k(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)h(\theta)$ is positive and integrates to a positive constant, thus allowing it to be normalized into a pdf.

Example 3.2.1 (obtained from Bayesian Methodology in Biostatistics)

Suppose we have a random sample from a normal population

$$x_1, x_2, \ldots, \sim N(\theta, 1)$$

and establish an improper prior

$$h(\theta) = 1 \quad -\infty < \theta < \infty$$

It follows that the posterior distribution is

$$k(\theta|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}\times 1 = \exp\left\{-\frac{1}{2}\sum_{i}^{n}(x_i-\theta)^2\right\}$$

$$= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left\{n\theta^2 - 2\theta\sum_{i=1}^{n}x_i\right\}\right\}$$

$$= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}[\theta-\bar{x}]^2\right\}.$$

Therefore,

$$\theta|\mathbf{x} \sim N\left(\bar{x}, \frac{1}{n}\right) \blacksquare$$

3.3 Uninformative Priors

When dealing with conjugate priors the mathematics involved can produce very nice and clean results, but if one is seeking to remain as objective as possible, a prior that minimally influences the posterior may be useful. This type of prior is called uninformative and is most often improper. Simply put, an uninformative prior "is a prior which treats all values of θ the same, that is, uniformly" (Hogg, McKean, and Craig, 2013). These types of

priors are most commonly used when there is no prior knowledge of the parameter on the part of the experimenter.

Example 3.3.1 (Obtained from Hogg, McKean, and Craig, 2013)

Suppose we have a normal distribution $N(\theta_1, \theta_2)$ in which both θ_1 and $\theta_2 > 0$ are unknown. Let us define an uninformative and improper prior for θ_1 to be $h_1(\theta_1) = 1$ where $-\infty < \theta_1 < \infty$. An improper prior for θ_2 is $h_2(\theta_2) = c_2/\theta_2$, where $0 < \theta_2 < \infty$ and c_2 is a constant. Note that the simple transformation $ln\theta_2$ is uniformly distributed between $-\infty < ln\theta_2 < \infty$, i.e. if

$$Y = ln\theta_2, \quad \theta_2 = e^Y, \quad and \ \frac{d\theta_2}{dy} = e^Y$$

then,

$$f_Y(y) = f_{\theta_2}[g^{-1}(y)] \times \frac{d\theta_2}{dy} = \frac{c_2}{e^y} * e^y = c_2, -\infty < ln\theta_2 < \infty,$$

Hence in this way $h_2(\theta_2)$ is an uninformative prior. Also, let us assume the parameters are independent. Then the joint improper uninformative prior, is

$$h_1(\theta_1)h_2(\theta_2) \propto \frac{1}{\theta_2}, \quad -\infty < \theta_1 < \infty \ and \ 0 < \theta_2 < \infty.$$

Now let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution initially described. Recall that $\bar{X}$ and $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are sufficient statistics. Finally, suppose we use the prior $h_1(\theta_1)h_2(\theta_2)$. It follows that the posterior distribution is given by,

$$k_{12}(\theta_1, \theta_2|\bar{x}, s^2) \propto \left(\frac{1}{\theta_2}\right)\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)\exp\left[-\frac{1}{2}\{(n-1)s^2 + n(\bar{x} - \theta_1)^2\}/\theta_2\right]$$

$$\propto \left(\frac{1}{\theta_2}\right)^{\frac{n}{2}+1}\exp\left[-\frac{1}{2}\{(n-1)s^2 + n(\bar{x} - \theta_1)^2\}/\theta_2\right].$$

To get the conditional pdf of $\theta_1|\bar{x}, s^2$, we integrate out θ_2

$$k_1(\theta_1|\bar{x}, s^2) = \int_0^\infty k_{12}(\theta_1, \theta_2|\bar{x}, s^2)\, d\theta_2.$$

To carry this out, let us change variables $z = 1/\theta_2$ and $\theta_2 = 1/z$, with Jacobian $\frac{d\theta_2}{dz} = -1/z^2$. Thus,

$$k_1(\theta_1|\bar{x}, s^2) \propto \int_0^\infty \frac{z^{\frac{n}{2}+1}}{z^2} \exp\left[-\frac{z}{2}\{(n-1)s^2 + n(\bar{x}-\theta_1)^2\}\right] dz.$$

Referring to the gamma distribution with $\alpha = n/2$ and $\beta = 2/\{(n-1)s^2 + n(\bar{x}-\theta_1)^2\}$, this result is proportional to

$$k_1(\theta_1|\bar{x}, s^2) \propto \{(n-1)s^2 + n(\bar{x}-\theta_1)^2\}^{-n/2}.$$

Let us change the variables to get more familiar results; namely, let

$$t = \frac{\theta_1 - \bar{x}}{s/\sqrt{n}} \ and\ \theta_1 = \bar{x} + ts/\sqrt{n},$$

with $\frac{d\theta_1}{dt} = s/\sqrt{n}$. This conditional pdf of t, given $\bar{x}$ and s^2, is then

$$k(t|\bar{x}, s^2) \propto \{(n-1)s^2 + (st)^2\}^{-n/2}$$

$$\propto \frac{1}{[1 + t^2/(n-1)]^{[(n-1)+1]/2}}.$$

That is, the conditional pdf of t given $\bar{x}$ and s^2 is a student t distribution with $n-1$ degrees of freedom. ■

3.4 Jeffreys Priors

When seeking a truly uninformative and objective prior one may run into problems i.e. the prior distribution may be uninformative when established, but may not be for another parameterization.

Example 3.4.1 (obtained from Bayesian Modeling and Inference, 2010)

Consider a binomial distribution $X \sim Binom(n, \theta)$ in which we put a prior on θ. We know that θ lies between 0 and 1. A uniform uninformative prior for θ would be $h(\theta) = 1$. Since θ lies between 0 and 1, we can use a new parameterization using the log-odds ratio: $\rho = \log\frac{\theta}{1-\theta}$. This is a perfectly valid parameterization, but it is no longer uninformative. ■

To establish an uninformative prior that is invariant, we look to a method based on the Fisher information function of a model. This method was originally devised by mathematician Herold Jeffreys.

Example 3.4.2 (obtained from Statistics, 2011)

Consider a model $X \sim f(X|\theta)$ where $logf(X|\theta)$ is twice differentiable in θ for every x. The Fisher information of the model at any θ is defined to be

$$I^F(\theta) = E_{X|\theta}\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2 = E_{X|\theta}\{l'_x(\theta)\}^2.$$

Which, under some regularity conditions, equals

$$I^F(\theta) = -E_{X|\theta}\frac{\partial^2}{\partial\theta^2}\log f(X|\theta) = -E_{X|\theta}l''_x(\theta).$$

If data has n components $(X_1, X_2, \dots, X_n)$ and the model is iid $X_i \sim g(x_i|\theta)$, then $f(X|\theta) = \prod_{i=1}^n g(x_i|\theta)$ and so

$$I^F(\theta) = \sum_{i=1}^{n}\left[-E_{X|\theta}\frac{\partial^2}{\partial\theta^2}\log g(x_i|\theta)\right] = nI_1^F(\theta)$$

where $I_1^F(\theta)$ is the single observation Fisher information of $X_i \sim g(x_i|\theta)$. The Jeffreys proposal for a non-informative prior pdf for the model $X \sim f(X|\theta)$ is

$$h(\theta) \propto \sqrt{I^F(\theta)}. \blacksquare$$

Example 3.4.3 (obtained from Statistics 2011)

Consider a random sample from a normal population with an unknown mean and a known variance. The Fisher information function of a single observation for μ is

$$I_1^F(\mu) = -E_{X_1|\mu}\frac{\partial^2}{\partial\mu^2}\frac{(X_1-\mu)^2}{2\sigma^2} = \frac{1}{\sigma^2}$$

and hence

$$I^F(\theta) = nI_1^F(\mu) = \frac{n}{\sigma^2}.$$

Therefore the Jeffreys prior for μ is

$$h(\mu) \propto \sqrt{\frac{n}{\sigma^2}} = constant, \quad -\infty < \mu < \infty. \blacksquare$$

4. Conclusion

This paper has outlined the most fundamental aspects of Bayesian Statistical Methods, and has established a foundation for further investigation. With a simple shift of definition at the most basic level, (i.e. considering the parameter a random variable rather than a fixed value), the effects are far reaching, allowing for a more reasonable explanation of parameter estimates in terms of probability rather than confidence.

There are, however, objections to the Bayesian point of view, primarily surrounding the issue of subjectivity. Clearly there is some danger when any level of subjectivity is allowed into the arena of scientific analysis. Given that there is already such widespread misinterpretation of statistical results, adding a subjective element directly into a theories mathematical framework may lead to much larger problems then we have now. However, it was shown earlier that there are Bayesian methods for establishing objective uninformative prior distributions, and that the subjective element need not always be present. When using Bayesian methods, subjectivity should be used in the right way for the right reasons, and someone who knows what they are doing should always do the analysis.

Hopefully this paper has given the reader much needed insight into a whole realm of statistical theory that is not presented in the standard higher education curriculum, thus allowing them to see that there is an alternative to the Frequentist approach. It is important that the statisticians of tomorrow are equipped with the tools necessary to extrapolate meaning from a data set in increasingly effective ways. The Bayesian method is a useful tool for statisticians, and is should not be ignored.

References

Bayesian Methodology in Biostatistics. Retrieved from http://isites.harvard.edu/fs/docs/icb.topic353981.files/unit2.pdf

Dagostini, (2003). *Bayesian Reasoning in Data Analysis: A Critical Introduction.* River Edge, NJ: World Scientific

Gelman, Andrew (2008). Objections to Bayesian Statistics. Retrieved from http://www.stat.columbia.edu/~gelman/research/published/badbayesmain.pdf

Hogg, McKean, and Craig, (2013). *Introduction to Mathematical Statistics.* Boston, MA: Pearson.

Lecture 13 and 14: Bayesian estimation theory (2012). Retrieved from http://www.eecs.tufts.edu/~khan/Courses/Spring2012/EE194/Lecs/Lec13_14.pdf

Lee, (2004). *Bayesian Statistics: An Introduction.* London, England: Hodder Arnold.

Reverend Thomas Bayes (2014). Retrieved from http://www.bayesian-inference.com/bayes.

STA 114: Statistics (2011). Retrieved from https://stat.duke.edu/courses/Fall11/sta114/jeffreys.pdf

Stat 260: Bayesian Modeling and Inference (2010). Retriaved from http://www.cs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture6.pdf