Stat 111 Spring 2011 Week 12 Problems - Objective Bayesian Inference

Bayesian inference is useful if you want to incorporate prior information into your analysis. But the most important argument for the Bayesian approach is that it provides a coherent strategy for making inference about any parametric problem: Write down a probability model for the data, specify prior distributions for the unknown parameters, then make inference based on the conditional (posterior) distributions of the parameters, given the data.

A common complaint about Bayesian inference is that different prior assumptions lead to different conclusions. *Objective Bayes* inference involves finding prior distributions that yield good frequentist properties (e.g., 90% interval estimates that cover the true parameter for 90% of data sets).

1. The average height for the 16 women who answered my Stat 1 survey this Spring is $\bar{y} = 65.0$ inches. Assume that this represents a simple random sample of all Swarthmore women, and that heights vary according to a Normal distribution. You wish to make inference about μ, the true mean height of all Swarthmore women. For now, assume we know that the standard deviation of Swarthmore women's heights is 3.0 inches.

   a) Say you're "certain" that the mean height is between 60 and 70 inches, but you don't want to say any more than that. You could specify a Uniform prior density between 60 and 70 inches. Derive the posterior density for $f(\mu|\bar{y})$ and find a 90% "credible interval" for μ. Explain precisely what this interval represents.

   b) For Bayesian inference, we may specify an improper prior density (i.e., a density that does not have a finite integral) as long as the corresponding posterior density is proper (i.e., has a finite integral). Show that a flat prior density on $-\infty$ to ∞ (i.e., $p(\mu) \propto c$) yields a proper posterior density in this problem, and compare it to the posterior density from part a. Improper prior densities do not present a paradox; rather, they are a mathematical convenience.

   c) Previous surveys of Swarthmore students might lead me to specify a Normal prior density with mean $\mu_o = 64.5$ and standard deviation $\tau = 1$ (I could be more precise than that, but I don't want to be too subjective). Work out the posterior density for the general problem of estimating the mean μ of a Normal distribution, when assuming the prior distribution $p_\mu(\mu)$ that is $N(\mu_o, \tau^2)$.

2. The goal in Objective Bayesian inference is to specify a prior density that is "non-informative." Uniform prior distributions make all values of a parameter equally likely, but that doesn't necessarily equate to not injecting any prior information into the problem. If you specify a Uniform prior density for a parameter θ, then the implied prior density for any non-linear function $\phi = g(\theta)$ is not Uniform, so some values of ϕ are *a priori* more probable than others.

a) What do we mean by a density function and a change of variables? Consider a random variable θ with density p_θ, and a transformation $\phi = g(\theta)$ with density p_ϕ. Find the relationship between these two densities by equating $P(\theta \in [\theta, \theta + d\theta)) = p_\theta(\theta)d\theta$, and $P(\phi \in [\phi, \phi + d\phi)) = p_\phi(\phi)d\phi$. This will involve finding a first order Taylor approximation to the function $g(\theta + d\theta)$ centered at θ. For example, let $\phi = \log(\theta)$.

b) The **Jeffreys' prior density** for a parameter θ is the density implied by specifying a Uniform density on a particular transformation $\phi = g(\theta)$. The transformation $\phi = g(\theta)$ is chosen such that the Fisher information for ϕ is the same for all data sets of the same size, regardless of the value of ϕ. This is a scale where it makes sense to say "non-informative" means "all values of ϕ are equally likely." Show this is the case for $\phi = \log(\sigma^2)$ for a sample from a $N(0, \sigma^2)$ distribution. Using the same axis, graph the log-likelihood function for σ^2 based on a sample of $n = 10$ with MLE $\hat{\sigma}^2 = 16$, and for a sample with $\hat{\sigma}^2 = 9$. Subtract off the maximum log-likelihood value so that both graphs have maximum value 0. Repeat this for $\phi = \log(\sigma^2)$. Notice that the curvature differs for the two σ^2 log-likelihoods, but not for the two ϕ log-likelihoods.

c) A Uniform prior density for $\log(\sigma^2)$ implies a density $p_{\sigma^2}(\sigma^2)d\sigma^2 \propto d\sigma^2/\sigma^2$. A prior density that satisfies the requirements we applied here (specifying a Uniform prior density on a transformation of θ that has constant information) is called "Jeffreys' prior" (Lord Jeffreys once wrote: "It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.") The problem simplifies to setting $p(\theta)$ proportional to the square root of the Fisher Information:

$$p(\theta) \propto \left(-E\left(\frac{\partial^2}{\partial^2\theta} \log(L(\theta)) \right) \right)^{1/2}.$$

Find the Jeffreys' prior densities for the mean and variance of a Normal distribution.

d) Multiply the prior densities from c) gives a non-informative joint prior distribution for μ and σ^2. We have seen that this prior implies a $t_{(n-1)}$ marginal posterior density for $(\mu - \bar{y})/(s/\sqrt{n})$. It is often not possible to get a closed-form solution like this. An alternative way to characterize a posterior density f is to simulate a large number of random draws from f. Outline a 2-step algorithm for simulating from the joint posterior distribution of μ and σ^2. Integrate μ out of the joint posterior distribution and recognize the resulting marginal posterior for $\sigma^2|y$, and the conditional posterior for $\mu|y, \sigma^2$. Explain how making sequential draws from these two distributions gives draws from the joint posterior, and that the sample of μ values represents draws from the marginal posterior density for $\mu|y$.

3. Consider the 2-level Normal hierarchical model with constant and known level-1 variance V:

$$\begin{aligned} Y_i \mid \theta_i &\overset{\text{indep}}{\sim} N(\theta_i, V), \quad i = 1, \ldots, k. \\ \theta_i \mid \mu, A &\overset{\text{i.i.d.}}{\sim} N(\mu, A). \end{aligned}$$

For example, each Y_i might represent the average point total for a particular team in their first n games. Assuming the same standard deviation for all teams, we have $V = \sigma/n$, with σ estimated based on $k(n-1)$ degrees of freedom. If this value is large, then V is essentially known.

a) It is well-established that a Uniform density on $(-\infty, \infty)$ is a reasonable non-informative prior density for the mean μ of a Normal random variable. Treating μ as known (for now) derive the Jeffreys prior density for $B = V/(V + A)$.

b) If an independent Uniform prior distribution is assumed for μ, then your answer to (a) is the implicit joint-prior density for μ and B. Making this prior specification, find the marginal posterior density $f(B|Y)$ and show that the posterior mode for B (the value of B that maximizes its marginal posterior density) is the Stein estimate: $\hat{B} = \min\{1,\ (k-3)V/\sum((y_i - \bar{y})^2)\}$.

c) Find the conditional posterior density for $\mu \mid y, B$ and for $\theta \mid y, B$. Use these to find the modified posterior standard deviation estimates for the θ_i's described in the article *Stein's Paradox Revisited.*