LECTURE 9

14.2. Example: Birth rates.

Over the course of the 1990s the General Social Survey gathered data on the educational attainment and number of children of 155 women who were 40 years of age at the time of their participation in the survey. These women were in their 20s during the 1970s, a period of historically low fertility rates in the United States. In this example we will compare the women with college degrees to those without in terms of their numbers of children. Let $Y_{1,1}, \ldots, Y_{n_1,1}$ denote the numbers of children for the n_1 women without college degrees and $Y_{1,2}, \ldots, Y_{n_2,2}$ be the data for women with degrees. For this example, we will use the following sampling models:

$$Y_{1,1}, \ldots, Y_{n_1,1}|\theta_1 \sim i.i.d.Poisson(\theta_1),$$

$$Y_{1,2}, \ldots, Y_{n_2,2}|\theta_2 \sim i.i.d.Poisson(\theta_2),$$

and group sums and means for the data are as follows:

Less than bachelors: $n_1 = 111$, $\sum_{i=1}^{n_1} Y_{i,1} = 217$, $\overline{Y_1} = 1.95$

Bachelors or higher: $n_2 = 44$, $\sum_{i=1}^{n_2} Y_{i,2} = 66$, $\overline{Y_2} = 1.50$.

In the case where $\{\theta_1, \theta_2\} \sim$ i.i.d. $gamma(a = 2, b = 1)$, we have the following posterior distributions:

$$\theta_1|\{n_1 = 111, \sum Y_{i,1} = 217\} \sim gamma(2 + 217, 1 + 111) = gamma(219, 112)$$

$$\theta_2|\{n_2 = 44, \sum Y_{i,2} = 66\} \sim gamma(2 + 66, 1 + 44) = gamma(68, 45)$$

Posterior means and modes for θ_1 and θ_2 can be obtained from their gamma posterior distributions.

14.3 Exponential families and conjugate priors.

The binomial and Poisson models are both instances of one-parameter exponential family models. A one-parameter exponential family model is any model whose densities can be expressed as $p(y|\varphi) = h(y)c(\varphi)e^{\varphi t(y)}$, where φ is the unknown parameter and $t(y)$ is the sufficient statistic.

Let us study conjugate prior distributions for general exponential family models, and in particular prior distributions of the form $p(\varphi|n_0, t_0) = \kappa(n_0, t_0)c(\varphi)^{n_0}e^{n_0 t_0 \varphi}$. Combining such prior information with information from $Y_1, \ldots, Y_n \sim i.i.d.p(y|\varphi)$ gives the following posterior distribution:

$$p(\varphi|y_1, \ldots, y_n) \propto p(\varphi)p(y_1, \ldots, y_n|\varphi) \propto c(\varphi)^{n_0+n}\, exp\left\{\varphi \times \left[n_0 t_0 + \sum_{i=1}^{n} t(y_i)\right]\right\} \propto$$

$$\propto p(\varphi|n_0 + n, n_0 t_0 + n\,\bar{t}(\mathbf{y})),$$

where $\bar{t}(\mathbf{y}) = \sum t(y_i)/n$. The similarity between the posterior and prior distributions suggests that n_0 can be interpreted as a "prior sample size" and t_0 as a "prior guess" of $t(Y)$. This interpretation can be made a bit more precise:

$$E[t(Y)] = E[E[t(Y)|\varphi]] = E[-c'(\varphi)/c(\varphi)] = t_0,$$

so t_0 represents the prior expected value of $t(Y)$. The parameter n_0 is a measure of how informative the prior is. There are a variety of ways of quantifying this, but perhaps the simplest is to note that, as a function of φ, $p(\varphi|n_0, t_0)$ has the same shape as a likelihood $p(\tilde{y_1}, \ldots, \tilde{y}_{n_0}|\varphi)$ based on n_0 "prior observations" $\tilde{y_1}, \ldots, \tilde{y}_{n_0}$ for which $\sum t(\tilde{y_i})/n_0 = t_0$. In this sense the prior distribution $p(\varphi|n_0, t_0)$ contains the same amount of information that would be obtained from n_0 independent samples from the population.

Example: Binomial model.

The exponential family representation of the binomial(θ) model can be obtained from the density function for a single binary random variable:

$$p(y|\theta) = \theta^y\,(1-\theta)^{1-y} = \left(\frac{\theta}{1-\theta}\right)^y (1-\theta) = e^{\varphi y}(1+e^{\varphi})^{-1},$$

where $\varphi = \log\left[\frac{\theta}{(1-\theta)}\right]$ is the log-odds. The conjugate prior for φ is thus given by

$$p(\varphi|n_0, t_0) \propto (1+e^{\varphi})^{-n_0}e^{n_0 t_0 \varphi},$$

where t_0 represents the prior expectation of $t(y) = y$, or equivalently, t_0 represents our prior probability that $Y = 1$. Using the change of variables formula, this translates into a prior distribution for θ such that $p(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1}$, which is a $beta(n_0 t_0, n_0(1-t_0))$ distribution. A weakly informative prior distribution can be obtained by setting t_0 equal to our prior expectation and $n_0 = 1$. If our prior expectation is $1/2$, the resulting prior is a $beta(1/2, 1/2)$ distribution, which is equivalent to Jeffreys prior distribution for the binomial sampling model. Under the weakly informative $beta(t_0, (1-t_0))$ prior distribution, the posterior would be

$$\{\theta|y_1, ..., y_n\} \sim beta(t_0 + \sum y_i, (1 - t_0) + \sum (1 - y_i)).$$

Example: Poisson model.

The Poisson (θ) model can be shown to be an exponential family model with

$t(y) = y$;

$\varphi = \log \theta$;

$c(\varphi) = \exp\left(e^{-\varphi}\right)$.

The conjugate prior distribution for φ is thus $p(\varphi|n_0, t_0) = \exp(n_0 e^{-\varphi}) e^{n_0 t_0 y}$, where t_0 is the prior expectation of the population mean of Y. This translates into a prior density for θ of the form $p(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1} e^{-n_0 \theta}$, which is a $gamma(n_0 t_0, n_0)$ density. A weakly informative prior distribution can be obtained with t_0 set to the prior expectation of Y and $n_0 = 1$, giving a $gamma(t_0, 1)$ prior distribution. The posterior distribution under such a prior would be

$$\theta|y_1, ..., y_n - gamma(t_0 + \sum y_i, 1 + n).$$

§15. Bayesian Inference for the Normal Distribution

The concepts introduced in the preceding sections can be illustrated by Bayesian treatment of some simple problems involving the normal distribution. We will consider inference concerning an unknown mean with known variance.

First, suppose that the prior distribution of μ is $N(\mu_0, \sigma_0^2)$ (as we will see, the use of a normal prior distribution is especially useful analytically). A single observation $X \sim$

$N(\mu, \sigma^2)$ is taken: the posterior distribution of μ is

$$p(\mu/x) = \frac{f(x/\mu)p(\mu)}{\int f(x/\mu)p(\mu)d\mu} \propto f(x/\mu)p(\mu).$$

Therefore, we have

$$f(x/\mu)p(\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right] \propto$$

$$\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right] = \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{x}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) + \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right]\right\}$$

Denote by

$$a = \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2},$$

$$b = \frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2},$$

and

$$c = \frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2},$$

then the last expression may then be written as

$$\exp\left[-\frac{a}{2}\left(\mu^2 - \frac{2b}{a}\mu + \frac{c}{a}\right)\right].$$

To simplify this, we use the technique of completing the square rewrite the expression as

$$\exp\left[-\frac{a}{2}\left(\mu - \frac{b}{a}\right)^2\right] \times \exp\left[-\frac{a}{2}\left(\frac{c}{a} - \left(\frac{b}{a}\right)^2\right)\right].$$

The second term does not depend on μ, and we thus have that

$$p(\mu/x) \propto \exp\left[-\frac{a}{2}\left(\mu - \frac{b}{a}\right)^2\right].$$

We see that the posterior distribution of μ is normal with mean

$$\mu_1 = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

We have proved the following theorem.

Theorem 14.1. Suppose that $\mu \sim N(\mu_0, \sigma_0^2)$ and $X \sim N(\mu, \sigma^2)$. Then the posterior distribution of μ is normal with mean

$$\mu_1 = \frac{\dfrac{x}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{1}{\sigma^2} + \dfrac{1}{\sigma_0^2}}$$

and variance

$$\frac{1}{\dfrac{1}{\sigma^2} + \dfrac{1}{\sigma_0^2}}.$$

According Theorem 14.1, the posterior mean is a weighted average of the prior mean and the data, with weights proportional to reciprocal variances.

Let us now consider how the prior distribution $\mu \sim N(\mu_0, \sigma_0^2)$ is altered by observing a sample $X = (X_1, ..., X_n)$, where the X_i are independent and $N(\mu, \sigma^2)$. As before, the posterior distribution is

$$p(\mu/x_1, ..., x_n) \propto f(x_1, ..., x_n/\mu)p(\mu).$$

Now, from the independence of the X_i, we get

$$f(x_1, ..., x_n/\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

Using the identity

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

we obtain

$$f(x_1, ..., x_n/\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \times \exp\left[-\frac{1}{2\sigma^2/n}(\bar{x} - \mu)^2\right].$$

Only the last term depends on μ, so

$$p(\mu/x_1, ..., x_n) \propto \exp\left[-\frac{1}{2\sigma^2/n}(\bar{x} - \mu)^2\right] p(\mu) \propto$$

$$\propto \exp\left[-\frac{1}{2\sigma^2/n}(\bar{x} - \mu)^2\right] \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right] \propto$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2/n}(\bar{x}-\mu)^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right]\right\} =$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2/n}(\bar{x}^2-2\mu\bar{x}+\mu^2)+\frac{1}{\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)\right]\right\} \propto$$

$$\propto \exp\left\{-\frac{a_1}{2}\left[\mu-\left(\frac{\bar{x}n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\times\frac{1}{a_1}\right]^2\right\}.$$

Therefore this posterior distribution is normal with mean

$$\mu_1 = \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

and variance

$$\frac{1}{a_1} = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}.$$

Example 20. Suppose that the prior distribution of μ is $N(2,4)$ and that the X_i are $N(\mu,1)$. If

a) $x_1 = 3.59$ and $x_2 = 5.52$, then

$\bar{x} = 4.55$ and the posterior mean is

$$\mu_1 = \frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = 0.89\bar{x}+0.11\mu_0 = 4.2695.$$

b) Now suppose that we observe $x_1 = 3.59$, $x_2 = 5.52$, $x_3 = 3.93$ and $x_4 = 4.71$, then

$\bar{x} = 4.44$ and the posterior mean is

$$\mu_1 = 0.94\bar{x}+0.06\mu_0 = 4.30.$$

c) Finally, suppose that we observe $x_1 = 3.59$, $x_2 = 5.52$, $x_3 = 3.93$, $x_4 = 4.71$, $x_5 = 4.4$, $x_6 = 5.06$, $x_7 = 3.68$ and $x_8 = 3.14$, so

$\bar{x} = 4.25$ and the posterior mean is

$$\mu_1 = 0.97\bar{x}+0.03\mu_0 = 4.18.$$