Bayesian Statistics: Normal-Normal Model

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

December 3, 2008

Reference: The material in this note is taken from Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* New York: Springer.

Suppose that we have a class of $n = 30$ students who have taken an exam, and the mean grade was $\overline{x} = 75$ with a standard deviation of $\sigma = 10$. We have taught the class many times before, and past test means have given us an overall mean μ of 70, but the class means have varied over time giving us a standard deviation of the class means of $\tau = 5$.

Our goal is to update our knowledge of μ, the unobservable population mean test score with the new test grade data; i.e., we wish to find $p(\mu|X)$ where X is the new test data. Using Bayes' rule:

$$p(\mu|X) \propto p(X|\mu)\; p(\mu) \tag{1}$$

where $p(X|\mu)$ is the likelihood function for the current data and $p(\mu)$ is the prior for the test mean. Assuming the current test scores are Normally distributed with a mean of μ and a variance of σ^2, then our likelihood function for X is

$$p(X|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}. \tag{2}$$

Our previous test scores have provided us with an overall mean of 70, but we are uncertain about μ's actual value, given that class means vary semester by semester (giving us $\tau = 5$). So our prior distribution for μ is:

$$p(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{ -\frac{(\mu - M)^2}{2\tau^2} \right\} \tag{3}$$

where M is the prior mean (=70) and τ^2 (=25) reflects the variation of μ around M. Plugging the likelihood and prior into Bayes' rule gives us:

$$p(\mu|X) \propto \frac{1}{\sqrt{\tau^2\sigma^2}} \exp\left\{ \frac{-(\mu - M)^2}{2\tau^2} + \frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} \right\}. \tag{4}$$

This posterior can be re-expressed as a Normal distribution, but it takes some algebra to do so. Since the terms outside the exponential are normalizing constants with respect to μ, we can drop them. We therefore focus on the exponential. Let's re-write the terms inside the exponential:

$$-\frac{1}{2}\left[\frac{\mu^2 - 2\mu M_M 2}{\tau^2} + \frac{\sum x^2 - 2n\overline{x}\mu + n\mu^2}{\sigma^2} \right]. \tag{5}$$

Any term that does not include μ can be viewed as a proportionality constant, can be factored out of the exponent, and can be dropped (recall that $e^{a+b} = e^a e^b$). Using algebra (and dropping constants with respect to μ), we obtain

$$-\frac{1}{2}\left[\frac{\sigma^2\mu^2 - 2\sigma^2\mu M - 2\tau^2 n\overline{x}\mu + \tau^2 n\mu^2}{\sigma^2\tau^2}\right] \tag{6}$$

$$-\frac{1}{2}\left[\frac{(n\tau^2+\sigma^2)\mu^2 - 2(\sigma^2 M + \tau^2 n\overline{x})\mu}{\sigma^2\tau^2}\right] \tag{7}$$

$$-\frac{1}{2}\left[\frac{\mu^2 - 2\mu\frac{(\sigma^2 M + n\tau^2\overline{x})}{(n\tau^2+\sigma^2)}}{\frac{\sigma^2\tau^2}{(n\tau^2+\sigma^2)}}\right] \tag{8}$$

$$-\frac{1}{2}\left[\frac{\left(\mu - \frac{\sigma^2 M + n\tau^2\overline{x}}{(n\tau^2+\sigma^2)}\right)^2}{\frac{\sigma^2\tau^2}{(n\tau^2+\sigma^2)}}\right]. \tag{9}$$

In other words, $\mu|X$ is Normally distributed with mean

$$\frac{\sigma^2 M + n\tau^2\overline{x}}{n\tau^2+\sigma^2} \tag{10}$$

and variance

$$\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}. \tag{11}$$

After a bit of algebra, the mean can be re-written as

$$\frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}M + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\overline{x} \tag{12}$$

and the variance can be re-written as

$$\frac{\frac{\sigma^2}{n}\tau^2}{\tau^2+\frac{\sigma^2}{n}}. \tag{13}$$

This is an important result. Note that the mean is a weighted average of the prior mean M and the data mean $\overline{x}$. The weight on the prior mean is inversely proportional to the variance of the prior mean ($1/\tau^2$), and the weight on the data mean is inversely proportion to the variance of the data mean (n/σ^2). This makes sense, right? If the prior mean is very precise relative to the data mean, then we should weight it highly. Alternatively, if the data mean is more precise, then it should be assigned a larger weight. In addition, also note that the variance of $\mu|X$ is smaller than the variance of the prior mean (τ^2) and smaller than the variance of the data mean (σ^2/n). That is, combining the information from the prior and the data gives us a more precise estimate than if we used either information source by itself.

To illustrate the ideas presented so far, consider the following scenario. Suppose that data is sampled from a Normal distribution with a mean of 80 and standard deviation of 10 ($\sigma^2 = 100$). We will sample either 0, 1, 2, 4, 8, 16, 32, 64, or 128 data items. We posit a prior distribution that is Normal with a mean of 50 ($M = 50$) and variance of the mean of 25 ($\tau^2 = 25$). Figure 1 shows the posterior distribution of $\mu|X$ when we take different sample sizes. (The horizontal axis

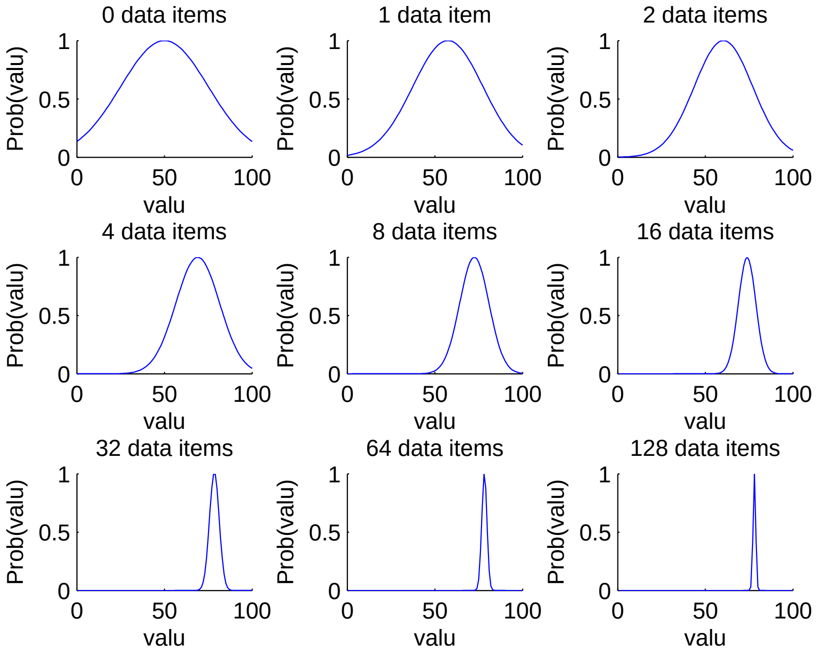

Figure 1: Posterior distribution of $\mu|X$. See text for explanation.

shows a value, and the vertical axis shows the probability assigned to that value by the posterior distribution. Actually, the probabilities have been linearly scaled so that the largest probability is always equal to 1.) Note that the upper left graph (0 data items) shows the prior distribution. With small sample sizes, the mean of the posterior distribution is a compromise between the mean of the prior distribution and the mean of the data. As sample sizes increase, the mean of the posterior distribution is closer to the mean of the data, and the variance of the posterior distribution shrinks.

This example is useful, but it can be regarded as unrealistic because we've assumed that the variance σ^2 is a known quantity. More realistically, we should try to estimate its value. A full probability model for μ and σ^2 would look like:

$$p(\mu, \sigma^2|X) \propto p(X|\mu, \sigma^2)\; p(\mu, \sigma^2). \tag{14}$$

We now need to specify a prior distribution for μ and σ^2. If we assume that these variables are independent, then $p(\mu, \sigma^2) = p(\mu)\; p(\sigma^2)$, and we can establish separate priors for each.

In this example, we assume noninformative priors for μ and σ^2. That is, we assume a uniform prior over the real line for μ and the same uniform prior for $\log(\sigma^2)$. We assign a uniform prior on $\log(\sigma^2)$ because σ^2 is a non-negative quantity, and the transformation to $\log(\sigma^2)$ stretches this new parameter across the real line. If we transform the uniform prior on $\log(\sigma^2)$ into a density for σ^2, we obtain $p(\sigma^2) \propto 1/\sigma^2$. Thus, the joint prior is $p(\mu, \sigma^2) \propto 1/\sigma^2$.

Using Bayes' rule, we can compute the posterior distributions for $\mu|X, \sigma^2$ and for $\sigma^2|X, \mu$. For the sake of brevity, we won't go through all the details here. Suffice it to say that $\mu|X, \sigma^2$ is Normally distributed with mean $\overline{x}$ and variance σ^2/n, and $\sigma^2|X, \mu$ has an inverse gamma distribution with parameters $a = n/2$ and $b = \sum(x_i - \mu)^2/2$.