Statistical Techniques Practice: Bayesian Statistics.

1. Use Matlab to draw graphs of Beta(a,b) distributions for different a and b, to get a sense of how the distributions depend on the parameters.

   Get a drawing pin (from me), and consider tossing it (but don't you dare toss it yet). Define the result of a toss to be "heads" if the point lands downwards, and "tails" otherwise.

   Write p for the probability that a toss will land point downwards.

   Still don't toss the drawing pin, but think about p, and choose a, b, so that a Beta(a,b) prior distribution approximates your subjective prior distribution for p.

   Now collect data. Toss the coin 100 times and keep track of the number of heads after 10, 50, and 100 tosses.

   Use Matlab to draw graphs of your prior distribution and your posterior distributions after 10, 50, and 100 tosses.

   Compare these plots with each other, and with those of your fellow drawing-pin-throwers.

2. Recall that if θ is normally distributed, $\theta \sim \mathrm{N}(\mu, \sigma^2)$, then the density function of θ has the form:

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)} \tag{1}$$

   where μ is the mean and σ^2 is the variance. You should particularly note that the exponent of the exponential function is *quadratic* in μ.

   Suppose we have some data, $\{y_1, \ldots, y_n\}$, which are *independent* samples from a normal distribution with *unknown* mean (μ), but *known* variance (σ^2). We wish to make inference about μ based on $\{y_i, \ldots, y_n\}$ and σ^2. Suppose further that our prior belief about possible values for μ is that they will follow a normal distribution: $\mu \sim \mathrm{N}(\mu_p, \sigma_p^2)$. In this case it is known that the posterior distribution of μ is also normal (and hence that the prior is conjugate to the likelihood). In fact, the posterior distribution for the mean is $\mathrm{N}(\mu_1, \sigma_1^2)$ where,

$$\mu_1 = \frac{\sigma_p^2 \cdot n \cdot \bar{y} + \sigma^2 \cdot \mu_p}{\sigma_p^2 \cdot n + \sigma^2}, \qquad \sigma_1^2 = \frac{\sigma^2 \cdot \sigma_p^2}{\sigma_p^2 \cdot n + \sigma^2}. \tag{2}$$

   (a) In Inter-Lab comparison for the validation of radiocarbon dating laboratories, a large sample of fixed age is sent to each lab. Each lab will make a number of measurements of this sample. The test sample used is not ancient, but the precise age is withheld from the labs.

   The labs estimate the variance of their own measurements. At one Lab (Oxford say) they make 200 repeat measurements of the age (measured in years before the present) of the sample. The measurements are recorded in the file rcages.txt. The Lab believes that each measured age is a draw from a normal distribution, with mean equal the true age, and variance 4yrs^2.

   The lab knows the sample is significantly older than 0 years, and from visual inspection of the material, is very unlikely to be as old as 300 years. The lab would like to estimate the unknown true age of the sample.

   Fix a prior (i.e. choose the mean and variance for your normally distributed prior) and use Matlab to plot the posterior distribution of the unknown true age of the sample based on the first n observations where $n = 5, 20, 50, 200$. For each value of n produce a plot showing your prior, the sample mean and the associated posterior. What do you notice about the plots as n gets larger? Explain your observations.

   Now change the variance of your prior (but not the mean) and repeat the process. Comment on any differences in the results.

(b) (optional, if time permits) Suppose you had no information at all about μ before taking your sample. One way to reflect this might be to choose μ_p somewhere finite and let $\sigma_p \to \infty$. What happens to the posterior in this instance?

(c) (optional, if time permits) Now we will prove the result used above, that if our prior distribution for μ is $\mathrm{N}(\mu_p, \sigma_p^2)$, then the posterior distribution of μ is also normal with mean and variance given by equation **??**. Note that in the calculations below, all the terms that do not depend on μ are just absorbed into the proportionality constant, so we can effectively ignore them.

i. Write down an expression for the prior distribution of μ, $\pi(\mu) = \mathrm{P}(\mu)$.

ii. Write $\bar{y}$ for the sample mean of $Y = \{y_1, \ldots, y_n\}$. Show that the likelihood of μ takes the form:

$$f(Y|\mu) = \mathrm{P}(Y|\mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right) . \tag{3}$$

(Hint: You will need to remember how to complete the square. You will also need to use the fact that the data are *independent* samples.)

iii. Hence, show that the posterior has the form:

$$f(\mu|Y) = \mathrm{P}(\mu|Y) \propto \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_p^2}\right) - 2\mu\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_p}{\sigma_p^2}\right) + \frac{\mu_p^2}{\sigma_p^2} + \frac{n\bar{y}^2}{\sigma^2}\right)\right) . \tag{4}$$

Note that the exponent of the exponential in this expression is quadratic in μ and thus the posterior is normally distributed.

iv. Let $a = \frac{n}{\sigma^2} + \frac{1}{\sigma_p^2}$, $b = \frac{n\bar{y}}{\sigma^2} + \frac{\mu_p}{\sigma_p^2}$ and $c = \frac{\mu_p^2}{\sigma_p^2} + \frac{n\bar{y}^2}{\sigma^2}$. Show

$$f(\mu|Y) = \mathrm{P}(\mu|Y) \propto \exp\left(-\frac{a}{2}\left(\mu - \frac{b}{a}\right)^2\right) . \tag{5}$$

Hence, show that the posterior distribution for the mean is $\mathrm{N}(\mu_1, \sigma_1^2)$ where,

$$\mu_1 = \frac{\sigma_p^2 \cdot n \cdot \bar{y} + \sigma^2 \cdot \mu_p}{\sigma_p^2 \cdot n + \sigma^2} , \qquad \sigma_1^2 = \frac{\sigma^2 \cdot \sigma_p^2}{\sigma_p^2 \cdot n + \sigma^2} . \tag{6}$$