Homework 1

Problem 1: Normal data with a normal prior

You will collect some data yourself. Data may not come from any already existing data set. Sample size should range from 4 to 10 observations. Data should be acceptably modeled as normal. Possible examples are travel times to (or from) school in the morning (or afternoon, but not both!), heart rate or blood pressure measurements from a blood pressure machine at the drug store, jogging times for a set distance.

The R command `dnorm(x, mean = 0, sd = 1)` can be used to calculate heights of the normal density, or you may calculate the density heights yourself $f(x|m, s) = (2\pi * s^2)^{-1/2} \exp(-.5(x-m)^2/s^2)$. Formulas for the posterior mean and posterior variance of μ are given in the lecture 1 notes.

1. Explain what your measurements will be.

2. Before you collect the data, decide on your prior. Please use a normal density, and specify your prior mean μ_0 and standard deviation τ. So μ_0 is your best guess at the average of all your blood pressure (or other) measures, and τ is an estimate of the standard deviation that the true value may differ from your guess. (Note: you get better at this with practice, you won't be penalized for being too ridiculous in your guessing, within reason.) Explain your reasoning (1 or 2 sentences).

3. Report the data and the sample mean and variance $(n-1)$ denominator.

4. Now specify the sampling standard deviation σ. Since we are doing a one parameter model, and since σ is usually *not* known, we need to do something because we are working with such a simple model. You may either

   (a) Pick a value for σ yourself, or

   (b) Set σ to the sample sd of your data set.

   (c) Specify the exact value for σ that you use in all your calculations (i.e. sqrt(2), 1.41, 1.414, or 1.4)

   Either way, this is commonly known as *cheating*; we often do this (Bayesians less often perhaps) in complicated models where treating the parameter as unknown complicates things substantially. If there were a later analysis, as we learn more about modeling and computation, we would relax the assumption of σ^2 known. Give your method of setting σ.

5. Calculate the posterior mean $\bar{\mu}$, posterior variance V, and posterior sd. Show the formulas for the posterior mean and variance with your data values in place of the symbols. Remember that in the likelihood, $\bar{y} \sim N(\mu, \sigma^2/n)$.

6. The *prior predictive* density is the density that you predict for a single observation before seeing any data. In this model, the prior predictive for a single observation is $y \sim N(\mu_0, \sigma^2 + \tau^2)$.

7. Construct a table with means, sds and vars for the (i) posterior for μ, (ii) the prior for μ, (iii) the prior predictive for y, and (iv) the likelihood of μ.

8. Plot on a single plot the (i) posterior for μ, (ii) the prior for μ, (iii) the prior predictive for y, and (iv) the likelihood of μ (suitably normalized so it looks like a density, ie a normal with mean $\bar{y}$ and variance σ^2/n) all on the same graph. *Interpret the plot.*

Problem 2: Count Data with a Gamma Prior

For $y_i|\lambda \sim \text{Poisson}(\lambda), i = 1, \ldots, n$, the conjugate prior is $\lambda \sim \text{Gamma}(a, b)$. The parameter b is the *rate* parameter and the mean of the $\text{Gamma}(a, b)$ distribution is a/b and the variance is a/b^2. The posterior given a sample of size n will be $\text{Gamma}(a + \sum_i y_i, b + n)$. You can calculate a `gamma(a,b)` density using `dgamma(x, shape=a, rate = b, log = FALSE)`, or by calculating the density yourself $f(x|a, b) = b^a * x^{(a-1)} \exp(-b * x)/\text{gamma}(a)$, where $\text{gamma}(a)$ is the gamma function.

You will collect some count data in this homework. You will specify two different priors, and work with both priors.

1. What is the support (place where density/function is non-negative) of: (i) prior, (ii) posterior, (iii) sampling density, (iv) likelihood?

2. In the prior $\text{gamma}(a, b)$, which parameter acts like a prior sample size? (Hint: look at the posterior, how does n enter into the posterior density?) You will need this answer later.

3. You will go (soon, but not yet!) to your favorite store entrance and count the number of customers entering the store in a 5 minute period. Collect it as 5 separate observations $y_1, \ldots, y_5$ of 1 minute duration each, this allows you to blink and take a break if needed. This will give you 5 data points.

4. Name your store, and the date and time.

5. We are now going to specify the parameters a and b of the gamma prior density. We will do this in two different ways, giving two different priors. We designate one set of prior parameters as a_1 and b_1; the other set of prior parameters are a_2 and b_2.

   (a) Before you visit the store, make a guess as to the mean number of customers entering the store in one minute. Call this m_0. This is the mean of your prior distribution for λ.

   (b) Make a guess s_0 of the prior sd associated with your estimate m_0. This s_0 is the standard deviation of the prior distribution for λ. Note: most people underestimate s_0.

   (c) Separately from the previous question 5b, estimate how many data points n_0 your prior guess is worth. That is, n_0 is the number (strictly greater than zero) of data points (counts of 5 minutes) you would just as soon have as have your prior guess of m_0.

   (d) Solve for a_1 and b_1 based on m_0 and s_0.

   (e) Separately solve for a_2 and b_2 using m_0 and n_0 only. You usually will not get the same answer each time. This is ok and is NOT wrong. (Note: if you do get the same answer, then please specify a second choice of a_2, b_2 to use with the remainder of this problem!)

6. Suppose we need to have a single prior, rather than two priors. Suggest 2 distinct methods to settle on a single prior.

7. Go to your store and collect your data as instructed in 3. Report it here.

8. Update both priors algebraically using your 5 data points. Give the two posteriors.

9. Give the posterior mean and sd for your two posteriors.

10. Plot your two prior densities on one graph. Plot your two posterior densities in another graph. (Use the algebraic formula, or you can use the `dgamma` function in R). In one sentence for each plot, compare the densities (talk about location, scale, shape and compare the two densities).

11. Plot each prior density/posterior density pair on the same graph. For each plot, compare the two densities in one sentence.

12. Extra Credit. (Recommended for Biostat grad students).

    (a) For this problem, treat the data as a single count y of customers that entered the store in 5 minutes. Define λ_1 as the 1 minute mean which you worked with previously. Define λ_5 as the 5 minute mean which you will work with now. Let a_5 and b_5 be the 5 minute prior parameters for λ_1 and similarly let a_1 and b_1 be 1 minute prior parameters from above.

    (b) Give algebraic formulas for the relationships between (i) λ_5 and λ_1, (ii) the *prior* mean of λ_5 and λ_1, (iii) prior variances, (iv) prior standard deviations, (v) prior a-parameters, and (vi) b-parameters. (Hint: Transformation-of-variables.)

    (c) Give the two priors for the parameter λ_5 that correspond to your priors for λ_1.

    (d) Give the two resulting posteriors for λ_5.

    (e) Explain the relationship between the posterior means of λ_5 and λ_1. Repeat for the posterior variance, posterior standard deviation, posterior a-parameters and finally posterior b parameters.

    (f) Do you need to redraw your plots (of priors and posteriors) that you drew in the previous problem? How could you alter them without redrawing to make them conform to the new data structure?

    (g) Do your conclusions change if you consider your data as a single 5 minute observation or as 5 one minute observations? That is, do your recommendations to the store on staffing levels change?