CONFIDENCE AND CREDIBILITY EXERCISE

(1) Consider again our simple model of data Y_t, $t = 1, \ldots, T$ on production from a plant with unknown capacity $\bar{Y}$. As before, the Y_t's are i.i.d. $U(0, \bar{Y})$ (i.e., uniform on the interval $(0, \bar{Y})$.

(a) Show that $Y_{\max}/\bar{Y}$ is a pivot for $\bar{Y}$ and find its pdf conditional on $\bar{Y}$. (Recall that if F is the cdf of an i.i.d. sequence of T random variables, the max of the sequence has cdf F^T.)

(b) Use the distribution you have found for the pivot to find an expression for a 95% confidence interval for $\bar{Y}$.

(c) Show that $\prod Y_t/\bar{Y}^T$ is also a pivot for $\bar{Y}$ and find its distribution conditional on $\bar{Y}$. [Recall from the last exercise that $\log(\bar{Y}/Y_t)$ is distributed as $\Gamma(1,1)$ and that sums of i.i.d. Γ's are Γ's.]

(d) Use the distribution you have found for this second pivot to generate an expression for a 95% confidence interval for $\bar{Y}$.

(e) Suppose our prior, at least over $(Y_{\max}, \infty)$, is proportional to $\bar{Y}^{-p}$. Find the formula for the implied 95% highest posterior density (HPD) interval. Does it, for some p, correspond to either of the confidence intervals?

(2) Often when we obtain data on a variable measuring an aggregate or count it is plausible that the data for larger observations are more accurate. The idea is that the larger entities are a cumulation of smaller elements and that in adding them up, errors tend to cancel out. This idea can be used to justify a model in which the variance of the observation is inversely related to its mean, or to be specific,

$$X_t \sim N\left(\mu, \frac{1}{\mu}\right), \; t = 1, \ldots, T .$$

(a) Assuming we have T i.i.d. observations on X_t for a single entity and that this one-parameter model for their distribution is correct, show that both

$$\sqrt{\mu}\left(\sum X_t - T\mu\right) \quad \text{and} \quad \mu \sum (X_t - \mu)^2$$

are pivots for μ and derive a 95% confidence region based on each. Assume $\sum x_t^2 = 72$, $\sum x_t = 18$ and $T = 6$.

(b) Assuming a flat prior on $(0, \infty)$, find a 95% HPD probability interval for μ. [You probably need to use the computer for this.]

(c) Is there any sense in which any of these intervals are better or worse than others?

(3) [You'll want to use either a hand calculator or a matrix-language computer program for this.] Suppose you have data on income, education and sex (this last a variable that is 1 for males, 0 for females) in a sample of 40 employed individuals collected in 1908. You consider modeling the data as a SNLM $Y = X\beta + \varepsilon$, with Y the vector of data on the natural log of income and X the matrix whose first column is a vector of ones, second column the education data (years of schooling), and third column the sex variable. The sufficient statistics are

$$X'X = \begin{bmatrix} 40 & 360 & 23 \\ 360 & 4240 & 280 \\ 23 & 280 & 23 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 42 \\ 514 \\ 36 \end{bmatrix}, \quad Y'Y = 80\,.$$

(a) Calculate the least-squares estimate of β and the posterior covariance matrix of β around this estimate under a $d\sigma/\sigma$ prior.

(b) Plot a 95% joint posterior HPD region for the coefficients of education and sex, again under the $d\sigma/\sigma$ prior.

(c) Suppose a previous study had concluded from these data that "there is no statistically significant evidence of sex differences in earnings, once education is accounted for, in these data." Would you agree with this summary of the evidence? Why or why not?