STA 215: Statistical Inference
HW 4 Due Wed Apr 04 2012

1. Consider the Gaussian linear model $Y_i = z_i^T\beta + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} N(0,\sigma^2)$. As always, let $y = (y_1,\cdots,y_n)^T$ denote the vector of observations on Y_i's, Z denote the $n\times p$ design matrix, $\hat\beta_{\text{LS}} = (Z^TZ)^{-1}Z^Ty$ and $s^2_{y|z} = \frac{1}{n-p}\sum_{i=1}^n (y_i - z_i^T\hat\beta_{\text{LS}})^2 = \|y - Z\hat\beta_{\text{LS}}\|^2/(n-p)$.

   (a) For the reference prior $\pi(\beta,\sigma^2) = \text{const}/\sigma^2$, show that the posterior distribution (given $Y = y$) is $N\chi^{-2}(\hat\beta_{\text{LS}}, Z^TZ, n-p, s^2_{y|z})$ whenever $n > p$

   [Hint: use the identity: $\|y - Z\beta\|^2 = \|y - Z\hat\beta_{\text{LS}}\|^2 + \|Z(\hat\beta_{\text{LS}} - \beta)\|^2$ to simplify the log-likelihood function. You should be able to write a "proof" in about 5 lines.]

   (b) For $\pi(\beta,\sigma^2) = N\chi^{-2}(m_0, K_0, r_0, s_0^2)$ where m_0 is a real number, K_0 is a $p\times p$ positive definite matrix, $r_0 > 0$ and $s_0 > 0$, show that the posterior is $N\chi^{-2}(m_n, K_n, r_n, s_n^2)$ where

$$m_n = (K_0 + Z^TZ)^{-1}(K_0m_0 + Z^Ty), \quad K_n = K_0 + Z^TZ$$
$$r_n = r_0 + n, \quad r_ns_n^2 = r_0s_0^2 + y^Ty + m_0^TK_0m_0 - m_n^TK_n^Tm_n.$$

   (c) The formulas in part (b) are correct even when $n \le p$. However, for $n > p$, the expression for $r_ns_n^2$ simplifies to $r_0s_0^2 + (n-p)s^2_{y|z} + (\hat\beta_{\text{LS}} - m_0)^T(K_0^{-1} + (Z^TZ)^{-1})^{-1}(\hat\beta_{\text{LS}} - m_0)$. Also m_n can be re-written as $m_n = (K_0 + Z^TZ)^{-1}(K_0m_0 + Z^TZ\hat\beta_{\text{LS}})$. Establish this for the special case: $Y_i \overset{\text{IID}}{\sim} N(\mu,\sigma^2)$.

2. Suppose the duration (in seconds) of a smile of a certain eight week old baby follows a *Unif*$(0,\theta)$ distribution with $\theta > 0$ unknown. The data X consists of n observed durations $X_1,\cdots,X_n$ from the baby. We are interested in a future observation $X^* = X_{n+1}$.

   (a) Suppose θ is assigned a $Pa(a,b)$ prior with pdf $\pi(\theta) = ab^a/\theta^{a+1}$, $\theta > b$ for some $a,b > 0$. Write down the expression for $P(X_1 > x)$ in terms of a,b and an arbitrary $x > 0$.

   (b) An expert quantifies her beliefs as

   - X_1 is equally likely to be smaller or larger than 10.
   - When $X_1 > 10$, it is equally likely to be smaller or larger than 20.

   Identify $a > 0, b > 0$ that match these quantifications.

   (c) Based on the prior chosen in part (b) and observations (10.4, 19.6, 12.8, 14.8, 1.3, 0.7, 5.8, 6.9, 8.9, 9.4), what is the posterior predictive probability that X^* would exceed 15?

   (d) Based on the same observations, what is the plug-in predictive probability that X^* would exceed 15 when θ is estimated by its MLE?

3. Weekly food expenditures $X_1,\cdots,X_n$ of n Duke undergraduate students are modeled as $X_i \overset{\text{IID}}{\sim} N(\mu,\sigma^2)$, with (μ,σ^2) assigned a $N\chi^{-2}(m_0,k_0,r_0,s_0^2)$ prior. During the lecture on 03/16, we determined that $m_0 = 150$, $k_0 = 1$, $r_0 = 0.7$ and $s_0 = 29.45$ match our (collective) prior belief about these expenditures.

   (a) Give a 95% posterior predictive interval for a future observable X^* based on the above choice of the prior and the following observations ($n = 22$):

```
125 140 200 200 190 100 140 250 125 180 110
125 120 130 140 150 120 100  95 195  95 130
```

(b) Get the same but with prior $\pi(\mu, \sigma^2) = \text{const}/\sigma^2$, the reference prior.

(c) Compare the two answers.

4. In an experiment, $n_1 = 12$ infant rats were assigned to a high protein diet while $n_2 = 7$ rats were assigned to a regular diet. For each rat, body weight gain between 28th and 84th days after birth were recorded. Let $U_1, \cdots, U_{n_1}$ denote these measurements for the high protein group and $V_1, \cdots, V_{n_2}$ denote the same for the regular diet group. Consider the model $U_i \overset{\text{IID}}{\sim} N(\mu_1, \sigma^2)$, $V_j \overset{\text{IID}}{\sim} N(\mu_2, \sigma^2)$, U_i's, V_j's are independent, with model parameters $\mu_1, \mu_2 \in (-\infty, \infty)$ and $\sigma^2 > 0$.

   We are interested in inferring whether the high-protein diet leads to a superior body weight gain than the low-protein diet, based on the following observations:

Diet	Weight gain (grams)
High	134 146 104 119 124 161 107 83 113 129 97 123
Low	70 118 101 85 107 132 94

   In HW 2 (#3(d)) we calculated the p-value for $H_0 : \mu_1 \leq \mu_2$, based on ML tests, to be 0.038 based on data given below. Here we would look at a Bayesian analysis with the reference prior $\pi(\mu_1, \mu_2, \sigma^2) = \text{const}/\sigma^2$.

   (a) Write down the name and parameters of the posterior distribution of (μ_1, μ_2, σ^2). Argue why the posterior probability of $\mu_1 \leq \mu_2$ is exactly equal to the ML-based p-value for $H_0 : \mu_1 \leq \mu_2$.

   (b) Consider a future observable U^* from the high-protein group and a future observable V^* from the low-protein group. Calculate the posterior predictive probability of $U^* \leq V^*$.

   (c) What would you conclude about the relative effects of high and low protein diets on body weight gain? Explain.

5. Consider $X = (X_1, \cdots, X_n)$ modeled as $X_i \overset{\text{IID}}{\sim} g(x_i|\mu, \lambda)$, $\mu > 0$ unknown and $\lambda > 0$ known where

$$g(x_i|\mu, \lambda) = I(x_i > 0)\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2} \exp\left\{-\frac{\lambda(x_i - \mu)^2}{2\mu^2 x_i}\right\}$$

   is the inverse-Gaussian pdf. It is known that the expectation under this pdf equals μ. Find the Jeffreys prior for μ.

6. To study whether early injection of IV fluids could be harmful to patients with penetrating injuries to the torso, data were collected at Ben Taub General Hospital in Houston under two treatment settings. In the early resuscitation group, $n_1 = 309$ patients were given fluids before they reached the hospital. Another $n_2 = 289$ patients in the delayed resuscitation group did not receive any fluid until they reached the operation theater. Let X_1, X_2 denote the number of survivors in the two groups (subscript 1 is for "early resuscitation").

   (a) Consider the model $X_1 \sim Bin(n_1, p_1)$, $X_2 \sim Bin(n_2, p_2)$, X_1 and X_2 are independent, with $\pi(p_1, p_2) = 1$ for $(p_1, p_2) \in [0, 1]^2$. Show that given observations $X_1 = 193$ and $X_2 = 203$, the posterior distribution of $p_1 - p_2$ is well approximated by a $N(a, b^2)$ distribution and find the values of a and b. [Hint: use the fact that a $Be(a, b)$ distribution, for large $a > 0, b > 0$ is well approximated by $N(m, s^2)$ where $m = a/(a + b)$ and $s^2 = m(1 - m)/(a + b + 1)$.

(b) Summarize the evidence against $H_0 : p_1 = p_2$ under the (approximate) posterior distribution in part (a).

(c) Toward a formal testing of $H_0 : p_1 = p_2$, denote the model in part (a) as M_1 and let M_0 denote the model: $X_1 \sim Bin(n, p)$, $X_2 \sim Bin(n, p)$, X_1, X_2 independent with p assigned the $Unif(0, 1)$ prior distribution. Calculate the Bayes factor of M_1 to M_0.

(d) Based on (b) and (c), what do you conclude about the relative effectiveness of the two resuscitation treatments?