STA 290 STATISTICAL LABORATORY **HOMEWORK 3**
Fall Semester 2002

Solutions should be written up using Latex; include graphs only where required.

Exercise (1)

Lung tumor samples from n patients are allocated into two tumor types: non-recurrent tumor (0) or recurrent tumor (1). There are n_0 non-recurrent and n_1 aggressive recurrent tumors. A specific protein is recorded as being present ($g_i = 1$) or not-present ($g_i = 0$) for each tumor. It is of interest to explore whether or not the presence/absence of the protein indicates whether or not the tumor is recurrent/non-recurrent. This is addressed in a model under which the g_i are independent Bernoulli's with $Pr(g_i = 1) = \pi_0$ for non-recurrent tumors and $Pr(g_i = 1) = \pi_1$ for recurrent tumors.
Write H_1 for the assumption (hypothesis) that $\pi_0 = \pi_1$, taking a common value π. Under H_1, take the prior for π to be $U(0, 1)$. Write H_2 for the alternative hypothesis $\pi_0 \neq \pi_1$. Under H_2, suppose that π_0 and π_1 are independent with uniform priors $U(0, 1)$.

Write $G = \{g_1, \ldots, g_n\}$ and suppose that the data indicates x_0 of the n_0 non-recurrent tumors have the protein, while x_1 of the n_1 recurrent tumors have the protein; write $x = x_0 + x_1$.

(a) Evaluate the marginal density $p(G|H_1) = \int_0^1 p(G|\pi)p(\pi)d\pi$ as a function of (n, x).

(b) Evaluate the related marginal density $p(G|H_2)$ as a function of (n_1, x_1) and (n_0, x_0).

(c) Suppose that you assign $Pr(H_1) = 0.5$. Give an expression for the posterior probability $Pr(H_1|G)$ in terms of $p(G|H_1)$ and $p(G|H_2)$.

(d) The clinical research protocol selects $n_0 = 34$ non-recurrent and $n_1 = 40$ aggressive recurrent tumors, and the analysis reports $x_0 = 5$ and $x_1 = 17$. Calculate the Bayes' factor $BF = p(G|H_1)/p(G|H_2)$ (it should be ≈ 0.13.) What is the posterior probability $Pr(H_1|G)$ in this case? Is this evidence for or against an association between the protein and the tumor type?

(e) Based on this specific data set, and assuming H_2, what are the posterior distributions for π_0 and π_1? What are the posterior means of π_0 and π_1?

(f) The incidence rate of aggressive tumors is about 15% – i.e., for a randomly selected patient, the probability of a recurrent tumor is about 0.15. A further patient is assessed for presence of the protein; and it is determined that the protein is present. Assuming H_2, give an expression for the probability, θ, that this patient has an aggressive tumor, in terms of π_0 and π_1.

(g) With the data values above, what is the MLE of θ?

(h) Compute the (approximate) posterior mean, median, and a 90% posterior credible interval for θ. Include a graph of the posterior distribution of θ (approximate).

Exercise (2)

Estimation from two normal experiments: an experiment was performed on the effects of magnetic fields on the flow of calcium out of chicken brains. The experiment involved two groups of chickens: a control group of 32 chickens, and an exposed group of of 36 chickens. One measurement was taken on each chicken, and the purpose of the experiment was to measure the average flow μ_c in untreated (control) chickens and the average flow in μ_t in treated chickens. The 32 measurements on the control group had a sample mean $\bar{X}_c$ of 1.013 and a sample standard deviation s_c of 0.24. The 36 measurements on the treatment group had a sample mean $\bar{X}_t$ of 1.173 and a sample standard deviation s_t of 0.20.

(a) Assuming control measurements were taken at random from a normal distribution with mean μ_c and variance σ_c^2, what is the posterior distribution of μ_c? Similarly, what is the posterior distribution of μ_t? In both cases, assume a uniform (improper) prior distribution on $(\mu_i, \log \sigma)_i)$ for $i \in \{c, t\}$

(b) What is the posterior distribution of the difference $\mu_t - \mu_c$? To get this, you may sample from the independent Student-t distributions you obtained in part (a) above. Plot a histogram of your samples and give an approximate 95% posterior interval for $\mu_t - \mu_c$.

Exercise (3)

EDA: The `fish` data under the Datasets link contains information on mercury concentrations in (ppm) for large mouth bass filets from two North Carolina rivers. The weight (grams) and length (cm) of each fish were recorded as well as an indicator variable for the river (0=Lumber, 1 = Wacamaw) and a number for the station on that river (0, 1, ... 15). Note station 1 on the Wacamaw is not the same as station 1 on the Lumber river, as station is *nested* in river. You may want to create a new factor variable for location with 32 levels. Each fish caught corresponds to a single row of the file. The goal is to develop a predictive model of mercury concentrations based on the other covariates. Note: S-Plus users will need to delete the lines with comments at the top of the data file before reading it in to S-Plus.

(a) Carry out an exploratory analysis of the data, looking for transformations that make the relationship between mercury and the other variables. Are the relationships between length, weight and mercury the same for both rivers/stations? Are there any stations that are outliers?

(b) Suppose we are only interested in comparing the proportion of fish with mercury concentration greater than 1.0 ppm by location (measuring station and river). Ignoring weight and length for now, analyze the data as follows:

   (a) Create a dummy variable for mercury concentration ($>$ 1.0 ppm vs.$<=$ 1.0 ppm).

   (b) Create a vector of total counts by station within river, and a vector of counts of fish with large mercury concentration by station and river. (these are your n and y binomial data; you may want to use the tabulate command for this).

   (c) Create an error.bar plot and comment on it. (see the rankings handout)

   (d) Analyze the posterior for the rankings and describe your findings. (For example, use 1000 simulations from the posterior, plot 95% credible intervals for the posterior ranks for all stations.

   (e) A concentration over 1 part per million is considered unsafe for human consumption. In light of this, what recommendations can you make for these stations/rivers? Write up a one page summary directed at fish managers (rather than statisticians).