FIRST YEAR EXAM 1997 – IN CLASS PORTION

Name _____

Notes:

1. This is a closed book exam. No notes are permitted.
2. It is to your advantage to show your work and explain your answers. Don't erase; draw a line through work you don't want graded.
3. You must answer 4 of the 6 problems in the test. Indicate which problem you do not want graded. If this is unclear, the first 6 problems will be graded.
4. Few complete answers are more valuable than many partial answers.
5. You have 3 hours to finish.

Problem	1	2	3	4	5	6	7	Total
Score								

1. In a Poisson process of unknown rate λ, x events are observed in time t. Consider two alternative models:

1. Assume t was fixed, so that x had a Poisson distribution with parameter λt.
2. Assume x was fixed, so that t had a gamma distribution $Ga(x, \lambda)$.

With t fixed, x/t is the classical minimum variance unbiased estimator (UMVUE) of λ;
Hint: The probability mass function for $x \sim Poi(\lambda t)$ distribution is:

$$Pr(x = j|\lambda, t) = \frac{1}{x!}(\lambda t)^x e^{-\lambda t}.$$

The p.d.f. for $t \sim Ga(x, \lambda)$ distribution is:

$$p(t|x, \lambda) = \frac{\lambda^x}{\Gamma(x)} t^{x-1} e^{-\lambda t}.$$

Also, you might want to use the following property of the Γ-function: $\Gamma(x + 1) = x \cdot \Gamma(x)$.

a. Show that under Model 2 the statistic $T = (x - 1)/t$ is an unbiased estimator of λ.

b. Show that under Model 2, T is uniformly minimum variance unbiased (UMVUE) for λ.

c. Conclude that using the UMVUE can violate the likelihood principle.

2. Let t_i be the number of misprints and N_i is the number of characters on page i in a certain book. Consider two alternative models to describe variation:

Model H_1:

$$\begin{aligned} t_i &\sim Poi(\lambda \cdot N_i), \quad i = 1, \ldots, n \\ \lambda &\sim Ga(1, 1/\beta) \end{aligned}$$

and model H_2:

$$\begin{aligned} t_i &\sim Bin(N_i, \theta) \\ \theta &\sim Be(1, \beta - 1), \end{aligned}$$

where $\beta > 1$ is a fixed hyperparameter.

Interpret β. Does it have the same meaning in both models?
Find the Bayes factor $B = p(y|H_2)/p(y|H_1)$ for choosing between models H_1 and H_2.

Hints: The p.d.f. for $t \sim Ga(a, b)$ distribution is:

$$p(x|a, b) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-xb}.$$

The p.d.f. for $x \sim Be(a, b)$ distributio

$$p(x|a, b) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1}$$

Neither $p(y|H_2)$ nor $p(y|H_1)$ take the form of any well-known distribution.

3. Data $x_1, \ldots, x_n$ are modelled as a random sample from the normal mixture

$$(1-\pi)N(x_i|\mu_0, 1) + \pi N(x_i|\mu_1, 1)$$

where μ_0, μ_1 and π are unknown. The histogram of a data set with $n = 200$ appears in each of the accompanying figures. A Gibbs sampling MCMC analysis iteratively generates approximate posterior samples from the posterior for (μ_0, μ_1, π). Two such analyses are summarised in the accompanying figures. In each case, π has prior $U(\pi|0, 1)$, independently of the μ_j; the analyses differ only in the choice of prior $p(\mu_0, \mu_1)$. In each case the MC sample size is 5000 after a period of burn-in.

Figure 1 displays some outputs from the MCMC analysis under a prior in which μ_0 and μ_1 are independent $U(\cdot| -10, 10)$. The figure gives plots of the marginal histograms for μ_0, μ_1, and π, the corresponding time series plots of the MCMC iterates, and a bivariate scatter plot of the (μ_0, μ_1) pairs with the $\mu_1 = \mu_0$ line superimposed.

Figure 2 displays similar plots from analysis in which μ_0 and μ_1 follow $U(\cdot| -10, 10)$ priors but then *conditioned* so that $\mu_0 < \mu_1$.

Discuss the figures and relate what is displayed to differences between the two analyses. You should describe how each of the figures differs between the two analyses, what the differences mean, and how/why they arise. Be clear and specific in describing features of each graph, and relating the features to aspects of the underlying model and analysis. (You **do not** need to discuss or develop technical aspects of the MCMC analysis.) Finally, briefly summarise your posterior inferences about values of the parameters μ_0, μ_1 and π.

4. After hurricane Fran many people in Durham were left without power. Two students, Chris and Robin were discussing their dilemmas. Chris was realitively lucky and had power restored after 48 hours, however Robin was still without power after 54 hours. If we are willing to assume that the time without power is exponentially distributed with mean λ and that Chris and Robin's outcomes are independent of each other, answer the following questions:

1. Write down the likelihood function of λ;
2. Find the maximum likelihood estimate of λ;
3. Find the maximum likelihood estimate for the probability that for another individual that the power will be out for more than 120 hours;
4. Using a non-informative prior distribution for λ, $p(\lambda) \propto 1$ for $\lambda > 0$, find the posterior distribution of λ given Chris and Robin's data;
5. A third student, Jo, joins Chris and Robin, and is interested in the probability that power will be restored by 120 hours given that the power hasn't yet been restored after 54 hours. Express this as a function of λ, say $\phi(\lambda)$;
6. Given Chris and Robin's data, what is the posterior mean for $\phi(\lambda)$.

5. Let $x = (x, y)$ have a pmf $\pi(x, y)$ defined over $\{0, 1\} \times \{0, 1\}$ by

(x, y)	(0,0)	(1,0)	(0,1)	(1,1)
$\pi(x, y)$	.4	.1	.1	.4

a. Although it is trivial to generate i.i.d. realizations from $\pi(x, y)$ directly, describe how you would construct a Gibbs sampler to generate dependent realizations from $\pi(x, y)$ by updating x and y in turn.

b. Suppose $(x^0, y^0) \sim \pi(x, y)$ and x^1 is obtained after updating x^0 a single time according to the x component update in your Gibbs sampler. Is (x^1, y^0) also a draw from $\pi(x, y)$?

c. Suppose $(x^0, y^0) \sim \pi(x, y)$ and you repeatedly update both x and y according to your Gibbs sampler so that you obtain the realization $(x^0, y^0), (x^1, y^1), \ldots, (x^T, y^T)$.

What is $\mu_x = E(\frac{1}{T}\sum_{t=1}^{T} x^t)$?

Will $\frac{1}{T}\sum_{t=1}^{T} x^t \to \mu_x$? Why? or Why not?

d. Suppose $(x^0, y^0) \sim \pi(x, y)$ and you repeatedly update only x according to your Gibbs sampler so that you obtain the realization $(x^0, y^0), (x^1, y^0), \ldots, (x^T, y^0)$.

What is $\mu_{x|y^0=0} = E(\frac{1}{T}\sum_{t=1}^{T} x^t | y^0 = 0)$?

What is $\mu_x = E(\frac{1}{T}\sum_{t=1}^{T} x^t)$?

Will $\frac{1}{T}\sum_{t=1}^{T} x^t \to \mu_x$? Why? or Why not?

e. One could transform (x, y) to (u, v) where

$$u = x \text{ and } v = \begin{cases} 0 & \text{if } x \neq y \\ 1 & \text{if } x = y \end{cases}$$

Describe a Gibbs sampler for the transformed variables (u, v).

f. Which of the two Gibbs samplers will result in a smaller standard error for $\frac{1}{T}\sum_{t=1}^{T} x^t$? Why?

6. The PrinStan Company offers a course for students wishing to improve their GRE scores. They claim to be able to raise scores by 50 points. We know that GRE scores for students who are not PrinStan clients follow a Normal distribution with mean 500 and standard deviation 100. Suppose that you have a random sample of 10,000 GRE scores, including some PrinStan clients. You may assume that the PrinStan clients are a random sample of all students taking the test. Under the following circumstances, how would you compare the PrinStan claim to the claim that the course is worthless?

1. PrinStan claims to raise each client's score by exactly 50 points and PrinStan tells you exactly which of the 10,000 scores belong to their clients.

2. PrinStan claims to raise each client's score by exactly 50 points and they tell you exactly how many scores belong to their clients, but not which ones.

3. PrinStan claims to raise each client's score by a random amount that is distributed with mean 50 and standard deviation 10 and they tell you exactly how many scores belong to their clients, but not which ones.

4. PrinStan claims to raise each client's score by a random amount that is distributed with mean 50 and standard deviation 10. You don't know exactly how many scores belong to their clients but you believe the number is between 500 and 1000.

FIRST YEAR EXAM 1997 – TAKE HOME PORTION

Name _____

Notes:

You have 12 hours to answer the questions related to this problem. The exam is due at Giovanni's office at 9pm (Boninsegna time) on 10 May 1997. Your solution should explain in simple language your model, results and recommendations. We are not looking for the most powerful analysis or anything like that. We are looking for an appropriate analysis that is clearly explained and that addresses the relevant questions.

Mercury levels in fish tissue for large mouth bass in the Wacamaw and Lumber Rivers

The Study

Rivers in North Carolina contain small concentrations of mercury which can accumulate in fish over their lifetimes. Because mercury cannot be excreted from the body, it builds up in the tissues. The concentration of mercury in fish tissue can be obtained at considerable expense by catching fish and sending samples to a lab for analysis. Directly measuring the mercury concentration in the water is impossible since it is almost always below detectable limits.

A study was recently conducted in the Wacamaw and Lumber Rivers in North Carolina to investigate mercury levels in tissues of large mouth bass. At several stations along each river, a group of fish were caught, weighed, and measured. In addition a filet from each fish caught was sent to the lab so that the tissue concentration of mercury could be determined for each fish. The data can be found in the file: `~higdon/fye/fish.dat`. In order, the recorded information for each fish is

column 1	river
column 2	station
column 3	length of fish in cm
column 4	weight of fish in grams
column 5	mercury concentration in parts per million

Data analysis

Researchers would like to address following questions with this study:

1. What physical attributes of the fish are associated with mercury concentration?
2. Are there differences in fish mercury levels between the rivers and/or stations?
3. A concentration over 1 part per million is considered unsafe for human consumption. In light of this, what recommendations can you make for fish caught from these rivers? Does the size of the fish matter? Does it matter which river the fish was taken from?
4. A second use of this data is to identify regions along the river where mercury may be entering one or both of the rivers. Is there any evidence that suggests a difference in mercury levels by station? Note that the large mouth bass tend to stay within a few hundred yards of their original locations.

Present a report that addresses each of the above questions. The presentation need not be pretty and can be given in outline form. But it should convey sensible answers to the questions. You will likely want to give

- relevant graphical summaries of the data,
- explanations of any models used,
- and explanations of your conclusions.