FIRST YEAR EXAM 1998 – 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. Please use a *separate* page for each problem (different people will grade different problems). Please put your name on each sheet of your exam.
4. You must answer 5 of the 6 problems in the test. Indicate which problem you do not want graded. If this is unclear, the first 5 problems will be graded.
5. Few complete answers are more valuable than many partial answers.
6. You have 3 hours to finish.

Problem	1	2	3	4	5	6	Home	Total
Score								

1. Determine in each case a sufficient one-dimensional statistic (that is, a single real valued function of the observations) based on a random sample of size n, $n > 1$;

1a from a Bernoulli population, $f(x;p) = p^x(1-p)^{1-x}, x = 0, 1$;

1b from a geometric population, $f(x;p) = p(1-p)^x, x = 0, 1, \ldots$;

1c from an exponential population, $f(x;\lambda) = \lambda e^{-\lambda x}, x > 0$;

1d from a uniform population, $f(x;\theta) = 1/\theta, 0 < x < \theta$.

2. A sign in an elevator reads: "Capacity, 3,000 lb or 20 persons." The distribution of weights of people who might ride the elevator has an unknown mean μ and a standard deviation $\sigma = 20$ lb. Suppose that the 'limit' of 20 people is based on the requirement that the probability that a full load exceeds 3000 pounds must be no more than 0.20. Find the presumed population mean μ.

3. A criminal suspect undergoing a polygraph test is either guilty ($\theta = 1$) or innocent ($\theta = 0$). His answer to the question "Are you innocent?" is "Yes." As a result of the test, the expert polygraph examiner declares the suspect to by lying, denoted by $y = 1$, or truth-telling, denoted by $y = 0$. Let $p_0 = Pr(y = 0|\theta = 0)$ and $p_1 = Pr(y = 1|\theta = 1)$ be the probabilities of correct determinations for innocent and guilty suspects, respectively.

3a. Assuming that you know the values of p_0 and p_1, describe how to calculate the probability π that the suspect is actually guilty if the examiner declares him to be lying. State any additional information which you need to assume to do this calculation.

To provide estimates of p_0 and p_1 needed to compute π, a pre-trial study is performed. Here the examiner administers the polygraph test to 20 "guilty" people and 18 "innocent" people, and each was asked "are you innocent?" All participants are instructed to say "yes" so that the guilty ones are lying, the innocent ones are not lying. The conditions of the test are otherwise exactly as used in testing a real suspect. Let x be the number out of the 20 liars that the examiner correctly identifies, and let z be the number out of the 18 truth-tellers that the examiner correctly identifies. Assume independence across participants.

3b State the distribution of x given p_1. State the distribution of z given p_0.

3c The pre-trial hearing led to data $x = 17$ and $z = 18$. What are the maximum likelihood estimates of p_0 and p_1?

3d Previous polygraph studies indicate that a typical examiner has about an 85% chance of correctly detecting liars, and about an 80% chance of correctly detecting truth-tellers. On this basis, discuss the relevance of the priors $p_1 \sim Beta(0.85, 0.15)$ and $p_0 \sim Beta(0.80, 0.20)$?

3e Assuming the priors in (d) above, what are the corresponding posteriors for p_0 and p_1? Find the posterior means.

3f Explain (briefly) how you would assess whether or not the examiner is in fact better at detecting liars than truth-tellers on the basis of this data $x = 17$ and $z = 18$.

4 Ramsey and Schafer (1997) analyze data from an observational study by Grayson (1990) on the survival of the Donner and Reed families who traveled from Illinois to California by covered wagon in 1846. A number of the adult group did not survive the journey and it is of interest to see if the survival status is related to the gender and age of the individuals. For the ith adult in the party, denote by y_i the survival status y_i (1 if survive and 0 if did not survive) of the ith adult, and let AGE_i and SEX_i (1 if male and 0 if female) denote the corresponding age and sex variables. Let p_i denote the probability that the ith adult survives the journey. Assume that a probit model of the form

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{SEX}_i$$

describes the probability of a given adult's survival.

4a. Describe a Gibbs sampling sampling scheme (perhaps based on an underlying latent trait) that might be used to simulate from the posterior distribution of the regression vector $\beta = (\beta_1, \beta_2, \beta_3)$. Specify all conditional distributions in detail.

4b. Suppose that you ran the Gibbs sampler described in part (a) to obtain simulated values of β, denoted by, say, $\beta^1, \beta^2, \ldots, \beta^s$. How could these values be used to approximate the distribution of the posterior-predictive
residuals for each individual? How would you interpret these distributions?

4c. Suppose you were given the MAP (posterior mode) $\tilde{\beta}$ and asymptotic covariance estimates of $\tilde{\beta}$ estimate for the regression model above as well as the reduced model without AGE. (Assume that these estimates were obtained with a proper, informative prior for the regression parameters in each model.) How could you approximate the Bayes factor for comparing the model that included AGE to the model that didn't?

4d. Let $\hat{\beta}$ denote the MLE estimate of β. Define the deviance residuals for the full model (with AGE and SEX).

5. Unknown to public health officials, a person with a highly contagious disease enters the population. During each period the infected individual either infects a new person, which occurs with probability p, or symptoms appear and the individual is discovered by public health officials, which occurs with probability $1-p$. Assume each infected individual behaves like the first.

5a. Compute the probability distribution of the number x of infected people in the population at the time of first discovery of a carrier.

5b. Assuming a uniform prior distribution $p \sim U(0,1)$, find the posterior mean for p given that we observe $x = 2$.

6. From a normal linear regression with 30 observations, we found that the largest studentized residual was 5.8 with a p-value of $4.78462e-06$. (All other p-values were greater than 0.22.) The Cook's distance for this case was .14.

6a. Explain how it is possible to have an observation that is a significant outlier, yet have a small Cook's distance.

6b. What would be the impact of deleting this observation (i.e. significant outlier, but small Cook's distance) on predictions? Little or a lot? Why?

6c. What would be the impact of deleting this observation on the length of prediction intervals?

FIRST YEAR EXAM 1998 – TAKE HOME PORTION

Name _____

Notes:

You have 12 hours to answer the questions related to this problem. The exam is due at Peter's office at 9pm (diego time) on Mo, 11 May 1998. Your solution should explain in simple language your model and results. 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.

Tropical rain forests have up to 300 species of trees per hectare. To gain insight into species responses, a sample of seeds were selected from eight species selected to represent the range of regeneration types which occur in this forest community. This representative community was then placed in experimental plots manipulated to mimic the natural variation in light conditions found in rain forests. Mammals were excluded from one half of each plot in order to assess their effects on the regeneration of rain forest trees. Six seeds of each type were planted on each half plot and an indicator of whether they survived was recorded. The two questions of specific interest are:

a. How are species different in their probabilities of survival?

b. Are there interactions that influence survival probabilities?

The response variable is

SURV $\in \{0, 1\}$ (No = 0, Yes = 1): Survival – Indicator of whether a given seedling was present at the end of the observation period.

The covariates are:

UND $\in \{0, 1\}$ (No = 0, Yes = 1): Indicator for under-story versus clearing

MAM $\in \{0, 1\}$ (0 = Absent, 1 = Present): Enclosure – Indicator whether there was an enclosure to prevent mammals from eating the seeds

LITTER $\in \{0, 1, 2, i\}$: categorical variable indicating the type of leave cover on the forest floor.

LIGHT $\in R^+$: measure of light levels at the forest floor

SPECIES $\in \{a, b, g, h, i, m, r, s\}$: categorical variable indicating the species. Species are coded as:

Code	Name	Size	Cotyledon type
a	Ardisia	3	H
b	C. biflora	7	H
g	Gouania	1	E
h	Hirtella	8	H
i	Inga	4	H
m	Maclura	2	E
r	C. racemosa	6	H
s	Strychnos	5	E

Size = 1 (smallest) to 8 (largest); E = Epigeal; H=Hypogeal.

Epigeal species rely on the cotyledons (the first leaves which a plant produces) for photosynthesis and production of energy to become established. Seed size tends to be small, with little reserves in the seeds.

Hypogeal species tend to have larger seeds, and rely on reserves in the seed to produce energy, thus if initial leaves are lost to predators, there may still be additional reserves that can be used to produce additional leaves. Larger seeds, however, may be easier to spot by predators.

Litter is not expected to have as big an impact on survival as the other variables. There are reasons to believe that there could be strong interactions between species type and UND, as some species are shade intolerant, and interactions between the species and MAM because of the seedling size and method of establishment. Higher order interactions are plausible.

Download the data set from `http://www.stat.duke.edu/~pm/surv.dat`. The data set has 9 columns:

1. PLOT_NUMBER
2. SUBPLOT_NUMBER (within the plot)
3. SPECIES
4. INDIV (seedling number within plot/subplot)
5. SURV
6. LIGHT
7. LITTER
8. MAM
9. UND

The first line gives the names of the data columns. The data set has 3071 data records. The first few lines are:

```
PLOT SUBPLOT SPECIES INDIV SURV  LIGHT LITTER MAM UND

1   2   a   1   0   10000   i   0   0
1   2   a   2   0   10000   i   0   0
1   2   a   3   0   10000   i   0   0
1   2   a   4   0   10000   i   0   0
1   2   a   5   0   10000   i   0   0
1   2   a   6   0   10000   i   0   0
1   2   b   1   1   10000   i   0   0
1   2   b   2   0   10000   i   0   0
1   2   b   3   0   10000   i   0   0
1   2   b   4   0   10000   i   0   0
1   2   b   5   0   10000   i   0   0
1   2   b   6   0   10000   i   0   0
1   2   g   1   0   10000   i   0   0
.....
```