NAME:

ISDS PHD PROGRAM **FIRST YEAR EXAM 1994**

You have four hours to complete the exam. You may not use any reference materials (notes, books, etc.).

Full credit can be obtained by answering six questions correctly. Few complete answers are more valuable than many partial answers. Should you attempt more than six questions, please indicate which question(s) you do not want to be graded. Also, you must **show** some **work** to get credit— *unsupported answers are not acceptable.* Blank sheets of paper are provided, for use as scratch paper or work sheets. Attach work sheets to the exam before returning it. It is to your advantage to write your solutions as clearly as possible, since you cannot receive credit for solutions we do not understand. You may need a hand calculator for some questions.

Good luck!

1. (STA 205) (a) Let $X_1, X_2, \ldots,$ be iid random variables with $EX_n = 0, Var(X_n) = 1.$ Prove

$$Y_n = \frac{(X_1 + \ldots + X_n)\sqrt{n}}{X_1^2 + \ldots + X_n^2} \Rightarrow \mathcal{N}(0,1), \quad n \to \infty.$$

(b) Let $X_1, X_2, \ldots$ be a sequence of iid random variables such that $EX_n = 0$ and $EX_n^2 = \sigma^2.$ Prove that the sequence

$$Y_n = (\sum_{i=1}^{n} X_i)^2 - n\sigma^2$$

is a martingale.

2. (STA 213) Jack and Jill are partners in a typing service. Jack makes typographical errors at an average rate of one per page while Jill makes errors at a rate of one per 4 pages. Assume that for each typist these errors occur independently and at a constant rate throughout the paper. If you have to make additional assumptions to answer then indicate this and **state them explicitly.**

(i) You submit a 4-page paper to Jack for typing. What is the probability he types it with no errors?

(ii) You submitted a 4-page paper to Jack for typing and learn that none of the 4 pages is errorless. (That is, each of the four pages contains at least one error.) What is the expected number of errors in the paper given this information?

(iii) You submit a 4-page paper to the partnership for typing without knowing whether Jack or Jill will type it. It comes back error-free. What is the probability that Jack typed it?

3. (STA 213) (a) Consider testing the null hypothesis $H_0 : \mu < 0$ against the alternative $H_1 : \mu > 0$ on the basis of a random sample $(X_1, \ldots, X_n)$ where X_i is $N(\mu, \sigma^2)$, σ^2 is unknown and $n \geq 2$. Show that the likelihood ratio test is a t test.

(b) Consider a random sample $(X_1, \ldots, X_n)$ from a variable that is uniform on the interval $(0, \theta)$.

(i) Find a one-dimensional sufficient statistic. Support your answer.

(ii) Show that $2\bar{X} = (2/n) \sum_{i=1}^{n} X_i$ is an unbiased estimate of θ.

(iii) Say why the estimate in part (ii) is unreasonable. [One way to do this is using a specific example with made-up data.]

(iv) Assume quadratic loss. Use the Rao-Blackwell theorem and part (i) to improve $2\bar{X}$ as an estimate of θ.

4. (STA 215) Let X_{in}, $i = 1, 2$, $n \geq 1$, be two exchangeable sequences of binary observations.

(i) Use the de Finetti representation theorem to write $p(x_{11}, ..., x_{1n}, x_{21}, ..., x_{2m})$ as a mixture of distributions depending on two parameters θ_1 and θ_2.

(ii) Is there a minimal sufficient statistic?

(iii) Is there a conjugate prior distribution for (θ_1, θ_2)?

(iv) Find an asymptotic approximation to the posterior distribution of

$$\varphi = \frac{\theta_1 - \theta_2}{\theta_1}.$$

(v) Use the result in (iv) to construct a 95% confidence interval on φ.

5. (STA 216) Consider the logistic regression model

$$Y \sim 1 + X$$

with binomially distributed Y, and the data set

i	N	Y	X
1.	25	10	6.0
2.	25	15	7.0

Here X_i is the pH of a test aquarium in which N_i fish are exposed, and Y_i denotes the number of survivors.

(i) Write down the likelihood function for the model parameter(s).

(ii) Find the maximum likelihood estimates of your parameters (analytically).

(iii) Evaluate and report the deviance residuals.

(iv) Does this data set offer any evidence against the suggested (logistic) model? How does the model fit compare, say, with a probit regression model? How much confidence would you have in predictions made from this model at, say, $X = 3$ or $X = 10$? Why?

6. (STA 244) Five replicate observations of a response variable y are made at each of three distinct values of a predictor variable x.

(i) When the linear regression model

$$y_{ij} = \beta_1 + \beta_2 x_i + d_{ij} \qquad i=1,2,3;\ j=1,2,3,4,5$$

is fit to the data by ordinary least squares, the following values are obtained:

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(\widehat{y}_{ij} - \overline{y})^2 = 16.8$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij} - \widehat{y}_{ij})^2 = 81.9$$

where $\overline{y} = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}$ and $\widehat{y}_{ij} = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$. Calculate the F statistic for a test of the hypothesis that $\beta_2 = 0$ in this model. Specify the degrees of freedom parameters for this F statistic.

(ii) When as an alternative the one-way ANOVA model

$$y_{ij} = \mu_i + d_{ij} \qquad i=1,2,3;\ j=1,2,3,4,5$$

is fit to the same data by least squares, we obtain

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(\widehat{y}_{ij} - \overline{y})^2 = 25.5$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij} - \widehat{y}_{ij})^2 = 73.2$$

where $\overline{y} = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}$ and $\widehat{y}_{ij} = \widehat{\mu}_i$. Calculate the F statistic for a test of the hypothesis that $\mu_1 = \mu_2 = \mu_3$ in this model. Specify the degrees of freedom parameters.

(iii) A third candidate model for this data is the quadratic regression model

$$y_{ij} = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + d_{ij} \qquad i=1,2,3;\ j=1,2,3,4,5.$$

Calculate the F statistic for a test of the hypothesis that $\beta_3 = 0$ in this model. Specify the degrees of freedom parameters.

7. (STA 290) A sociologist is interested in whether registered voters in Durham support a proposed ban on smoking in all public buildings. She has a list of all potential phone numbers of Durham residents, along with the knowledge that there are N households in Durham with at least one registered voter, and a total of N^* registered voters. To select the sample of respondents, she uses random digit dialing, with replacement, from this list. When a household is contacted, she asks for a list of registered voters in the household, randomly selects one of them, and asks whether this person supports the ban. The responses are recorded as support = 1 or does not support = 0.

(i) What assumptions do you need to make to ensure that the sociologist is studying the issue she wishes to study.

(ii) A standard estimator of the mean of a population proportion is the sample proportion. Create a small example of a population to demonstrate that this estimator may be a *biased* estimator of the population proportion (under the sociologist's sampling plan).

(iii) Provide an unbiased estimator of the population proportion for the sociologist's sampling plan. Show that your estimator is unbiased.

8. (STA 294) Suppose we want to estimate the total number (τ) of specific trees (say, *Mora Excelsa*) in a small geographical region of the closed evergreen forest of Trinidad, West Indies. Suppose that the study region in the forest is divided into eight segments. The information you may need about the population are summarized in the following table.

Segment	Proportion of total study area	Number of trees
1	0.10	53
2	0.20	76
3	0.14	87
4	0.15	62
5	0.13	14
6	0.08	52
7	0.05	90
8	0.15	30

Table 0.1: Number of trees from *Mora Excelsa* community by segment

Suppose that you draw 3 segments **with replacement** from the eight segments as described above by using *Probability Proportional to the Size* (PPS) scheme.

(i) How many different samples are possible (disregarding the orders they appear)?

(ii) What is the probability that segment number 4 is in your sample?

(iii) What is the probability that your sample (s) will consist only segments 3 and 8?

(iv) Derive an unbiased estimate of τ. Evaluate the estimate at $s = \{3, 8, 8\}$,

(v) Give an estimate of the variance of your estimate in (iv).

9. (STA 376) Consider the following Poisson process model with a change point: [1]

$$\begin{aligned}
k &\sim \text{discrete uniform on } 1,\ldots,n,\\
Y_i|k &\sim Poisson(\theta t_i),\quad i=1,\ldots,k,\\
Y_i|k &\sim Poisson(\lambda t_i),\quad i=k+1,\ldots,n\\
\theta &\sim G(a_1,b_1),\\
\lambda &\sim G(a_2,b_2),\\
b_1 &\sim IG(c_1,d_1),\\
b_2 &\sim IG(c_2,d_2).
\end{aligned}$$

(i) Describe a Markov chain Monte Carlo implementation to simulate a Monte Carlo sample from the posterior $p(\theta,\lambda,b_1,b_2,k \mid y_1,\ldots y_n)$. Use draws from the exact conditional posterior whenever practicable. State the explicit algorithm, specifying all involved distributions.

(ii) Implementing the Markov chain Monte Carlo from part (a), how would you determine when to terminate simulations? Describe any of the convergence diagnostics described in the literature (many answers are possible and acceptable). A conceptual description is fine, no computational details required.

---

[1]

$$X\sim Poisson(\lambda):\ P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda},\ x=0,1,2,\ldots,\quad \lambda>0.$$

$$X\sim G(\alpha,\beta):\ f(x|\alpha,\beta)=\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)},\quad x\geq 0,\alpha>0,\beta>0.$$

$$X\sim IG(\alpha,\beta):\ f(x|\alpha,\beta)=\frac{e^{-1/(x\beta)}}{\Gamma(\alpha)\beta^\alpha x^{\alpha+1}},\quad x\geq 0,\alpha>0,\beta>0.$$

10. (Misc) A simple version of signal detection models in observational astronomy hypothesizes Poisson arrival times for nuclear particles generated from astronomical sources, such signals being obscured by " background" noise. Specifically, assume the number of background radiated particles arriving at a detector is Poisson with the known rate b, so $X \sim \mathcal{P}(b)$. Further observations are made to investigate an hypothesis that there exists a source (e.g., a supernova) of particles; if such a source exists, the detected radiation from it will be weak and so assessment of whether or not it exists obscured by the background noise. So we suppose particles arrive at rate $b + \beta$ where (as before) b represents background rate and where β represents the rate for the source. Then $\beta = 0$ is consistent with no source, $\beta > 0$ consistent with source existence. Write $H = \{$ source exists $\}$, and suppose

(A) $X \sim \mathcal{P}(b + \beta)$ with density $p(x|\beta) = (b + \beta)^x e^{-(b+\beta)}/x!$, over $x = 0, 1, \ldots$;

(B) $P(H) = P(\beta > 0) = w$, the prior probability of source existence; and

(C) a prior density $p(\beta|H)$ over $0 < \beta$ for the source rate assuming the source exists.

We now observe $X = x$ particles in the unit of time.

(i) Show that $P(H|x)$ is determined in odds form by $O(H|x) = \frac{P(H|x)}{1-P(H|x)} = \frac{w}{(1-w)} r(x)$ where

$$r(x) = \int_0^\infty (1 + \beta/b)^x e^{-\beta} p(\beta|H) d\beta$$

(ii) Deduce that $x = 0$ always provides evidence against source existence, i.e. decreases the odds on H.

(iii) Show that, whatever the prior $p(\beta|H)$ is, $r(x) \leq (1 + \hat{\beta}/b)^x e^{-\hat{\beta}}$, where $\hat{\beta}$ is the MLE of β under H. *(Hint: a maximum always exceeds an average).*

(iv) Find $\hat{\beta}$ as a function of x and b. If $b = 0.73$ (a background rate in a recent study of supernova SN1987A), compute the upper bounds on $r(1)$ and $r(2)$. Comment on the potential evidence in favour of source existence if only one or two counts are recorded.

11. (Misc) There is often interest in regression models in which it is assumed that one regression function applies in a certain range of x and another applies in a different range. For example, a general segmented linear regression with iid $N(0, \sigma^2)$ errors is of the form

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \epsilon, \quad x \leq r \\ &= \gamma_0 + \gamma_1 x + \epsilon, \quad x > r. \end{aligned}$$

The regression line is continuous if $\beta_0 + \beta_1 r = \gamma_0 + \gamma_1 r$ and discontinuous otherwise. Assume that r is known.

Let $I(x)$ be the usual indicator function with value 0 if $x \leq r$ and 1 if $x > r$ and define the function $x_+ = \max(0, x)$. Consider fitting the regression model

$$Y = \delta_0 + \delta_1 x + \delta_2 I(x - r) + \delta_3 (x - r)_+ + \epsilon, \tag{0.1}$$

where the errors are again iid $N(0, \sigma^2)$.

(i) Give the relations between $\beta_0, \beta_1, \gamma_0, \gamma_1$, and $\delta_0, \delta_1, \delta_2, \delta_3$; and determine if the two models are equivalent.

(ii) Describe how fitting the model (0.1) may be used to test the null hypotheses

(a) of continuity of segmented regression, $\beta_0 + \beta_1 r = \gamma_0 + \gamma_1 r$

(b) of identity of the two segments $\beta_0 = \gamma_0, \beta_1 = \gamma_1$.

(iii) If the errors are iid, mean 0, but not necessarily normal, are the tests in (ii) still valid?