FIRST YEAR EXAM

Monday May 8, 2006; 9:30 - 12:30

NOTES: PLEASE READ CAREFULLY BEFORE BEGINNING EXAM!

1. Do not write solutions on the exam; please write your solutions on the paper provided.
2. Put the problem number and your assigned code on the top of **each page**.
3. Write only on **one side** of the page (solutions on the reverse side of the page will be ignored).
4. Start each problem on a new page.
5. It is to your advantage to show your work and explain your answers. Draw a line through work you do not want graded; you do not need to erase.
6. Exam credit is not uniformly distributed across questions, nor are questions necessarily expected to be of comparable difficulty. You should attempt all questions.
7. You have 3 hours to finish.
8. This is a closed book exam. No notes are permitted.
9. The Takehome Exam is due at 10 AM May 10th and should be handed in to Krista Moyle in Room 223 Old Chemistry.

1. (a) X and Y are random variables with joint density $f(x, y) = \frac{3}{16}(4 - 2x - y)$ for $x > 0$, $y > 0$, and $2x + y < 4$.

   i. Find the conditional pdf of Y given X.

   ii. Find $\Pr[Y \geq 2 \,|\, X = .5]$

   (b) X is a random variable with density $p(x) = x/2$ for $x \in (0, 2)$. Let $Y = X(2 - X)$. Find the pdf of Y.

2. We count the number X of successes in a fixed number $n \in \mathbb{N}$ of independent trials, each with probability $\theta \in \Theta = [0,1]$ of success, in an effort to learn about the hypotheses

$$H_0 : \ \theta \le \frac{1}{2} \qquad H_1 : \ \theta > \frac{1}{2}.$$

(a) Denote by $P = P(X)$ the p-value for the likelihood ratio test, which will reject H_0 at level α if $P(X) \le \alpha$. How large must n be for it to be *possible* to reject H_0 at level $\alpha = 0.01$? Why?

(b) Compute the value of $P = P(9)$ if we observe $X = 9$ successes in $n = 10$ tries. Would we accept or reject H_0 at level $\alpha = 0.01$?

(c) Find the posterior probability of H_0 in a Bayesian analysis with uniform prior distribution with density function $\pi(\theta) \equiv 1$, $\theta \in \Theta$, if we observe $X = 5$ successes in $n = 5$ tries.

(d) Find an exact 95% one-sided confidence interval $[L, 1]$ for θ, if we observe $X = 5$ successes in $n = 5$ tries. Give L to four decimal places, or give an exact expression for L. Show your work.

3. The random variable $X \sim \mathsf{Un}(0, 2)$ has a uniform distribution on the interval $[0, 2]$ while $Y \sim \mathsf{Ex}(1)$ has an exponential distribution, each with mean one; X and Y are independent. Let $Z \equiv (X \vee Y)$ be their maximum.

   (a) Find the probability $\mathsf{P}[X < Y]$.

   (b) Find the conditional probability $\mathsf{P}[X < Y \mid Z = 1.0]$.

   (c) Give the C.D.F. for *and* describe in words the conditional distribution of Z, given that $X = 1.0$.

4. The moment generating function (mgf) of a random variable X is

$$M_X(t) = \mathrm{E}(e^{tX})$$

(a) If X has density $f_X(x) = \lambda e^{-x\lambda}$ for $x > 0$, find $M_X(t)$.

(b) Show that if $X \sim \mathrm{Bern}(\theta)$ then $M_X(t) = \theta e^t + (1-\theta)$.

(c) Show that if $X \sim \mathrm{Bin}(n, \theta)$ then $M_X(t) = [\theta e^t + (1-\theta)]^n$. State the two theorems needed to derive this result from part (b).

(d) Show that if $Y \sim \mathrm{Poi}(\lambda)$ then $M_Y(t) = e^{\lambda(e^t-1)}$.

(e) For some positive constant c, let $\{X_n\}$ be a sequence of random variables with distributions $X_n \sim \mathrm{Bin}(n, c/n)$ and let $Y \sim \mathrm{Poi}(c)$. Show that $\lim_{n\to\infty} M_{X_n}(t) = M_Y(t)$.
Note : Recall the following results on e, 1) $e = \lim(1+1/n)^n$ and 2) $e^x = \sum_{i=1}^{\infty} x^i/i!$.

5. A physical experiment to detect subatomic particles can detect two types of particles: type A and type B. A new source of particles is to be monitored and it will generate particles of only one type; the scientists do not know which. It is known that type A particles are about 3 times as common as type B, and that the rate of flux per unit time of particles of type B is about 10 times that of type A.

   Let μ denote the rate of flux of type A particles; prior experiments have led to a current gamma prior for μ, namely $\mu \sim \mathsf{Ga}(a, am)$. In the new experiment, the sampling model for the arrival time t of the first particle is

   $$p(t|\mu) = 0.75\mu \exp(-\mu t) + c\mu \exp(-10\mu t), \qquad t > 0.$$

   NOTE: Recall that $x \sim \mathsf{Ga}(a, b)$ has mean equal a/b and has density function $p(x) = A \cdot x^{a-1} \exp(-bx)$ where $A = b^a/\Gamma(a)$ is the normalising constant.

   (a) What is the value of c?

   (b) Briefly describe the meaning of this sampling model and how it relates to the physical problem.

   (c) Show that the posterior density function $p(\mu|t)$ has the form

   $$p(\mu|t) = q(t)p_A(\mu|t) + (1 - q(t))p_B(\mu|t), \qquad \mu > 0,$$

   where $p_A(\mu|t)$ and $p_B(\mu|t)$ are two gamma densities, and $q(t)$ is some number depending on t, as well as (a, m).

   (d) How can you interpret $q(t)$?

   (e) What is $\mathsf{E}(\mu|t)$?

6. Let $\mathbf{Y}$ be a n dimensional Normal random vector, $\mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n)$, with mean $\mathsf{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ and covariance matrix that is the identity, where $\mathbf{X}$ is the $n \times p$ design matrix of rank p and $\boldsymbol{\beta}$ is the p dimensional vector of regression coefficients. Suppose that the prior for $\boldsymbol{\beta}$ is Zellner's g prior, $\boldsymbol{\beta} \sim \mathsf{N}(0, g(\mathbf{X}^T\mathbf{X})^{-1})$ leading to the posterior distribution for $\boldsymbol{\beta}|\mathbf{Y}$ that is $\mathsf{N}(g/(1+g)\hat{\boldsymbol{\beta}}, g/(1+g)(\mathbf{X}^T\mathbf{X})^{-1})$ where $\hat{\boldsymbol{\beta}}$ is the usual maximum likelihood estimate of $\boldsymbol{\beta}$.

   Let $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

   (a) Find the distribution of $\boldsymbol{\mu}$ given $\mathbf{Y}$.

   (b) As an estimator of $\boldsymbol{\mu}$, is $\mathsf{E}[\boldsymbol{\mu} \mid \mathbf{Y}]$ unbiased? (Explain/show)

   (c) Suppose that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, so that the $\boldsymbol{\beta}$ are uncorrelated given $\mathbf{Y}$. Are μ_j and μ_k independent for all j and k (given $\mathbf{Y}$)? (μ_i is the ith elemnet of the vector $\boldsymbol{\mu}$)

   (d) Find $\mathsf{E}[(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^T(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \mid \mathbf{Y}]$ where $\tilde{\boldsymbol{\mu}}$ is the posterior mean of $\boldsymbol{\mu}$.

First Year Exam - Takehome

Turn into Krista in Room 223 by 10 AM May 10, 2006

Monitoring programs to track long-term changes in population size are important for applied ecological studies. Such monitoring programs often have multiple objectives that include monitoring trends, estimating abundance, and estimating the effects of covariates. Your job is to analyze trend, abundance, and the effects of covariates using a dataset of counts of springbok antelope around 25 watering holes. If you think the dataset cannot answer all these questions, then say why.

The data come from aerial surveys. On each survey date, an airplane flies a route over the collection of sites (watering holes) at an altitude of 200–300 meters. Springbok are counted at each site. A survey normally includes counts at all 25 sites but occasionally some sites could not be counted because of poor weather. Each site was circled until the observer was confident that an accurate count had been made. For larger groups of springbok, color photographs were taken and springbok were counted later from the photos. Several surveys, usually 7–10, were made each year. You have data from 1990–2002 for sites 12–21, 23, and 24. The other sites were excluded because they usually don't have many antelope. Within a year, springbok are faithful to a single site; i.e., if a springbok goes to site i on one day, it will return to site i on other days.

Many studies have demonstrated effects of date and time-of-day on springbok counts; therefore, these covariates are included in the dataset. Whether they are relevant in this particular dataset remains unknown, until you tell us.

Conduct a statistical analysis of the Springbok data and summarize your findings in a typed two-page report. You may include a supplemental appendix of no more than five pages with any other key figures, output or more technical expressions to support your analysis that are not included in the main text. All figures and computer output should be clearly labelled and annotated. Any results in the appendix should be referenced in the body of the report.

The data are in a spreadsheet file `http://www.stat.duke.edu/programs/grad/fye/springbok.xls`. Headers and first 3 lines:

LOCNUMBER	SITEI	YEAR	DATE	HourFromNoon	COUNTS
12	1	1990	28	0.800	50
12	1	1990	29	0.117	41
12	1	1990	30	-0.167	43