Department of Statistical Science Duke University

FIRST YEAR EXAM - SPRING 2012

Monday, May 7th 2012

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.
   Do not erase anything– just draw a line through work you do not want graded.
6. You have 3 hours to finish the written exam: Questions 1-6 inclusive. Attempt all questions; note that credit is not necessarily equally allocated across questions.
7. This is a closed book exam. No notes are permitted.

1. (a) Show that if $\xi \sim N(0,1)$ then $P(|\xi| > z) \leq \exp(-z^2/2)$ for any $z > 0$.

   (b) Let $\{\xi_n\} \overset{\text{iid}}{\sim} N(0,1)$ be independent standard normal random variables. For what values of $c > 0$ (if any) is it true that only finitely-many of the events

   $$A_n = \{|\xi_n| > c\sqrt{\log n}\}$$

   will occur?

2. Emile Borel proposed the game *La Relance*. In this game two players (say Bart and Lisa) each ante a dollar to play, and then each draws, independently and privately, a random number from the $U[0,1]$ distribution. Bart goes first. He observes his draw $X = x$, and decides whether to fold (and lose his dollar to Lisa) or bet b dollars. Then Lisa observes her draw, $Y = y$, and decides whether to fold (and lose her dollar to Bart) or match Bart's bet. If neither player folds, then they compare their hands and the person with the largest number wins all of the money (i.e., $2 + 2b$ dollars).

   (a) Bart believes that Lisa bets if and only if $Y > c$ for some value c. His prior on c is $U(0,1)$. Show that Bart's expected *net gain* (amount won – amount put in, in dollars), if he bets after observing $X = x$, is $V_x = 1/2 + (1+b)(x^2 - 1/2)$.

   (b) Suppose $b = 5$. What are the values of $x \in (0,1)$ for which Bart's optimal decision rule is "*to bet*", so as to maximize his expected net gain?

   (c) What is Bart's expected net gain from the game (again with $b = 5$) if he plays by his optimal betting rule?

3. A set of n counts $X = (X_1, \cdots, X_n)$ are modeled as

$$\Pr(X_i = 0 \,|\, \gamma_i = 0, \lambda, \mu) = 1,\ X_i \,|\, (\gamma_i = 1, \lambda, \mu) \sim Poi(\mu),\ (\text{indep. across } i)$$

where $\gamma = (\gamma_1, \cdots, \gamma_n)$ – a set of n latent binary counts, $\lambda \in [0, 1]$ and $\mu > 0$ are assigned the hierarchical prior:

$$\gamma_i \,|\, (\lambda, \mu) \overset{\text{iid}}{\sim} Bern(\lambda),\ \lambda|\mu \sim Be(c\mu, 1),\ \mu \sim Ga(a, b)$$

for some positive constants a, b, c. [$Be(r, 1)$ has pdf $r\lambda^{r-1}$, $0 \leq \lambda \leq 1$]

(a) Show that the conditional prior pdf of μ given λ is $Ga(a + 1, b - c \log \lambda)$.

(b) Write down the conditional posterior distributions of $\lambda|(\gamma, \mu, x)$, $\mu|(\gamma, \lambda, x)$ and $\gamma|(\lambda, \mu, x)$ given data $x = (x_1, \cdots, x_n)$ on X. Answer in terms of named distributions with explicit formulas for their parameters and with appropriate use of conditional independence.

(c) Label the above model as M_1. Now consider the following *reduced* model with a fixed $\lambda = 1$:

$$M_0 : X_i \overset{\text{iid}}{\sim} Poi(\mu),\ \mu \sim Ga(a, b)$$

with same a, b as in M_1. Would the Bayes factor of M_0 to M_1 equal 1 if we had all $x_i > 0$? Justify your answer.

4. (a) Two scalar random quantities x, y have a joint distribution with complete conditionals $(x|y) \sim Ga(x|a, ay)$ and $(y|x) \sim Ga(y|a + b, ax + c)$ for some known, positive parameters a, b, c.
What are the margins $p(x)$ and $p(y)$? Show your reasoning.

   (b) A first-order Markov process on the real line has homogenous transition distribution with density $p(x|x')$ defined by

$$x \sim \begin{cases} N(x|\phi x', v), & \text{with probability } a, \\ N(x|0, s), & \text{with probability } 1 - a, \end{cases}$$

   where $|\phi| < 1$, $s = v/(1 - \phi^2)$ and for some probability a. Typical uses will have larger values of a.

      i. Give the expression for the transition p.d.f. $p(x|x')$.

      ii. Show that process is ergodic and identify the unique stationary distribution. *Note: Just quote any relevant theory/results you use here; you do not need to work the minute details through if you can just draw on standard theory.*

      iii. Is the process reversible? Either prove or disprove.

5. Consider the following one-way ANOVA model for $j = 1, \ldots, m$ groups with $i = 1, \ldots, n_j$ observations in group j

$$Y_{ij} = \alpha + \tau_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Suppose the prior on α is $N(0, \sigma_\alpha^2)$, and independently of α, the m dimensional vector $\boldsymbol{\tau} = (\tau_1, \ldots \tau_m)^T$ is assigned a $N(\mathbf{0}_m, \sigma_\tau^2(\mathbf{I}_m - \mathbf{P}_m))$ prior distribution, where $\mathbf{I}_m$ is the identity matrix of dimension m and $\mathbf{P}_m = \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$ is the orthogonal projection matrix onto the linear space spanned by $\mathbf{1}_m$, the vector of ones of length m. The following questions relate only to the prior distribution.

(a) What is the distribution of $\sum_{j=1}^m \tau_j$?

(b) Find the distribution of $\boldsymbol{\mu} = \alpha\mathbf{1}_m + \boldsymbol{\tau}$ and express $\mathsf{Cov}(\boldsymbol{\mu})$ in the form $a\mathbf{I}_m + b\mathbf{P}_m$; give expressions for a and b. What is the correlation between μ_j and $\mu_{j'}$? [Express in terms of σ_α^2, σ_τ^2 and any other quantities.]

(c) What are the conditions (if any) on σ_α and σ_τ for the covariance of $\boldsymbol{\mu}$ to be positive definite? [Hint: find the eigenvalues using properties of projection matrices.]

(d) Split $\boldsymbol{\tau} = (\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)})^T$ where $\boldsymbol{\tau}^{(1)} = (\tau_1, \cdots, \tau_r)^T$ and $\boldsymbol{\tau}^{(2)} = (\tau_{r+1}, \cdots, \tau_m)^T$ with r an integer between 1 and m. What is the conditional distribution of $\boldsymbol{\tau}^{(1)}$ given $\boldsymbol{\tau}^{(2)}$? [Give name and simplified expressions for parameters. You may find the following identity useful: $(\mathbf{I}_k - \frac{k}{m}\mathbf{P}_k)^{-1} = \frac{m}{r}\mathbf{I}_k - \frac{k}{r}(\mathbf{I}_k - \mathbf{P}_k)$, with $k = m - r$.]

6. (a) Let $Y = \sigma(\rho\,|U| + \sqrt{1-\rho^2}\,V)$ where U, V are independent $N(0,1)$ variables and $\rho \in (-1,1), \sigma \in (0,\infty)$ are constants. Demonstrate that the pdf of Y is given by

$$f(y|\rho,\sigma) = \frac{a}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \Phi\left(\frac{\rho y}{b\sigma\sqrt{1-\rho^2}}\right), \quad -\infty < y < \infty$$

for some positive constants a, b. Identify the numerical values of a and b. In above, $\Phi(z)$ denotes the standard normal CDF.

(b) Consider n observations $Y_1, \cdots, Y_n$, modeled as $Y_i \overset{\text{iid}}{\sim} f(y_i|\rho,\sigma), \rho \in (-1,1), \sigma \in (0,\infty)$. Does a maximum likelihood estimate of $(\rho,\sigma) \in (-1,1) \times (0,\infty)$ always exist? Justify.

(c) For the same model, suppose we are interested in testing $H_0 : \rho = 0$ (against $H_1 : \rho \neq 0$). What is the value of c such that the test that rejects H_0 if

$$\frac{\sqrt{n}\,|\bar{Y}|}{s_Y} > c$$

has size 5%? Identify c as a specific quantile of a named distribution.

(d) For $n = 100$, the power of the above 5% test at $(\rho,\sigma) = (0.2, 1)$ is calculated to be 0.36. What can you say about the power of this test at $(\rho,\sigma) = (0.2, 5)$? Justify.

FYE 2012
May 07-09 2012

Department of Statistical Science
Duke University

Take Home Data Analysis Problem

In many developing countries, people get their drinking water from wells. Sometimes these wells are contaminated with the chemical arsenic, which when consumed in high concentrations causes skin and bladder cancer, as well as cardiovascular disease. Fortunately, in many cases people living near contaminated wells have the opportunity to get water from nearby uncontaminated wells.

In one study, several researchers measured the concentrations of arsenic in wells in a particular region of a developing country. They labeled wells as safe or unsafe based on the measurements. The researchers encouraged people drinking from unsafe wells to switch to safe wells. Several years later, the researchers returned to the area with the goal of seeing who had switched from unsafe to safe wells. They recorded the following information on a sample of 3020 individuals who had wells at their homes that were unsafe.

Variable	Description
switch	=1 if respondent switched to a safe well, =0 if still using own unsafe well
arsenic	amount of arsenic in well at respondent's home
educ	years of schooling of the head of household
dist	distance in meters to the nearest known safe well
assoc	=1 if any members of household are active in community organizations, =0 otherwise

These variables were collected because, *a priori*, it is reasonable to hypothesize that people living far from the nearest safe well are less likely to switch; that people with high levels of arsenic are more likely to switch; and that people with high education and community involvement are more likely to switch. It is also reasonable to think that some variables will moderate the effects of others.

Present a three page (maximum) report addressing the question: what predicts why people switch wells? Your report should discuss all relevant aspects of your analysis (exploratory and modeling) with graphical and numerical summaries that are important for communicating results. The report should be written so that policymakers could understand and apply the findings. While you may include code and other plots in a supplemental appendix, you should not assume that graders will read beyond the main report; all relevant material should be within the three page limit.

Go to `http://www.stat.duke.edu/~st118/fye12takehome.txt` to access the data file in tab delimited format.

Submit your report electronically to Karen Herndon by email (`karen@stat.duke.edu`). Your report file should be named `fye12_codename.pdf`. Your report should not contain your name or any other identifier. It MUST include your assigned code name.

Distribution	Notation	$f(x) =$ **pdf (pmf)**	Support	Mean	Variance	
Beta	$Be(a,b)$	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$x \in (0,1)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	
Bernoulli	$Bern(p)$	$f(x) = p^x q^{(1-x)}$	$x \in \{0,1\}$	p	$p\,q$	$(q = 1-p)$
Binomial	$Bin(n,p)$	$f(x) = \binom{n}{x} p^x q^{(n-x)}$	$x \in \{0,\cdots,n\}$	$n\,p$	$n\,p\,q$	$(q = 1-p)$
Chi-square	$\chi^2(\nu)$	$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}\, e^{-x/2}$	$x \in \mathbb{R}_+$	ν	2ν	
Exponential	$Ex(\lambda)$	$f(x) = \lambda\, e^{-\lambda x}$	$x \in \mathbb{R}_+$	$1/\lambda$	$1/\lambda^2$	
Gamma	$Ga(\nu,\lambda)$	$f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1}\, e^{-\lambda x}$	$x \in \mathbb{R}_+$	ν/λ	ν/λ^2	
Geometric	$Geo(p)$	$f(x) = p\, q^x$	$x \in \mathbb{Z}_+$	q/p	q/p^2	$(q = 1-p)$
		$f(y) = p\, q^{y-1}$	$y \in \{1,\ldots\}$	$1/p$	q/p^2	$(y = x+1)$
HyperGeo.	$HG(n,M,N)$	$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$x \in 0,\cdots,n$	$n\,p$	$n\,p\,(1-p)\frac{N-n}{N-1}$	$(p = \frac{M}{N})$
Logistic	$Lo(\mu,\beta)$	$f(x) = \frac{e^{-(x-\mu)/\beta}}{\beta[1+e^{-(x-\mu)/\beta}]^2}$	$x \in \mathbb{R}$	μ	$\pi^2\beta^2/3$	
Log Normal	$LN(\mu,\sigma^2)$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\log x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}_+$	$e^{\mu+\sigma^2/2}$	$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$	
Neg. Binom.	$NB(\alpha,p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha\, q^x$	$x \in \mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$	$(q = 1-p)$
		$f(y) = \binom{y-1}{y-\alpha} p^\alpha\, q^{y-\alpha}$	$y \in \{\alpha,\ldots\}$	α/p	$\alpha q/p^2$	$(y = x+\alpha)$
Normal	$N(\mu,\sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	$x \in \mathbb{R}$	μ	σ^2	
Pareto	$Pa(\alpha,\epsilon)$	$f(x) = \alpha\,\epsilon^\alpha / x^{\alpha+1}$	$x \in (\epsilon,\infty)$	$\frac{\epsilon\,\alpha}{\alpha-1}$	$\frac{\epsilon^2\alpha}{(\alpha-1)^2(\alpha-2)}$	
Poisson	$Poi(\lambda)$	$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$x \in \mathbb{Z}_+$	λ	λ	
Snedecor F	$F(\nu_1,\nu_2)$	$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \times x^{\frac{\nu_1-2}{2}} \left[1+\frac{\nu_1}{\nu_2}\,x\right]^{-\frac{\nu_1+\nu_2}{2}}$	$x \in \mathbb{R}_+$	$\frac{\nu_2}{\nu_2-2}$	$\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}$	
Student t	$t(\nu)$	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}[1+x^2/\nu]^{-(\nu+1)/2}$	$x \in \mathbb{R}$	0	$\nu/(\nu-2)$	
Uniform	$U(a,b)$	$f(x) = \frac{1}{b-a}$	$x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Weibull	$Wei(\alpha,\beta)$	$f(x) = \alpha\beta\, x^{\alpha-1}\, e^{-\beta\, x^\alpha}$	$x \in \mathbb{R}_+$	$\frac{\Gamma(1+\alpha^{-1})}{\beta^{1/\alpha}}$	$\frac{\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)}{\beta^{2/\alpha}}$	