Ph. D. Qualifying Exam
Friday, August 22, 2008

Please submit solutions to at most **seven** problems. You have four hours. No one is expected to answer all the problems correctly. Partial credit will be given. All problems are worth an equal amount of credit.

Put your solution to each problem on a separate sheet of paper.

Applied Statistics

Problem 1. Consider the following unbalanced random one-way model:

$$Y_{ij} \;=\; \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \dots, k; \;\; j = 1, \dots, n_i,$$

where $\{\alpha_i, i = 1, \dots, k\}$ are iid $N(0, \sigma_\alpha^2)$, $\{\epsilon_{ij}, i = 1, \dots, k; j = 1, \dots, n_i\}$ are iid $N(0, \sigma_\epsilon^2)$, and the α_i's and ϵ_{ij}'s are independent. Define

$$\text{SSA} = \sum_{i=1}^k n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^k \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{\sum_{i=1}^k n_i},$$

$$\text{SSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2,$$

where

$$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n_i}; \qquad y_{\cdot\cdot} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{\sum_{i=1}^k n_i}.$$

(a) Find the expectations and variances of $y_{i\cdot}$ and $y_{\cdot\cdot}$. What are the distributions of $y_{i\cdot}$ and $y_{\cdot\cdot}$?

(b) Find the expectation of SSA.

(c) What is the distribution of SSE? Find the expectation and variance of SSE.

Problem 2. Consider the linear regression model:

$$Y = X\beta + \xi,$$

where $Y = (y_1, \dots, y_n)'$, $\xi = (\xi_1, \dots, \xi_n)'$, $\beta = (\beta_1, \dots, \beta_p)'$, and X is an $n \times p$ full-rank matrix. $\{\xi_i, i = 1, \dots, n\}$ are assumed to be iid $N(0, \sigma^2)$ variables. The least squares estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-p}\left[Y'(I - H)Y\right],$$

where H is the projection matrix $H = X(X'X)^{-1}X'$.

(a) Suppose that $A = (a_{ij})$ and $B = (b_{ij})$ are two $n \times n$ matrices. Show that trace$(AB) =$ trace(BA) where trace$(A) = \sum_{i=1}^{n} a_{ii}$.

(b) Let $Z = (Z_1, \ldots, Z_n)'$ be a random vector with $\mathrm{E}(Z) = \boldsymbol{\mu}$ and $\mathrm{Cov}(Z) = V$. Define $Q = Z'AZ$. Using the result in (a), show that

$$\mathrm{E}(Q) = \mathrm{trace}(AV) + \boldsymbol{\mu}' A \boldsymbol{\mu}.$$

(c) Using the result in (b), find $\mathrm{E}(\hat{\sigma}^2)$. Is $\hat{\sigma}^2$ an unbiased estimate of σ^2?

Problem 3. In a generalized linear model (GLM), the response variable Y is assumed to have a density function with the exponential dispersion form:

$$f(y; \theta, \phi) \quad = \quad \exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}.$$

Consider the gamma density

$$f(y; \lambda, \beta) = \begin{cases} \frac{\beta^\lambda}{\Gamma(\lambda)} y^{\lambda - 1} e^{-\beta y}, & y > 0, \\ 0, & y \le 0, \end{cases}$$

where $\beta > 0$ and $\lambda > 0$.

(a) Write the gamma density in the exponential dispersion form. Identify θ, $b(\theta)$, $a(\phi)$, and $c(y, \phi)$ in terms of β and λ.

(b) Let $l(\theta, y) = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$. Derive the mean and variance of Y (in the general case) from the relations $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ and $E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0$. What are the mean and variance when Y has the gamma density?

Problem 4.

(a) For the following autoregressive (AR) process in which $\{a_t\}$ is a white noise process, check the stationarity condition. Then calculate the autocorrelation ρ_k for $k = 0, 1, 2, 3$ (hint: one root of the following polynomial is -2):

$$X_t + 2.1X_{t-1} + 1.6X_{t-2} + 0.4X_{t-3} = a_t.$$

(b) For the following moving-average (MA) process in which $\{a_t\}$ is a white noise process, check the invertibility condition. Then calculate the autocorrelation ρ_k for $k = 0, 1, 2, 3$:

$$X_t = 1 + a_t + 1.5a_{t-1} + 0.6a_{t-2}.$$

Probability

Problem 5.

(a) Let $\mathcal{A}$ be a collection of sets on a space Ω. Show that $\mathcal{A}$ is a σ-field if and only if $\mathcal{A}$ is a π/λ system. (Carefully state the definitions of a σ-field and λ-systems.)

(b) Let P and Q be two probability measures on $(\mathbb{R}, \mathcal{B})$ with

$$P(-\infty, x] = Q(-\infty, x] \quad \text{for all } x \in \mathbb{R}.$$

Show that $P(B) = Q(B)$ for all $B \in \mathcal{B}$, the collection of all Borel sets on $\mathbb{R}$. (Carefully state what result on π/λ systems you use.)

Problem 6.

(a) Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i| < \infty$, $i = 1, 2, \cdots$. Show that

1. the sequences X_i and $Y_i := X_i 1\{|X_i| \le i\}$, $i = 1, 2, \dots$ are tail equivalent and
2. $\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}Y_i \right| \to 0 \quad \text{as } n \to \infty.$

(b) Let S_n be a Binomial random variable with parameters n and p_n and $\mathrm{Var}(S_n) \to \infty$ as $n \to \infty$. Show that

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}(S_n)}}$$

converges in distribution to a standard normal random variable as $n \to \infty$ by writing S_n as a sum of independent random variables and verifying Lyapounov's condition.

Problem 7. Let $X_1, X_2, \dots$ be i.i.d. with common distribution $N(1,1)$. Let $V_n = \frac{1}{n}\sum_1^n X_i$ and $M_n =$ median of $\{X_1, \dots, X_n\}$ be the mean and median based on the first n random variables.

(a) Write down the limiting distributions of $\sqrt{n}(V_n - 1)$ and $\sqrt{n}(M_n - 1)$.

(b) Let

$$T_n = \begin{cases} V_n & \text{if } |V_n| \le n^{-\frac{1}{4}} \\ M_n & \text{if } |V_n| > n^{-\frac{1}{4}} \end{cases}$$

in other words,

$$T_n = V_n I(|V_n| \le n^{-\frac{1}{4}}) + M_n I(|V_n| > n^{-\frac{1}{4}}).$$

Obtain the limiting distribution of $\sqrt{n}(T_n - 1)$. *Hint: First compute the limit of* $E(I(|V_n| \le n^{-\frac{1}{4}})) = P(|V_n| \le n^{-\frac{1}{4}})$.

Problem 8. A clumsy climber is on a ladder with 6 rungs. At any time, he is either on the ground (state 0) or on one of the 6 rungs (states 1 to 6). Let i denote his current state, and suppose that his state at the next time is determined by the following rule: If $1 \le i \le 5$, he either steps up one rung, stays where he is, steps down one rung, or falls to the ground, with probabilities .4, .3, .2, .1, respectively. If $i = 0$, he steps up one rung (with probability 1), and if $i = 6$, he steps down one rung (with probability 1). Note that, if $i = 1$, he ends up on the ground (either by "stepping" or by "falling") with probability .3. Let X_n denote the climber's state at time n.

(a) Write down the transition probability matrix for this Markov chain.

In the remaining parts, write out matrix expressions for your answers. Any matrices appearing in your answers should be written out explicitly.

(b) Suppose the climber is currently on rung i where $i \ge 1$. What is his expected number of visits to rung j (where $j \ge 1$) before visiting the ground (state 0)?

(c) Let τ be the first time (after time zero) that the climber reaches either the ground (state 0) or the top of the ladder (state 6). That is, $\tau = \inf\{n \ge 1 : X_n \in \{0, 6\}\}$. Assuming $1 \le i, j \le 5$ and $m > 0$, evaluate the following:

$$E_i \sum_{k=0}^{m} I(X_k = j, \tau > k) \qquad \text{(Note that the sum is only up to } m.\text{)}$$

(d) If the climber starts from the ground (state 0), what is the probability he will reach the top of the ladder (state 6) before ever returning to the ground?

Problem 9. Suppose we observe $X_1, \ldots, X_n$ iid from the density $f(x \,|\, \theta) = \frac{1}{2}e^{-|x-\theta|}$ where θ is unknown.

(a) Find a minimal sufficient statistic for θ.

(b) Is the statistic you found above complete?

(c) What would be your answers to (a) and (b) if instead $f(x \,|\, \theta) = \frac{1}{2\beta}e^{-|x-\alpha|/\beta}$ where $\theta = (\alpha, \beta)$ is unknown?

Problem 10. Suppose that we have two independent random samples: $X_1, \ldots, X_n$ are exponential(θ) and $Y_1, \ldots, Y_m$ are exponential(μ).

(a) Find the likelihood ratio test (LRT) of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$.

(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum X_i}{\sum X_i + \sum Y_i} \,.$$

(c) Find the distribution of T when H_0 is true.

Problem 11.

For the exponential family, the density function may be written:

$$f(y, \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\} \tag{1}$$

(a) Find $\dfrac{\partial}{\partial \theta} f(y, \theta)$.

(b) Use this result to find $E(a(Y))$ where Y has density $f(y, \theta)$.

(c) Find the score function in terms of the functions $a(y), b(\theta), c(\theta), d(y)$ and their derivatives.

(d) Show that the expected value of the score function is 0.

Problem 12. The data matrix $\mathbf{x}$ for a random sample of size $n = 4$ from a bivariate normal distribution is given by $\mathbf{x} = \begin{pmatrix} 2 & 1 \\ 2 & 3 \\ 1 & 2 \\ 3 & 2 \end{pmatrix}$.

(a) Describe and sketch a 95% confidence region for the mean vector μ of this distribution. (Use the attached table.)

(b) Test $\mu = (3.5 \;\; 2.5)^T$ at level $\alpha = 0.05$.

Problem 13. For $i = 1, \ldots, n$ exchangeable subjects/observations, the CPO (conditional predictive ordinate) of the observation/subject i is given by the cross-validated probability:

$$CPO_i = E_\theta \left[f_i(y_{i,obs} \,|\, \theta) \,|\, \tilde{y}_{-i,obs} \right] = \int_\Theta f_i(y_{i,obs} \,|\, \theta) p(\theta \,|\, \tilde{y}_{-i,obs}) \, d\theta$$

where $\tilde{y}_{-i,obs}$ is the observed data vector minus the i-th observation, $f_i(y_{i,obs} \,|\, \theta)$ is the sampling density of observation i evaluated at the observed value $y_{i,obs}$, and $p(\theta \,|\, \tilde{y}_{obs}) \propto \pi(\theta) \prod_{j=1}^n f_j(y_{j,obs} \,|\, \theta)$ is the posterior density given the observed data $\tilde{y}_{obs}$. An important measure of influence of the i-th observation on the posterior of θ is the K-L distance

$$KL(p, p_{-i}) = \int_\Theta \left[\log \left\{ \frac{p(\theta \,|\, \tilde{y}_{obs})}{p(\theta \,|\, \tilde{y}_{-i,obs})} \right\} \right] p(\theta \,|\, \tilde{y}_{obs}) \, d\theta \,.$$

(a) Show that

$$(CPO_i)^{-1} = E_\theta \left[\{ f_i(y_{i,obs} \,|\, \theta) \}^{-1} \,|\, \tilde{y}_{obs} \right] = \int_\Theta \{ f_i(y_{i,obs} \,|\, \theta) \}^{-1} p(\theta \,|\, \tilde{y}_{obs}) \, d\theta \,.$$

(b) How can you approximate CPO_i if you have N (large number) of samples from the full posterior $p(\theta \,|\, \tilde{y}_{obs})$?

(c) Show that

$$KL(p, p_{-i}) = \log \left[\int_\Theta \{ f_i(y_{i,obs} \,|\, \theta) \}^{-1} p(\theta \,|\, \tilde{y}_{obs}) \, d\theta \right] + \int_\Theta \left[\log \{ f_i(y_{i,obs} \,|\, \theta) \} \right] p(\theta \,|\, \tilde{y}_{obs}) \, d\theta$$

(d) What is the relationship between CPO_i and $KL(p, p_{-i})$?

Problem 14. Assume that one has a random sample of n observations $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$ where

$$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \text{and} \quad y_i = \begin{cases} 0 \text{ if the observation is not an event} \\ 1 \text{ if the observation is an event} \end{cases}$$

Suppose you fit a logistic model to the data and derive $\hat{\boldsymbol{\beta}}$, the maximum likelihood estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$ (the true vector of parameters for the logistic model). Let

$$\hat{p_i} = \frac{1}{1 + \exp\{-\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i\}}$$

denote the estimated probability that the i^{th} individual develops disease. Show that in this case, the following must be true:

$$\sum_{i=1}^{n} (y_i - \hat{p_i}) = 0$$

That is, the expected number of cases must equal the observed number of cases.

Problem 15. Assume that an investigator wants to do a clinical trial testing the difference between two normal means. She is willing to assume:

- σ^2 is the variance, assumed equal for the two groups.
- $\Delta = \mu_1 - \mu_2$ is the minimum difference that she thinks is clinically important.
- She is willing to accept a type I error of α.

Because of cost considerations, the total sample size $n = n_1 + n_2$ is fixed. (Here n_1 and n_2 are the number to be randomized to the first and second group, respectively.)

What are the values of n_1 and n_2 that maximize the power of the study?

Problem 16.

(a) When considering a sample with k age groups the data are:

Age Group	# at Risk	# Events	Rate= $\frac{\text{\# Events}}{\text{\# at Risk}}$
1	n_1	d_1	r_1
2	n_2	d_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_k	d_k	r_k
Total	n	d	r

Show that the crude rate, r, can be decomposed as:

$$r = \sum_{i=1}^{k} p_i r_i$$

Where $p_i = n_i/n$ is the proportion of people in the i^{th} age-group.

(b) For two sample data, the data may be viewed as:

Age Group	# at Risk	# Events	Rate= $\frac{\text{\# Events}}{\text{\# at Risk}}$
Sample 1			
1	n_{11}	d_{11}	r_{11}
2	n_{21}	d_{21}	r_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_{k1}	d_{k1}	r_{k1}
Total	n_1	d_1	r_1
Sample 2			
1	n_{12}	d_{12}	r_{12}
2	n_{22}	d_{22}	r_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_{k2}	d_{k2}	r_{k2}
Total	n_2	d_2	r_2

Decompose the difference, $r_1 - r_2$ into two sums, the first involving the difference in age specific rates, $r_{i1} - r_{i2}$, and the second involving differences in the proportion of individuals in the different age groups, $p_{i1} - p_{i2}$

(c) Comment on what this means and suggest a way how we might deal with this.

Problem 17. Let the survival time $T > 0$ be an integer-valued (discrete) random variable with finite mean residual lifetime $\mu_k = E[T - k \,|\, T > k]$ and discrete hazard rate $h_k = P[T = k \,|\, T \geq k]$ for $k = 0, 1, 2, \ldots$

(a) Show that $\mu_0 - (1 - h_1)\mu_1 = 1$.

Hint: One approach is to write down the terms in the sums for μ_0 and μ_1 in two rows, multiply each term of μ_1 by $(1 - h_1)$, and subtract the rows.

(b) Can you generalize the result of (a) to $\mu_k - (1 - h_{k+1})\mu_{k+1}$?

(c) Use the results of (a) and (b) to show that the Geometric distribution is the only discrete distribution with constant mean residual lifetime.

(d) Show that the exponential distribution is the only continuous distribution for which the mean residual lifetime $r(t)$ is constant for all $t > 0$.

Computational Statistics

Problem 18. Let x and y be two continuous random variables such that the marginal probability density of x is $N(a_0, \sigma_0^2)$ and the conditional probability density of y given x is $N(bx, \sigma_1^2)$.

(a) We know that the posterior density $P(x|y)$ is also Gaussian. Derive an expression for its mean.

(b) We can use this result to form the "update step" in the Kalman filter as follows. Let the predictive density of a process x_t, given all the previous observations $(y_1, y_2, \ldots, y_{t-1})$, be Gaussian with mean $\mu_{t|t-1}$ and variance $\sigma^2_{t|t-1}$. What is the mean of the posterior density of x_t given the updated data $(y_1, y_2, \ldots, y_{t-1}, y_t)$? We can assume that the conditional density of y_t, given x_t, is Gaussian with mean bx_t and variance ν^2.

Assume all quantities to be scalar in this problem.

Problem 19. Consider a multinomial distribution parameterized by θ according to:

$$(x_1, x_2, x_3, x_4) \sim \mathcal{M}(n; 0.5 + 0.25\theta, 0.25(1-\theta), 0.25(1-\theta), 0.25\theta) \ .$$

Our goal is to develop an EM algorithm for estimating θ from the given values of $x_1, \ldots, x_4$.

(a) Define two new random variables y_1 and y_2 such that $x_1 = y_1 + y_2$, where $(y_1, y_2) \sim \mathcal{M}(x_1; 0.5/T, 0.25\theta/T)$ with $T = 0.5 + 0.25\theta$. Set the complete data to be $y = (y_1, y_2, x_2, x_3, x_4)$. The vector y has multinomial distribution according to:

$$y \sim \mathcal{M}(n; 0.5, 0.25\theta, 0.25(1-\theta), 0.25(1-\theta), 0.25\theta) \ .$$

What is the log-likelihood of the complete data?

(b) Here we will derive an EM algorithm to estimate θ.

1. **E-Step**: First, what is the expected value of the log-likelihood of the complete data with respect to the density function $f(y_1, y_2|\theta_m, x_1, x_2, x_3, x_4)$? Call that function: $Q(\theta|\theta_m, y)$. (Drop terms that do not depend on θ)
2. **M-Step**: Next, solve for

$$\theta_{m+1} = \operatorname*{argmax}_{\theta} Q(\theta|\theta_m, y).$$

(Hint 1: For a multinomial random variable $x \equiv (x_1, x_2, \ldots, x_k) \sim \mathcal{M}(n; p_1, p_2, \ldots, p_k))$, the likelihood function is given by:

$$f(x|p_1, p_2, \ldots, p_k) \propto p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \ .$$

(Hint 2: You can use the fact that $E[y_2|\theta_m, x_1, x_2, x_3, x_4] = \frac{\theta_m}{2+\theta_m} x_1$.)