Department of Statistics
University of Florida

PhD Qualifying Exam

January 3, 2005

Instructions:

1. You have exactly four hours to answer questions in this examination.
2. There are 8 problems of which you must answer 6.
3. Only your first 6 problems will be graded.
4. Write only on one side of the paper, and start each question on a new page.
5. Clearly label each part of each question with the question number and the part, e.g., **1(a)**.
6. Write your **number** on every page.
7. Do not write your name anywhere on your exam.
8. You must show your work to receive credit.
9. While the eight questions are equally weighted, within a given question, the parts may have different weights.

1. Let $X_1, \cdots, X_p$ be $p(\geq 3)$ independent $\mathrm{N}(\theta_i, \sigma^2)$ random variables, where both $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T \in R^p$, and $\sigma^2(> 0)$ are unknown. We write $\boldsymbol{X} = (X_1, \cdots, X_p)^T$, and $\boldsymbol{x} = (x_1, \cdots, x_p)^T$. Let $h(\cdot)$ be a real-valued function satisfying $E_{\boldsymbol{\theta},\sigma^2}|\partial^2 \log h(\boldsymbol{X})/\partial X_i^2| < \infty$ and $E_{\boldsymbol{\theta},\sigma^2}[\partial \log h(\boldsymbol{X})/\partial X_i]^2 < \infty$ for all $i = 1, \cdots, p$. For estimating $\boldsymbol{\theta}$ by $\boldsymbol{a}$, suppose that the loss incurred in is given by $L_{\sigma^2}(\boldsymbol{\theta}, \boldsymbol{a}) = ||\boldsymbol{\theta} - \boldsymbol{a}||^2/\sigma^2$. Also, let U be a random variable distributed independently of the X_i's such that $U \sim \sigma^2\chi^2_m/(m + 2)$.

(a) Show that $\boldsymbol{T} = (X_1 + U(\partial \log h(\boldsymbol{X})/\partial X_1), \cdots, X_p + U(\partial \log h(\boldsymbol{X})/\partial X_p))^T$ improves on $\boldsymbol{X}$ for estimating $\boldsymbol{\theta}$ if

$$2\sum_{i=1}^{p} \frac{\partial^2 \log h(\boldsymbol{x})}{\partial x_i^2} + \sum_{i=1}^{p}\left(\frac{\partial \log h(\boldsymbol{x})}{\partial x_i}\right)^2 < 0$$

for almost all $\boldsymbol{x} \in R^p$.

(b) Take $h(\boldsymbol{x}) = ||\boldsymbol{x}||^{-(p-2)}$. Show that with this choice of h, the estimator $\boldsymbol{T}$ given in (a) improves on $\boldsymbol{X}$ for estimating $\boldsymbol{\theta}$.

(c) Find an unbiased estimator of the risk improvement given in (b).

2. Let $X_1, \cdots, X_n$ denote a random sample from an unknown (not necessarily normal) distribution with finite fourth moment. Let μ and σ^2 denote the mean and variance of this distribution. Also, let $\sqrt{\beta_1} = \mu_3/\sigma^3$ and $\beta_2 = \mu_4/\sigma^4$, where μ_3 and μ_4 denote the third and fourth central moments of this distribution. Suppose that one mistakenly assumes that the random sample is generated from a $\mathrm{N}(\mu, \sigma^2)$ distribution. In the following, one assumes the normal likelihood, but expectation is taken with respect to the true distribution which is not necessarily normal.

(a) Show that

$$\begin{bmatrix} E\left(\frac{\partial \log L}{\partial \mu}\right)^2 & E\left(\frac{\partial \log L}{\partial \mu}\right)\left(\frac{\partial \log L}{\partial \sigma^2}\right) \\ E\left(\frac{\partial \log L}{\partial \mu}\right)\left(\frac{\partial \log L}{\partial \sigma^2}\right) & E\left(\frac{\partial \log L}{\partial \sigma^2}\right)^2 \end{bmatrix} = n\begin{bmatrix} \frac{1}{\sigma^2} & \frac{\sqrt{\beta_1}}{2\sigma^3} \\ \frac{\sqrt{\beta_1}}{2\sigma^3} & \frac{\beta_2 - 1}{4\sigma^4} \end{bmatrix}.$$

(b) Show that

$$\begin{bmatrix} E\left(-\frac{\partial^2 \log L}{\partial \mu^2}\right) & E\left(-\frac{\partial^2 \log L}{\partial \mu \partial \sigma^2}\right) \\ E\left(-\frac{\partial^2 \log L}{\partial \mu \partial \sigma^2}\right) & E\left(-\frac{\partial^2 \log L}{\partial \sigma^4}\right) \end{bmatrix} = n\mathrm{Diag}\left(\frac{1}{\sigma^2}, \frac{1}{2\sigma^4}\right).$$

(c) Find the asymptotic distribution of the MLE of (μ, σ^2) under the true distribution.

(d) Comment on the results obtained.

3. Let $\{X_n, n \geq 1\}$ be a sequence of independent mean 0 square integrable random variables. Set $S_n = \sum_{j=1}^n X_j$ and $s_n^2 = \sum_{j=1}^n EX_j^2, n \geq 1$.

Prove that if

(1) $0 < s_n^2 \to \infty$,

(2) $\dfrac{X_n}{s_n} \to 0$ a.c., and

(3) $\displaystyle\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{[|X_j|>\varepsilon s_j]} X_j^2 \, dP = 0$ for some $\varepsilon > 0$,

then $\frac{S_n}{s_n} \to N(0,1)$.

4. Let $\{X_n, n \geq 1\}$ be a sequence of independent and identically distributed random variables and set $S_n = \sum_{j=1}^n X_j, n \geq 1$.

(a) Suppose that EX_1 exists. Prove that if $\frac{S_n}{n} \xrightarrow{P} c$ for some constant c in $(-\infty, \infty)$, then $\frac{S_n}{n} \to c$ a.c. and $c = EX_1$.

Warning. You cannot assume that X_1 is integrable; this is in effect part of the conclusion.

(b) Suppose that EX_1 does not exist. Identify the almost certain value of

$$\overline{\lim_{n\to\infty}} \frac{|S_n|}{n}.$$

Prove your answer.

(c) Demonstrate by an example that if EX_1 does not exist, then $\frac{S_n}{n} \xrightarrow{P} c$ for some constant c in $(-\infty, \infty)$ can hold. Justify your example.

5. Let $\boldsymbol{X} = (X_1, X_2)'$ be a random vector with two elements. Its density function is of the form

$$f(x_1, x_2) = \omega_1 f_1(x_1, x_2) + \omega_2 f_2(x_1, x_2)$$

where $\omega_1 + \omega_2 = 1$, $\omega_i \geq 0$, $i = 1, 2$, and f_i is the density function for the bivariate normal $N(\mathbf{0}, \mathbf{\Sigma}_i)$, where

$$\mathbf{\Sigma}_i = \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}, i = 1, 2.$$

(a) Find the moment generating function of $\boldsymbol{X}$.

(b) Find the moment generating functions for the marginal distributions of X_1 and X_2 using part (a). Deduce that X_1 and X_2 are each normally distributed as $N(0, 1)$.

(c) Show that $\boldsymbol{X}$ itself is not normally distributed if $\rho_1 \neq \rho_2$.

(d) What can you conclude from parts (b) and (c)?

(e) Show that the covariance of X_1 and X_2 can be zero if $\rho_1 \neq \rho_2$, but X_1 and X_2 are not independent.

6. Consider the two-way model,

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where $i = 1, 2, \cdots, 6$; $j = 1, 2, \cdots, 8$, α_i, β_j, and ϵ_{ij} are independently distributed as normal variates with zero means and variances σ_α^2, σ_β^2, and σ_ϵ^2, respectively. Let $\gamma_1 = \frac{\sigma_\alpha^2}{\sigma_\epsilon^2}$, $\gamma_2 = \frac{\sigma_\beta^2}{\sigma_\epsilon^2}$.

(a) Show how to obtain an exact confidence interval on $2\sigma_\alpha^2 + 3\sigma_\beta^2 + \sigma_\epsilon^2$ with a confidence coefficient of at least $1 - \alpha$.

(b) Find the uniformly minimum variance unbiased estimator of

$$\theta = \frac{1 + 8\gamma_1}{1 + 6\gamma_2}$$

(c) Find the expected values of $R(\beta \mid \mu)$ and $R(\beta \mid \mu, \alpha)$.

(d) If α_i were fixed and β_j random, what would the BLUE of $LSM(\alpha_i)$, the least-squares mean for α_i, be? Please provide justifications for your answer.

7. Let

$$Y = \begin{cases} 1, & \text{if a person has a disease,} \\ 0, & \text{otherwise,} \end{cases}$$

and suppose that presence of the disease depends on covariates $\boldsymbol{x}$ and on a separate exposure variable, E, with levels $j = 1, \ldots, k$, through the model

$$\log \frac{P(Y = 1|\boldsymbol{x}, E = j)}{P(Y = 0|\boldsymbol{x}, E = j)} = a(\boldsymbol{x}) + \lambda_j, \quad j = 1, \ldots, k,$$

for some function $a(\boldsymbol{x})$ (possibly unknown).

(a) Controlling for $\boldsymbol{x}$, what is the odds ratio comparing the odds of having the disease in two different exposure categories, j and j'.

(b) Suppose that the data on exposure is collected retrospectively; that is, diseased and non-diseased individuals are sampled and their exposure categories are determined. Let

$$Z = \begin{cases} 1, & \text{if a person is sampled,} \\ 0, & \text{otherwise,} \end{cases}$$

and suppose that

$$P(Z = 1|Y, \boldsymbol{x}, E) = P(Z = 1|Y, \boldsymbol{x}),$$

i.e., the selection probability for an individual given disease status, Y, and covariates, $\boldsymbol{x}$, is conditionally independent of exposure, E.

Derive an expression for $P(Y = 1|Z = 1, \boldsymbol{x}, E = j)$, and show that, controlling for $\boldsymbol{x}$, the odds-ratio comparing the odds of disease for sampled individuals in two different exposure categories, j and j', is the same as that derived in part (a).

8. The standard Cauchy (or Cauchy$(0, 1)$) distribution has density $f(u) = \frac{1}{\pi}\frac{1}{1+u^2}$, $-\infty < u < \infty$, cumulative distribution function $F(u) = \frac{1}{2} + \frac{1}{\pi}\arctan(u)$, and characteristic function $\varphi(t) = e^{-|t|}$. If U has a standard Cauchy distribution, then the distribution of $\sigma U + \mu$, where $-\infty < \mu < \infty$, and $\sigma > 0$, is known as a Cauchy distribution with location parameter (median) μ and scale parameter σ. Denote this distribution by the notation Cauchy(μ, σ).

(a) If U_1 and U_2 are independent random variables, $U_i \sim$ Cauchy(μ_i, σ_i), $i = 1, 2$, and a_1 and a_2 are real numbers, what is the distribution of $a_1U_1 + a_2U_2$? Justify (prove) your answer.

(b) Suppose that Y is a binary response following a generalized linear mixed model (GLMM) for $E(Y|\boldsymbol{U})$ (i.e., $g\big(E(Y|\boldsymbol{U})\big) = \eta$) with link function $g = F^{-1}$ (where F is the cdf of the Cauchy$(0, 1)$ distribution as given above) and linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{U}$, where $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{z} \in \mathbb{R}^q$ are known covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression parameters, and $\boldsymbol{U} = (U_1, \ldots, U_q)$ is a vector of independent random variables, $U_i \sim$ Cauchy$(0, \sigma_i)$, $i = 1, \ldots, q$. Show that the marginal model for Y is a GLM, identify its link function, and express its coefficient vector in terms of the elements of the GLMM.