Department of Statistics

Ph.D. Qualifying Examination

August 19, 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 the 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 each 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. Assume $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Provide complete theoretical development for the following problems:

   a. Let $\boldsymbol{A}$ be a matrix of full row rank. Derive the distribution of $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}$, where $\boldsymbol{b}$ is a vector of constants.

   b. Derive the mgf of $y = \boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x}$, where $\boldsymbol{B}$ is a given symmetric matrix.

   c. Derive $E(y)$.

2. Assume $y_{ijk} = \mu + \alpha_i + c_{ij} + \beta_j + e_{ijk}$, $i = 1, 2, 3, 4, 5$, $j = 1, 2, 3$, and $k = 1, 2$, where μ, α_i and β_j are constants, $c_{ij} \sim \mathrm{NID}(0, \sigma_c^2)$ and $e_{ijk} \sim \mathrm{NID}(0, \sigma_e^2)$. Denote

$$\boldsymbol{y}^T = (y_{111}, y_{112}, y_{121}, y_{122}, y_{131}, y_{132}, \cdots, y_{511}, y_{512}, y_{521}, y_{522}, y_{531}, y_{532}).$$

   a. Write an expression for $E(\boldsymbol{y})$.

   b. Write an expression for $\boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{y})$.

   c. Write an ANOVA table, including Source of variation, df, SS, MS, and EMS.

   d. Set up a test statistic for testing $\mathrm{H}_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$.

   e. Justify the motivation for the test statistic in (d), giving complete theoretical details.

3. Suppose that $Y_1, Y_2, ..., Y_N$ are independent observations from the exponential dispersion family,

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

   a. Let $\ell_i = \log f(y_i; \theta_i, \phi)$. Show that $\mu_i = E(Y_i) = b'(\theta_i)$.

   b. Using $-E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2$, show that $\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi)$.

   c. Using all N observations, find an expression for the log likelihood function.

   d. Now suppose you specify a generalized linear model $\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j$, and you use the canonical link function. Subsitute the model into the log likelihood function and identify the sufficient statistics for estimating the model parameters.

   e. The likelihood equations for a generalized linear model are

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)} \frac{d\mu_i}{d\eta_i} = 0, \quad j = 1, \cdots, p.$$

   Explain how the choice of link function for the GLM affects the likelihood equations and the asymptotic variance of the parameter estimates.

4. a. Explain what is meant by quasi-likelihood methods. For a univariate response, how is quasi-likelihood inference different from maximum likelihood inference? When are they equivalent?

   b. Give an example of a situation with count data in which quasi-likelihood methods might be useful. Define the model, explain how the analysis would differ from maximum likelihood with a standard Poisson regression model, and explain how to implement the quasi-likelihood method.

c. Refer to the previous part. Specify an alternative parametric model that, like quasi likelihood, allows a departure from a standard Poisson model for count data.

d. Now suppose the response is multivariate, as in a longitudinal study. Explain the sense in which generalized estimating equations (GEE) methodology is a multivariate version of quasi likelihood. Summarize the basic elements of that approach.

5. Let $X_1, \cdots, X_n$ be iid uniform$(0, \theta]$, where $\theta(> 0)$ is unknown.
a. Find the maximum likelihood estimator of θ.
b. Find the generalized likelihood ratio test (GLRT) for testing $\mathrm{H}_0 : \theta = \theta_0$ against the alternatives $\mathrm{H} : \theta \neq \theta_0$.
c. Show that $-2\log_e \lambda$ has an exact chi-squared distribution under H_0, where λ denotes the GLRT criterion.
d. Find the power function of the GLRT.

6. Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be two $p(\geq 3)$-component vectors such that conditionally on $\boldsymbol{\theta}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are iid $\mathrm{N}(\boldsymbol{\theta}, \boldsymbol{I}_p)$. Here $\boldsymbol{\theta} \in \mathcal{R}^p$ (the p-dimensional Euclidean space). The objective is to predict $\boldsymbol{Z}$ on the basis of $\boldsymbol{X}$. Throughout we assume squared error loss, i.e. $L(\boldsymbol{Z}, \boldsymbol{a}) = ||\boldsymbol{Z} - \boldsymbol{a}||^2$.
a. Consider the sequence $\{\pi_m, m \geq 1\}$ of priors for $\boldsymbol{\theta}$, where π_m is $\mathrm{N}(\boldsymbol{0}, m\boldsymbol{I}_p)$. Show that the predictive distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ (i.e. the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$) under the prior π_m is $\mathrm{N}((1 - B_m)\boldsymbol{X}, (2 - B_m)\boldsymbol{I}_p)$, where $B_m = (1 + m)^{-1}$.
b. Show that $(1 - B_m)\boldsymbol{X}$ is the Bayes predictor of $\boldsymbol{Z}$ under the prior π_m.
c. Show that $\boldsymbol{X}$ is a minimax predictor of $\boldsymbol{Z}$.
d. Show that the frequentist risk (i.e. conditional on $\boldsymbol{\theta}$) of the Stein predictor $(1 - \frac{p-2}{||\boldsymbol{X}||^2})\boldsymbol{X}$ of $\boldsymbol{Z}$ is smaller than that of $\boldsymbol{X}$. [NOTE: You need not explicitly calculate the risk of this predictor.]

7. Let $S_n = \sum_{i=1}^n X_i, n \geq 1$ where $\{X_n, n \geq 1\}$ is a sequence of independent mean 0 random variables and let $\{\lambda_n, n \geq 1\}$ be a bounded sequence of positive constants. Suppose that $X_i \in \mathcal{L}_{2+\lambda_n}, 1 \leq i \leq n,\ n \geq 1$. Prove that if

$$\sum_{i=1}^{n} E|X_i|^{2+\lambda_n} = o(s_n^{2+\lambda_n}),$$

where $s_n^2 = \sum_{i=1}^n EX_i^2, n \geq 1$, then

$$\frac{S_n}{s_n} \xrightarrow{d} N(0, 1).$$

Be sure to point out where and how you use the boundedness assumption regarding $\{\lambda_n, n \geq 1\}$.

8. Let $S_n = \sum_{i=1}^n X_i, n \geq 1$ where $\{X_n, n \geq 1\}$ is a sequence of independent random variables and let $\{b_n, n \geq 1\}$ be a sequence of positive constants with $b_n \uparrow \infty$.
a. Prove that if $\frac{S_n}{b_n} \to 0$ almost certainly, then $\frac{\max_{1 \leq j \leq n} |S_j|}{b_n} \to 0$ almost certainly.
b. Prove that if $\frac{S_n}{b_n} \xrightarrow{P} 0$, then $\frac{\max_{1 \leq j \leq n} |S_j|}{b_n} \xrightarrow{P} 0$.