**BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS**

**THEORY, SECTION 2**

**(9:00 AM- 1:00 PM**
**Friday, August 12, 2011)**

INSTRUCTIONS:

a) This is a **CLOSED-BOOK** examination.

b) The time limit for this Examination is four hours.

c) Answer any TWO (2) (BUT ONLY TWO) of the THREE (3) questions that follow.

d) Put the answers to different questions on separate sets of paper.

e) Put your code letter, **NOT YOUR NAME**, on each page. The same code will be used for Section 1 and Section 2 of the PhD Theory Exam. Please keep the code confidential and do not share this information with any students or faculty.

f) Return the examination with a signed statement of the UNC honor pledge, separately from your answers. The pledge statement is given on the last page of the exam handout.

g) In the questions to follow, you are required to answer only what is asked, and not to tell all you know about the topics involved.

1. For a given $i = 1, \ldots, n$, let $X_i$ and $Y_i$ be independent exponential random variables with means $1/(\psi\lambda_i)$ and $1/\lambda_i$, respectively. Assume further that the bivariate random vectors $(X_i, Y_i)$ are independent, for $i = 1, \cdots, n$. The parameter of interest is $\psi$, with the $\lambda_i$'s being unknown parameters which may vary for $i = 1, \ldots, n$. Note: in this problem, asymptotics refers to $n \to \infty$.

   (a) Write the log-likelihood function $L_1(\psi, \lambda_1, \cdots, \lambda_n)$ based on $(X_i, Y_i), i = 1, \ldots, n$. Derive the score equation that defines the maximum likelihood estimator for $\psi$ based on $L_1$. Denote that equation by $U_1(\psi) = 0$.

   (b) Are the standard regularity conditions for the consistency and asymptotic normality of the maximum likelihood estimators for $\psi, \lambda_1, \cdots, \lambda_n$, based on $L_1(\psi, \lambda_1, \cdots, \lambda_n)$ satisfied in this problem? Explain.

   (c) Assuming the regularity conditions for the maximum likelihood estimators from $L_1$ are satisfied, derive an explicit expression for the asymptotic variance of $\hat{\psi}$ via the Fisher information matrix from $L_1$.

   (d) Show that $U_1(\psi)$ depends on the data only through the ratios $T_i = X_i/Y_i$. Derive the pdf of $T_i$ and show that it does not depend on $\lambda_i$.

   (e) Use the density of $T_1, \cdots, T_n$ to obtain a likelihood function, $L_2(\psi)$. Compare the score equation derived for $\psi$ from $L_2$ with the function $U_1(\psi)$ derived in part (a). Is the maximum likelihood estimator for $\psi$ from $L_2$ identical to that from $L_1$? Derive the asymptotic variance for the maximum likelihood estimator based on $L_2$ using standard asymptotic calculations and compare with that in part (c). Discuss.

   (f) Let $g_i(\psi) = Y_i - \psi X_i, i = 1, \ldots, n$ and consider estimation of $\psi$ by solving

$$\sum_{i=1}^{n} w_i g_i(\psi) = 0$$

   for $\psi$, where $w_i, i = 1, \ldots, n$ are finite constants. Determine the asymptotic variance of the estimator thus obtained and find the optimal $w_i$'s (up to a proportionality constant). Compare the efficiency of this optimal estimator to that based on $w_i = 1, i = 1, \ldots, n$ and to that from $U_1(\psi)$. Is the optimal estimator usable in practice?

2. In this problem, we consider the univariate density

$$p(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \qquad 0 < y < 1, \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function, and $\alpha > 0$ and $\beta > 0$. One may reparameterize (1) in terms of $(\mu, \phi)$, such that

$$p(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1, \tag{2}$$

where $0 < \mu < 1$, $\phi > 0$, $E(y) = \mu$ and $\text{Var}(y) = \mu(1-\mu)(1+\phi)^{-1}$.

(a) Find explicit expressions for $(\mu, \phi)$ in terms of $(\alpha, \beta)$.

(b) Let $Y_1, \ldots, Y_n$ be a random sample from the density (2). Show that the joint density for $Y_1, \ldots, Y_n$ belongs to the multivariate exponential family of distributions, identify the canonical statistics and parameters, determine its rank, and find the joint complete sufficient statistics for $(\mu, \phi)$.

(c) Now, suppose that $Y_1, \ldots, Y_n$ are independent random variables, where each $Y_i$, $i = 1, \ldots, n$, follows the density in (2) with unknown mean $\mu_i$ and unknown precision $\phi$. Suppose that $X_i$ is a $p \times 1$ vector of covariates, with $g(\mu_i) = \beta^T X_i$ , where $\beta$ is a $p \times 1$ vector of unknown regression coefficients and $g(\cdot)$ is an arbitrary known link function. Define $\xi = (\beta, \phi)$.

(i) Derive the score function for $\xi$ and show that the expectation of the score function equals 0 at the true value of $\xi$.

(ii) Show that the Fisher information matrix of $\xi$ is given by

$$I(\xi) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\phi} \\ I_{\phi\beta} & I_{\phi\phi} \end{pmatrix},$$

where $I_{\beta\beta} = \phi X^T W X$, $I_{\beta\phi} = I_{\phi\beta}^T = X^T T c$, $I_{\phi\phi} = \text{tr}(D)$, $c = (c_1, \ldots, c_n)^T$ with $c_j = \phi\left\{\psi'(\mu_j\phi)\mu_j - \psi'((1-\mu_j)\phi)(1-\mu_j)\right\}$, $T = \text{diag}\left(\frac{1}{g'(\mu_1)}, \ldots, \frac{1}{g'(\mu_n)}\right)$, $g'(z) = \frac{d}{dz}g(z)$, $\psi(z) = \frac{d}{dz}\log\Gamma(z)$, $\psi'(z) = \frac{d}{dz}\psi(z)$, $D = \text{diag}(d_1, \ldots, d_n)$ with $d_j = \psi'(\mu_j\phi)\mu_j^2 + \psi'((1-\mu_j)\phi)(1-\mu_j)^2 - \psi'(\phi)$, and $W = \text{diag}(w_1, \ldots, w_n)$ with

$$w_j = \phi\left\{\psi'(\mu_j\phi) + \psi'((1-\mu_j)\phi)\right\} \frac{1}{\{g'(\mu_j)\}^2}.$$

(d) Let $\hat{\xi} = (\hat{\beta}, \hat{\phi})$ denote the maximum likelihood estimator of $\xi$. Derive from first principles the asymptotic distribution of $\hat{\xi}$, properly normalized.

(e) The multivariate generalization of the distribution in (2) is called the Dirichlet distribution, which may be defined as as follows. Let $r_1, \ldots, r_k$ be independent random variables, with $r_j \sim \text{gamma}(\alpha_j, 1)$, $j = 1, \ldots k$. The gamma$(a, b)$ density is given by $f(r) = \frac{\Gamma(a)}{b^a} r^{a-1} \exp(-br)$ for $r > 0$, $a > 0$, $b > 0$. Define $s = \sum_{j=1}^{k} r_j$ and $q_j = \frac{r_j}{s}$, $j = 1, \ldots, k$. The joint density of $(q_1, \ldots, q_{k-1})$ is called the Dirichlet density. Derive the joint density of $(q_1, \ldots, q_{k-1})$.

3. Consider the linear model

$$Y = X\beta + \epsilon, \tag{1}$$

where $X$ is a $n \times p$ covariate matrix, $\beta$ is a $p \times 1$ vector of regression coefficients, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, where the $\epsilon_i$'s are i.i.d. $N(0, \sigma^2), i = 1, \ldots, n$. In this problem, both $\beta$ and $\sigma^2$ are unknown.

(a) Suppose that $\text{rank}(X) = r \leq p$ and we wish to test

$$H_0 : \ell^T\beta = \theta_0 \quad \text{versus} \quad H_1 : \ell^T\beta \neq \theta_0, \tag{2}$$

where $\ell \in C(X)$, $C(X)$ denotes the column space of $X$, and $\theta_0$ is a specified constant. Derive a UMPU size $\alpha$ test for the hypotheses in (2). Determine the exact distribution of the test statistic under $H_0$ and $H_1$ as well as an explicit expression of the critical value to make the test size $\alpha$.

(b) Consider the model in (1) and the hypothesis in (2).

(i) Derive an explicit closed-form expression for the asymptotic power function of the UMPU test in part (a)

(ii) Suppose that $p = 2$, $X$ is $n \times 2$ where the first column consists of a vector of ones, $\beta = (\beta_0, \beta_1)^T$, $\ell = (0, 1)$, $\text{rank}(X) = 2$, and $\sum_{i=1}^n x_i = n/2$, where $(x_1, \ldots, x_n)^T$ denotes the second column of $X$. Use the asymptotic power function of part (i) to derive an explicit closed form sample size formula for an $\alpha$ level test with prespecified power.

(c) Consider the model in (1) and suppose that $\text{rank}(X) = p$. We wish to test

$$H_0 : R\beta = b_0 \quad \text{versus} \quad H_1 : R\beta \neq b_0, \tag{3}$$

where $R$ is an $s \times p$ specified matrix of constants of rank $s \leq p$ and $b_0$ is a specified $s \times 1$ vector. Derive the size $\alpha$ likelihood ratio test for this hypothesis and determine the exact distribution of the likelihood ratio statistic (or a monotonic function of it) under $H_0$ and $H_1$. In carrying out this derivation, you need to derive all relevant estimates under $H_0$ and $H_1$.

(d) Derive the score test for the hypothesis and setup of part (c), and state its asymptotic distribution under $H_0$.

(e) Consider the model in (1) and suppose that $\text{rank}(X) = r \leq p$. Derive an exact 95% confidence region for $R\beta$, where $R$ is an $s \times p$ matrix of constants of rank $s \leq r$, and all rows of $R$ are contained in $C(X)$.

(f) Consider the model in (1) and suppose that $\text{rank}(X) = r \le p$. Derive a UMPU size $\alpha$ test for testing $H_0 : \sigma^2 \le \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$, where $\sigma_0^2$ is a specified constant. Determine the exact distribution of the test statistic under $H_0$ and determine an explicit expression of the critical value to make the test size $\alpha$.

**2011 PhD Theory Exam, Section 2**

Statement of the UNC honor pledge:

*"In recognition of and in the spirit of the honor code, I certify that I have neither given nor received aid on this examination and that I will report all Honor Code violations observed by me."*

(Signed) ______________________________

NAME

(Printed) ______________________________

NAME