Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part I
August 27, 2013
12:30-4:30pm, Room 133 SMI

- There are a total of FOUR (4) problems in this exam. Please do a total of THREE (3) problems.
- Each problem must be done in a separate exam book.
- Please turn in THREE (3) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. Suppose $\boldsymbol{X} = (X_1, \ldots, X_K)'$ follows a multinomial distribution Multi$(n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ with $\pi_k > 0$, $\forall k$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ be a parameter vector with $q < K-1$. Let $h : \mathcal{R}^q \to \mathcal{R}^K$ be a known re-parameterization function such that $\boldsymbol{\pi} = h(\boldsymbol{\theta})$. Assume the parameter space of $\boldsymbol{\theta}$ is an open set. Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$. Assume the re-parameterization function h is chosen so that $\widehat{\boldsymbol{\theta}}$ exists and is unique.

   For the following three questions, please show the details of derivation and calculation.

   (a) Write down the likelihood function of $\boldsymbol{\theta}$. Derive the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, as $n \to \infty$, while K is fixed. Express the limiting covariance matrix in terms of $\boldsymbol{\theta}$ and h. Additionally, carefully state any extra conditions used to obtain the result.

   (b) To estimate $\boldsymbol{\pi}$, we have two estimators: $\widehat{\boldsymbol{\pi}}_1 = h(\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\pi}}_2 = \boldsymbol{X}/n$. Let $g : \mathcal{R}^K \to \mathcal{R}$ be any continuously differentiable function. Show that $\sqrt{n}\Big[g(\widehat{\boldsymbol{\pi}}_1) - g(\boldsymbol{\pi})\Big] \to N(0, \sigma_1^2)$ and $\sqrt{n}\Big[g(\widehat{\boldsymbol{\pi}}_2) - g(\boldsymbol{\pi})\Big] \to N(0, \sigma_2^2)$, as $n \to \infty$, while K is fixed. Express σ_1^2 and σ_2^2 in terms of $\boldsymbol{\theta}$, h and g.

   (c) Show that $\sigma_2^2 > \sigma_1^2$.

2. Suppose that $X_1, \cdots, X_n$ are independent random variables, and X_i follows a binomial distribution with m trials and success probability p_i, where

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_i, \qquad i = 1, \cdots, n,$$

β_0 and β_1 are parameters, z_i is a known covariate, and $z_1 < \cdots < z_n$.

(a) Find the likelihood function of (β_0, β_1). Prove that if $0 < X_1 + \cdots + X_n < mn$, the MLE, $(\widehat{\beta}_0, \widehat{\beta}_1)$, of (β_0, β_1) exists and is unique.

(b) State some appropriate conditions and then derive a non-degenerate limiting distribution of $\sqrt{n}[(\widehat{\beta}_0, \widehat{\beta}_1) - (\beta_0, \beta_1)]$ under the conditions, as $n \to \infty$, while m is fixed.

(c) A prior distribution Π of (β_0, β_1) has a joint cumulative distribution function given by

$$\Pi(\beta_0 \le u, \beta_1 \le v) = \frac{e^u}{1+e^u} 1(v \ge 0), \qquad -\infty < u, v < \infty,$$

where $1(\cdot)$ is an indicator function. Derive the posterior distribution of (β_0, β_1). Find the Bayesian estimator of $(p_1, \cdots, p_n)$ under the loss function

$$L((p_1, \cdots, p_n), (a_1, \cdots, a_n)) = \sum_{j=1}^{n} (p_j - a_j)^2/[p_j(1-p_j)].$$

(d) Assume that $\beta_1 = 0$. A prior distribution of β_0 has a cumulative distribution function

$$G(u) = \frac{e^u}{1+e^u}, \qquad -\infty < u < \infty.$$

Find the Bayesian estimator of p_1 under the loss function $L(p_1, a) = (p_1 - a)^2/[p_1(1-p_1)]$. Is the Bayesian estimator a minimax estimator of p_1 under the same loss function? Prove or disprove your answer.

3. Consider random variables $X_1, X_2, \ldots$ from a common distribution with density $f(x)$ and cumulative distribution function $F(x)$. The distribution satisfies $P(0 \leq X_i \leq 1) = 1$. In addition, we have random variables $U_1, U_2, \ldots \sim \text{Uniform}(0, 1)$. All random variables are mutually independent.

   (a) In terms of these random variables, define

   $$N = \min\{n \geq 1 : U_n \leq 1 - X_n\}$$

   and determine the probability mass function of N.

   (b) Show that the cumulative distribution function $G(y)$ of $Y = X_N$ equals

   $$G(y) = \frac{F(y) - m(y)}{1 - m(1)}$$

   where $m(y) = \int_0^y x f(x)\, dx$, for $y \in [0, 1]$.

   (c) Suppose we have independent and identically distributed copies $Y_1, Y_2, \ldots, Y_m$ of Y from above, and we seek to estimate the the distribution corresponding to f and F. Construct estimators using:

      i. parametric model: $f(x) = \frac{\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}$ for positive parameters θ_1, θ_2.

      ii. nonparametric model entailing no finite-dimensional constraints.

   (d) Assume that instead of $Y_1, Y_2, \ldots, Y_m$ of Y, we only observe realizations of copies $N_1, N_2, \ldots, N_m$ of N. What properties of the distribution can be estimated based on $N_1, N_2, \ldots, N_m$?

4. Suppose that we have observations $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$, where $\{X_1, \ldots, X_n\}$ come from Class A, and $\{Y_1, \ldots, Y_n\}$ come from Class B.

The following notation and assumptions will be used in parts (a)–(c). Let c_1 and c_2 be finite constants. Assume that $\{u_j\}_{j=1}^n \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\{v_j\}_{j=1}^n \overset{\text{i.i.d.}}{\sim} N(0,1)$, $e \sim N(0, \sigma_e^2)$, where $\sigma_e^2 \in (0, \infty)$. Assume that $\{u_j\}_{j=1}^n$, $\{v_j\}_{j=1}^n$ and e are mutually independent.

For both parts (a) and (b), use the following t-statistic:

$$T = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{2}{n}}\, S}, \quad \text{for } n \geq 2,$$

where $\overline{Y}$ and $\overline{X}$ are sample means and $S^2 = \frac{\sum_{j=1}^n (X_j - \overline{X})^2 + \sum_{j=1}^n (Y_j - \overline{Y})^2}{2n-2}$ is the pooled estimator of variance.

(a) Assume that for $j = 1, \ldots, n$,

$$X_j = c_1 + u_j + \frac{e}{2}, \qquad Y_j = c_2 + v_j + \frac{e}{2}. \tag{1}$$

We wish to test $H_0 : c_1 = c_2$, using the above t-statistic T. Derive the null distribution of this test statistic.

(b) Assume that for $j = 1, \ldots, n$,

$$X_j = c_1 + u_j - \frac{e}{2}, \qquad Y_j = c_2 + v_j + \frac{e}{2}. \tag{2}$$

We wish to test $H_0 : c_1 = c_2$, using the above t-statistic T. Derive the null distribution of this test statistic.

(c) Comment on the suitability of the t-statistic T for the testing problems in (a) and (b).