Department of Statistics

PhD Qualifying Exam

August 2008

University of Florida

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 your chosen identifying number on every page.
5. Do not write your name anywhere on your exam.
6. Write only on one side each sheet of paper. For each problem you do, start the problem on a new page. At the end of the exam, for each problem, staple together all pages for that problem.
7. Clearly label each part of each question with the question number and the part, e.g., **1(a)**.
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.
10. Do not write near the upper left corner of the page where the pages will be stapled together.

1. Let $\beta_1, \beta_2, \beta_3$ be the interior angles of a triangle, so that $\beta_1 + \beta_2 + \beta_3 = 180$ degrees. Suppose we have available estimates Y_1, Y_2, Y_3 of $\beta_1, \beta_2, \beta_3$, respectively. We assume that $Y_i \sim N(\beta_i, \sigma^2)$, $i = 1, 2, 3$ (σ is unknown) and that the Y_i's are independent.

(a) What is the "best" estimate of β_1? (Part of the question is to explain what is meant by "best".)

(b) Construct a $(1-\alpha)$ level confidence interval for β_1.

2. Consider the one-way ANOVA model

$$Y_{ij} = \beta_i + \epsilon_{ij}, \qquad i = 1, \ldots, k, \; j = 1, \ldots, n_i$$

with $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Derive the F-test for testing

$$H_0 : \beta_i = \alpha \times i \text{ for } i = 1, \ldots, k \qquad \text{vs.} \qquad H_1 : \text{not } H_0.$$

3. Let P_n and Q_n be probability measures on $(\mathbb{R}, \mathscr{R})$ ($\mathscr{R}$ represents the Borel sets on $\mathbb{R}$) having densities f_n and g_n, respectively, with respect to Lebesgue measure λ. In view of Scheffé's theorem, one might expect that if $f_n(x) - g_n(x) \to 0$ as $n \to \infty$ for λ-almost all x, then

$$\sup_{A \in \mathscr{R}} |P_n(A) - Q_n(A)| \to 0, \qquad (*)$$

but in fact, as part (c) of this problem shows, $(*)$ can fail even under much stronger conditions.

(a) Prove that $(*)$ holds if $f_n(x) - g_n(x) \to 0$ for λ-almost all x and there exists a Lebesgue-integrable function h such that $|f_n - g_n| \le h$ for all $n \ge 1$.

(b) Prove that $(*)$ holds if the sequences $\{P_n : n \ge 1\}$ and $\{Q_n : n \ge 1\}$ are (uniformly) tight and $f_n - g_n \to 0$ uniformly on compact sets, i.e.,

$$\sup_{x \in [-M, M]} |f_n(x) - g_n(x)| \to 0 \quad \text{as } n \to \infty \text{ for all } M \ge 0.$$

(c) Give an example to show that $(*)$ can fail even when $\sup_{x \in \mathbb{R}} |f_n(x) - g_n(x)| \to 0$ as $n \to \infty$.

4. Let $X_1, X_2, \ldots$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$, and let $Y_n = \sum_{k=1}^n kX_k$. Prove that for suitably chosen sequences of constants $\{a_n : n \ge 1\}$ and $\{b_n : n \ge 1\}$,

$$\frac{Y_n - a_n}{b_n} \rightsquigarrow Z,$$

where $Z \sim N(0, 1)$ and $\rightsquigarrow$ denotes convergence in distribution.

5. Suppose that a population of individuals consists of two sub-populations or groups, G_1 and G_2, with $100\pi\%$ of the population belonging to G_1 and $100(1-\pi)\%$ belonging to G_2, where π is known.

(a) Assume that measurements $\boldsymbol{X}$ made on individuals have the following distributions in the two groups:

$$G_1 : \boldsymbol{X} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$$
$$G_2 : \boldsymbol{X} \sim N_p(\boldsymbol{\mu}_2, \Sigma).$$

Let $\boldsymbol{x}$ be an observation made on an individual drawn at random from the combined population and let $Y = 1$ if the individual is from G_1 and $Y = 0$ if the individual is from G_2.

(i) Show that the conditional (or "posterior") odds that the individual belongs to G_1 given $\boldsymbol{x}$ have the form

$$\text{odds}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}), \qquad (*)$$

and express α and $\boldsymbol{\beta}$ in terms of π, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Σ.

(ii) Suppose that we are given training data consisting of measurements of $\boldsymbol{X}$ for n_1 individuals drawn at random from G_1 and n_2 individuals drawn at random from G_2. Assuming that the model of part (a) holds, express the MLEs of α and $\boldsymbol{\beta}$ in terms of the MLEs of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Σ (you may assume that the latter are known).

(b) Suppose that instead of the full model described in part (a), we assume only that the conditional odds $Y = 1$ given $\boldsymbol{X} = \boldsymbol{x}$ has the form $(*)$. Note that the sampling probability for an individual depends on Y and $\boldsymbol{X}$ only through Y.

(i) Assuming that the total population size is N, what is the sampling probability for an individual in G_1 (i.e., with $Y = 1$) and what is it for an individual in G_2 (i.e., with $Y = 0$)? Let the variable Z take the value 1 for individuals that are included in the sample and 0 otherwise. Show that the odds that $Y = 1$ for a sampled individual with covariate vector $\boldsymbol{x}$ have the form

$$\text{odds}(Y = 1|Z = 1, \boldsymbol{X} = \boldsymbol{x}) = \exp(\alpha^* + \boldsymbol{\beta}^T\boldsymbol{x}), \qquad (**)$$

and express α^* in terms of α, π, n_1, and n_2.

(ii) If we run ordinary logistic regression on the combined $(y, \boldsymbol{x})$ data from the two groups to estimate parameters, how can the "intercept" estimate be adjusted if we wish to estimate the α in $(*)$ rather than the α^* in $(**)$

(c) Comment very briefly on the relative advantages and disadvantages of the methods of estimating α and $\boldsymbol{\beta}$ described in parts (a) and (b).

6. Suppose that $X \sim \mathrm{Bin}(n,p)$, $0 < p < 1$, and let $\theta = \arcsin(\sqrt{p})$ (or $\theta = \sin^{-1}(\sqrt{p})$ if you prefer that notation).

(a) Consider the estimator

$$\tilde{\theta} = \arcsin\left(\sqrt{\frac{X+c}{n+2c}}\right),$$

where $c \geq 0$ is a fixed constant not depending on n. Regardless of the value of c, the estimator $\tilde{\theta}$ is variance stabilized in the sense that the variance of its asymptotic distribution does not depend on p. Prove this, i.e., show that the limiting distribution of $\sqrt{n}(\tilde{\theta}-\theta)$ does not depend on p. *Hint: First consider $\sqrt{n}(\tilde{p}-p)$, where $\tilde{p} = (X+c)/(n+2c)$. Also, in case you've forgotten, $\frac{d}{dw}\arcsin(w) = 1/\sqrt{1-w^2}$.*

(b) Argue that

$$E(\tilde{\theta}) = \theta + \frac{(4c-1)b_p}{n} + O(n^{-3/2}),$$

where b_p depends on p but not on n, so that the choice $c = 1/4$ reduces the asymptotic bias of $\tilde{\theta}$ by an order of magnitude. *Your proof does not have to be completely rigorous in its handling of the $O(n^{-3/2})$ term.*

7. Starting with a density f with mean 0 and covariance matrix $\sigma^2 I$, create the location family $\{f(x-\theta) : |\theta| < \infty\}$. Let $\boldsymbol{X}_{p\times 1} \sim f(x-\theta)$ and consider a prior on θ to be $\theta \sim f^{*n}$, the n-fold convolution of f with itself. (The *convolution* of f with itself is $f^{*2}(x) = \int f(x-y)f(y)dy$. The *$n$-fold convolution* is $f^{*n}(x) = \int f^{*(n-1)}(x-y)f(y)dy$.) An equivalent formulation is to let $U_i \sim f(u)$, $i = 0,\ldots,n$ iid, $\theta = \sum_1^n U_i$, , and $\boldsymbol{X} = U_0 + \theta$.

(a) Show that the Bayes rule against squared error loss is $\frac{n}{n+1}\boldsymbol{X}$. Note that n is a prior parameter.

(b) Calculate the mean squared error of $\frac{n}{n+1}\boldsymbol{X}$, and compare it to the mean squared error of $\boldsymbol{X}$. Under what circumstances would you prefer $\frac{n}{n+1}\boldsymbol{X}$?

(c) Show that, marginally, $|\boldsymbol{X}|^2/(p\sigma^2)$ is an unbiased estimator of $n+1$. Use this fact to construct an empirical Bayes estimator of θ that resembles a Stein estimator.

8. Let $X \sim N(\theta, 1)$ and $L(\theta, \delta) = (\theta - \delta)^2$.

(a) Define *generalized Bayes estimator* and show that X is a generalized Bayes estimator.

(b) Define *limit of Bayes estimators* and show that X is a limit of Bayes estimators. In particular, exhibit a sequence of proper Bayes estimators δ^{π_n} that satisfy (i) $\delta^{\pi_n} \to X$ and (ii) $R(\theta, \delta^{\pi_n}) \to R(\theta, X)$. Hence, X is both generalized Bayes and a limit of Bayes estimators.

(c) For the prior measure $\pi(\theta) = e^{a\theta}, a > 0$

(i) Show that the generalized Bayes estimator is $X + a$.

(ii) For $a > 0$, show that there is no sequence of proper priors for which $\delta^{\pi_n} \to X + a$.

(Hint: You might want to use the fact that the Bayes estimators have the form $x + \nabla \log m(x)$, where m is the marginal distribution. If you do use this fact you must first prove it.)