Ph. D. Qualifying Exam
Monday, August 18, 2003

Please submit solutions to at most **seven** problems. You have four hours. No one is expected to answer all the problems correctly. Partial credit will be given. All problems are worth an equal amount of credit.

Put your solution to each problem on a separate sheet of paper.

Applied Statistics

Problem 1. Suppose that $\{X_i, i = 1, \ldots, n\}$ are independent random variables with $\mathrm{E}(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. In ANOVA and regression, transformations are often used to stabilize the variances of $\{X_i\}$.

(a) If $\sigma_i = g(\mu_i)$ and the transformation $Y_i = f(X_i)$ is used to stabilize the variances, find an approximate relationship between $f(\cdot)$ and $g(\cdot)$ using the Taylor expansion.

(b) If $\sigma_i = c\mu_i^\alpha$, which transformation $f(\cdot)$ will approximately stabilize the variance $\mathrm{Var}(X_i) = \sigma_i^2$?

(c) In practice, α is usually unknown. If $\hat{\sigma}_i^2$ and $\hat{\mu}_i$ are available, give an empirical way to find the transformation $f(X_i)$.

Problem 2. In a complete factorial experiment, consider the following random-effects model:

$$Y_{ij} \;=\; \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \ldots, a; \;\; j = 1, \ldots, b,$$

where μ is the overall mean, α_i is the random effect corresponding to the ith level of factor A, and β_j is the random effect due to the jth level of factor B. The α_i's are iid $N(0, \sigma_\alpha^2)$ variables, the β_j's are iid $N(0, \sigma_\beta^2)$ variables, and the ϵ_{ij}'s are iid $N(0, \sigma_\epsilon^2)$ variables. Furthermore, $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\epsilon_{ij}\}$ are assumed to be independent.

(a) What are the distributions of Y_{ij}, $Y_{i\cdot}$, and $Y_{\cdot j}$? Find the covariance $\mathrm{Cov}(Y_{ij}, Y_{kl})$.

(b) Find unbiased estimates for the variances σ_α^2, σ_β^2, and σ_ϵ^2 based on the observations $\{Y_{ij}, \quad i = 1, \ldots, a; \;\; j = 1, \ldots, b\}$.

Problem 3. Consider the linear regression model:

$$Y = X\beta + \xi,$$

where $Y = (y_1, \ldots, y_n)'$, $\xi = (\xi_1, \ldots, \xi_n)'$, $\beta = (\beta_1, \ldots, \beta_p)'$ and X is an $n \times p$ full-rank matrix. The process $\{\xi_i\}$ is generated by the model:

$$\xi_i - \phi\xi_{i-1} = a_i, \qquad (**)$$

where $|\phi| < 1$ and $\{a_i, i = 0, \pm 1, \pm 2 \ldots, \}$ are iid $N(0, \sigma^2)$ variables.

(a) Show that $\xi_i = \sum_{j=0}^{\infty} \phi^j a_{i-j}$ is a solution of equation (**) above.

(b) Based on the expression in (a) for ξ_i, calculate autocorrelations $\gamma_k = \mathrm{Cov}(\xi_i, \xi_{i+k})$ for $k \geq 0$. Show that $\gamma_k = \phi\gamma_{k-1}$ for any $k \geq 0$.

(c) Suppose that ϕ is known. How do you estimate β in this setting? Discuss the properties of your estimate such as mean, covariance matrix, and distribution of $\hat{\beta}$. Compare your estimate with the least squares estimate of β.

Problem 4. In a generalized linear model, the response variable Y is assumed to have a density function with the form:

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

(a) Identify θ, $b(\theta)$, $a(\phi)$, and $c(y, \phi)$ when Y has a binomial distribution and when Y has a Poisson distribution.

(b) Let $l(\theta, y) = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$. Derive the mean and variance of Y from the relations $E\left(\frac{\partial l}{\partial \theta}\right) = 0$ and $E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0$.

Probability

Problem 5. Let $X_1, X_2, \ldots$ be independent. Show that $\sup X_n < \infty$ a.s. if and only if $\sum_n P(X_n > M) < \infty$ for some $M < \infty$.
Hint : Use Borel-Cantelli Lemmas.

Problem 6. Let X_n be a sequence of random variables and let $a_n \to \infty$ be an increasing sequence of constants such that $a_n X_n \to_d X$, as $n \to \infty$. Let g be a continuously differentiable function at 0. Show that

$$a_n (g(X_n) - g(0)) \to_d g'(0)X.$$

Problem 7. Let $X_1, X_2, \ldots$ be i.i.d. with a uniform distribution on $(0,1)$. Consider the statistic

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \min(X_i, X_j).$$

Obtain the asymptotic distribution of U_n.
Hint: The df and pdf of X_1 are easy. So it will be easy to compute all moments etc. that you may require.

Theoretical Statistics

Problem 8. Consider the following model:

$$\begin{aligned} Y_1 &= \alpha_1 + \alpha_2 + e_1 \\ Y_2 &= 2\alpha_2 + e_2 \\ Y_3 &= -\alpha_1 + \alpha_2 + e_3 \end{aligned}$$

where $e_i \sim$ independent $N(0, \sigma^2)$. Derive the F-test for testing $H_0 : \alpha_1 = 2\alpha_2$.

Problem 9. Let $X_1, X_2, \ldots, X_n$ be iid with a **shifted** geometric distribution

$$P_\theta(X = x) = \left(\frac{1}{2}\right)^{x-\theta+1}, \quad x = \theta, \theta+1, \theta+2, \ldots, \quad -\infty < \theta < \infty\,.$$

(a) Find a minimal sufficient statistic for θ (and prove that your statistic has both these properties).

(b) Find the MLE for θ.

(c) Show that the statistic you found in part (a) is complete.

Problem 10. Suppose that we have two independent random samples: $X_1, X_2, \ldots, X_n$ are iid $N(0, \sigma^2)$, and $Y_1, Y_2, \ldots, Y_m$ are iid $N(0, \tau^2)$.

(a) Find the likelihood ratio test statistic (LRT) for $H_0 : \sigma^2 = \tau^2$ versus $H_1 : \sigma^2 \neq \tau^2$.

(b) Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^m Y_i^2}\,.$$

Problem 11. For a given $\theta \in \Re$, let $f(x|\theta)$ be a density function given by:

$$f(x|\theta) = \frac{\exp(-x+\theta)}{(1+\exp(-x+\theta))^2}\ , \quad \text{for} \ \ x \in \Re.$$

Assuming that a uniform random number generator between 0 and 1 is available, suggest a method for generating exact samples from f. Write an algorithm (e.g. in matlab code) to implement this method.

Problem 12. Let X be a random variable whose mean we are interested in estimating using Monte Carlo methods. Let Y be another random variable and define $Z = E[X|Y]$.

(a) Show that $E[Z] = E[X]$, but variance$(Z) \leq$ variance(X).

(b) Suppose that Y is an exponential random variable with mean one, and given $Y = y$, X is an exponential random variable with mean y. Use variance reduction by conditioning to set up an efficient Monte Carlo method to estimate $Pr\{XY \leq 3\}$. State the procedure algorithmically.

Biostatistics

Problem 13. Assume that Y is a random variable taking values 0,1 and that X is a random variable such that:
If $Y = 0$, X has density f_0, and
If $Y = 1$, X has density f_1.

(a) Assume that the proportion of people in the population for whom $Y = 1$ is p. Derive a general expression for

$$\Pr(Y = 1|X = x)$$

(b) Show that if f_i is the $N(\mu_i, \sigma^2)$ density for $i = 0, 1$, then

$$\Pr(Y = 1|X = x) = \frac{1}{1+\exp\{-(\alpha+\beta x)\}}$$

and give expressions for α and β. That is, the logistic model arises naturally.

(c) Assume you have a random sample of individuals, n_0 of them have $Y = 0$, and n_1 have $Y = 1$. What are the maximum likelihood estimates of α and β?

Problem 14. Suppose that an investigator wants to do a clinical trial testing the difference between two normal means. She is willing to assume that the variance is equal for the two groups and is σ^2. She specifies a minimum difference that she thinks is clinically important: $\Delta = \mu_1 - \mu_2$, and specifies the level of type I error she is willing to accept, α. Because of cost considerations $n = n_1 + n_2$ is fixed ($n_1 =$ the number to be randomized in the first group, $n_2 =$ the number to be randomized in the second group.) What are the values of n_1 and n_2 that maximize the power of the study?