DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST, 2003

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all six questions. Each will be graded from 0 to 10.

b. Use a different booklet for each question. Write the problem number and your code number (**NOT YOUR NAME**) on the outside cover.

c. Keep scratch work on separate pages in the same booklet.

d. If you use a "well known" theorem in your solution to any problem, it is your responsibility to make clear which theorem you are using and to justify its use.

e. You may use calculators as needed.

---

1. Engineers A and B collected replicated data of the form (x_i, Y_{ij}), $i = 1, \ldots, k$, $j = 1, \ldots, m$. A plotted $\bar{Y}_{i\cdot}$ vs. x_i and claimed that there is a nonlinear relationship between the response variable Y and the control variable x. B used ordinary least squares to fit a linear model $Y = \beta_0 + \beta_1 x + e$ and claimed that a straight line model was adequate to fit the data because he found $R^2 > 0.9$.

(a) Assuming $Y_{ij} = m(x_i) + e_{ij}$ for some function m and that the e_{ij} are i.i.d. $N(0, \sigma^2)$, how would you settle the dispute between A and B? Are either of them using correct reasoning to support their claims?

(b) Suppose that there had been no replication ($m = 1$). What guidance, if any, could you provide to A and B?

2. Let $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be data from a balanced one-way random effects ANOVA, where the a_i are i.i.d. $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$.

In terms of sample averages and statistics calculated in the usual ANOVA table, find $1 - \alpha$ confidence intervals for μ, σ_e^2 and σ_a^2/σ_e^2.

3. A population $\mathcal{U}$ consists of N clusters with M_i elements in the ith cluster. Altogether the population contains $K = \sum_{i=1}^{N} M_i$ elements. A simple random sample $\mathcal{S}$ of n clusters is selected, and a variable y is measured on each element of the selected clusters, yielding data $\{y_{ij},\ i \in \mathcal{S},\ j = 1, \ldots, M_i\}$. Consider the following estimators of the population total $t_y = \sum_{i \in \mathcal{U}} \sum_{j=1}^{M_i} y_{ij} = \sum_{i \in \mathcal{U}} t_i$, where t_i is the ith cluster total:

(i) the simple expansion estimator

$$\hat{t}_1 = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i,$$

(ii) the ratio to size estimator

$$\hat{t}_2 = K \frac{\sum_{i \in \mathcal{S}} t_i}{\sum_{i \in \mathcal{S}} M_i}.$$

(a) Is either of these two estimators unbiased? Explain your answer.

(b) Give expressions for the variances of these estimators, assuming both n and N are large. Your answer should be exact if possible, otherwise approximate. When would one expect $\hat{t}_1$ to be less accurate than $\hat{t}_2$?

4. Random variables Y_{ij}, $1 \le i < j \le 3$, are observed, where the Y_{ij} are independent $N(\beta_i - \beta_j, \sigma^2)$. The parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ and σ^2 are unknown. The parameters β_i can be regarded as effects of a factor B.

(a) Assuming that all three combinations of (i, j) are observed, write a linear model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ to describe the data, where $\mathbf{Y} = (Y_{12}, Y_{13}, Y_{23})^T$. Write the $\mathbf{X}$ matrix explicitly.

(b) Are any of the individual parameters β_i estimable? Is an unbiased estimator of σ^2 available? Prove your answer.

5. Consider the quadratic regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + e$$

where, as usual, the e's are i.i.d. $N(0, \sigma^2)$ random errors.

(a) If the x_{ij} are all ± 1, verify that the parameters β_0 and β_{jj}, $j = 1, 2$, are not estimable.

(b) Suppose that n observations are available for each combination of x values with $x_1 = \pm 1$, $x_2 = \pm 1$ and m additional observations are available at $(x_1, x_2) = (0, 0)$. Show that β_0 and $\beta_0 + \beta_{11} + \beta_{22}$ are estimable, but that β_{11} and β_{22} are not individually estimable.

(c) Propose a test of H_0: $\beta_{11} = \beta_{22} = 0$ and give the distribution of your test statistic under H_0.

6. A simple random sample of households $\mathcal{S}$ is selected from a very large population. The data will be used to estimate the proportion $p_{\mathcal{U}}$ of households with a certain attribute. It is believed that $p_{\mathcal{U}}$ is between 30% and 70%. What sample sizes are needed to meet the following requirements for precision?

(a) The population proportion $p_{\mathcal{U}}$ is to be estimated with a standard error of no more than 3%.

(b) The proportions $p_{\mathcal{U}_k}$ in each of the three income classes—under \$25,000, \$25,000 to \$50,000, and over \$50,000 ($k = 1, 2, 3$, respectively)—are each to be estimated with a standard error of no more than 3%.

(c) The differences of proportions $(p_{\mathcal{U}_j} - p_{\mathcal{U}_k})$ for each pair of classes in (b) are to be estimated with a standard error of no more than 3%.

Income statistics indicate that the proportions in the three classes above are 50%, 40% and 10%.

You should provide separate answers for each of parts (a), (b), (c). The finite population correction may be neglected.