DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY, 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. Let $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be data from a 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. $N(0, \sigma_e^2)$.

(a) Write out the usual ANOVA table and compute the expected mean squares, $E(MS_A)$ and $E(MS_E)$.

(b) Find the distribution of the statistic $F = MS_A/MS_E$ under general conditions.

(c) Find a $1 - \alpha$ confidence interval for the intraclass correlation coefficient

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

2. A questionnaire is to be sent to a sample of high schools to find out which schools provide certain facilities, such as a computer laboratory or a course in Russian. The ith school has an enrollment of M_i students and the total number of students is $K = \sum_{i=1}^{N} M_i$. For a certain facility, it is desired to estimate the proportion of students attending a school with the facility:

$$p_{\mathcal{U}} = \frac{\sum_w M_i}{\sum_{i=1}^{N} M_i},$$

where $\sum_w$ is a sum over the schools *with* the facility.

A sample of n schools is selected *with* replacement and with probability proportional to M_i. For one facility of interest, it was found from the sample that a schools had the facility.

(a) Show that $\hat{p} = a/n$ is an unbiased estimator of $p_{\mathcal{U}}$ and that

$$\text{Var}\,(\hat{p}) = \frac{p_{\mathcal{U}}(1 - p_{\mathcal{U}})}{n}.$$

(b) Show that an unbiased estimator of $\text{Var}\,(\hat{p})$ is

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1 - \hat{p})}{n}.$$

[*Hint:* Let $t_i = M_i$ if the ith school has the facility and 0 otherwise.]

3. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with rank $p \leq n$, $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var-Cov}(\mathbf{e}) = \sigma^2\mathbf{I}$. Let $\boldsymbol{\xi}_i$ denote the ith column of X. Suppose $\{\hat{\beta}_1, \ldots, \hat{\beta}_p\}$ is a set of least squares estimates under the general model. Show that $\{\hat{\beta}_1, \ldots, \hat{\beta}_m\}$, $m < p$ are also least squares estimates under the null hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_p = 0$ if and only if $\boldsymbol{\xi}_i \perp \sum_{j=m+1}^{p} \hat{\beta}_j \boldsymbol{\xi}_j$, $i = 1, \ldots, m$.

4. In an agricultural study, the weight in pounds (Y) and age in weeks (x) were recorded for samples of turkeys selected from three different treatment groups. The following (full) model was fitted to the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_1 z_{ij} + \alpha_2 w_{ij} + e_{ij},$$

where $i = 1, 2, 3$ indexes treatment groups, $j = 1, \ldots, J_i$ indexes turkeys within group, $z_{ij} = I\{i = 1\}$, $w_{ij} = I\{i = 2\}$, and $I\{\cdot\}$ denotes the indicator function of an event. The sample sizes were $J_1 = 4$, $J_2 = 4$, and $J_3 = 5$. Least squares analysis of this model yielded $R^2 = 97.94\%$. By contrast, when the simple linear regression model (reduced model)

$$Y_{ij} = \beta_0^* + \beta_1^* x_{ij} + e_{ij}$$

was fitted to the data, it was found that $R^2 = 64.77\%$.

(a) The experimenters claimed that the large differences in R^2 showed that the treatment differences were significant. Can this statement be verified? If so, calculate an appropriate test statistic and give its distribution under the null hypothesis of no treatment differences. If not, explain why not.

(b) How would you test whether the mean difference between Groups 1 and 2 was nonzero, assuming this comparison had been planned in advance? Would the same testing procedure be used if this comparison was suggested by examination of the data?

5. A stratified population has L strata with N_h units in stratum h. Assume that independent simple random samples of size n_h are selected from stratum h, $h = 1, \ldots, L$. The *combined ratio estimator* of the population total $t_{y\mathcal{U}}$ is

$$\hat{t}_{rc} = t_{x\mathcal{U}} \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h}.$$

Argue that $\hat{t}_{rc}$ is approximately unbiased and derive a formula for its variance when all the n_h are large.

6. Independent observations Y_{ij}, $i = 1, 2$, $j = 1, 2$, were modeled as a two factor ANOVA:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where the e_{ij} are independent random variables with a common $N(0, \sigma^2)$ distribution. Representing the data in vector form, the following decomposition was calculated:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 50 \\ 50 \\ 50 \\ 50 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix} + \begin{bmatrix} 5 \\ -5 \\ 5 \\ -5 \end{bmatrix} + \begin{bmatrix} 3 \\ -3 \\ -3 \\ 3 \end{bmatrix}$$

(a) Compute the ANOVA table for the data.

(b) Compute statistics for testing the hypotheses H_A: no Factor A effect and H_B: no Factor B effect. What are the the distributions of the test statistics under the null hypothesis?

(c) Is there some test of whether this additive model fits this data? Would there exist a test if there had been three levels of Factor A and two levels of Factor B?