DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY, 2004

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. Consider the one-way ANOVA model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ errors. Let $a_1, ..., a_k$ be fixed constants, chosen before the data were observed.

(a) At level α, test H_0: $\sum_{i=1}^{k} a_i\theta_i = 0$ versus H_1: $\sum_{i=1}^{k} a_i\theta_i \neq 0$.

(b) Specialize your test to compare treatment 1 to the average of treatments 2 and 3.

(c) Would your answer to (a) change if the constants $a_1, \ldots, a_k$ had been chosen after looking at the data? If so, how? If not, why not?

2. In a study of which of three machines is preferable for an industrial process, six employees were selected at random and each employee operated each machine three times. The result was Y_{ijk}, the output of the kth run of employee j on machine i, for $i = 1, 2, 3$, $j = 1, \ldots, 6$, $k = 1, 2, 3$. It was assumed that the following linear mixed model described the data:

$$Y_{ijk} = \mu_i + b_j + c_{ij} + e_{ijk},$$

where μ_1, μ_2, μ_3 are fixed but unknown parameters, $b_j \sim N(0, \sigma_b^2)$, $c_{ij} \sim N(0, \sigma_c^2)$, and e_{ijk} $N(0, \sigma_e^2)$. The usual ANOVA table for a balanced two way layout with replication was calculated, yielding the following results.

Source	d.f.	Mean Square	E(MS)
Machines		877.63	
Employees		248.38	
Interaction		42.65	
Residual		0.93	

(a) Find the missing d.f. and E(MS) values, indicating any functions of the fixed parameters implicitly by notation such as $Q(\mu_1, \mu_2, \mu_3)$.

(b) How can one test for differences among machines?

(c) Find a point estimate $\hat{\mu}_i$ for μ_i and find a point estimate for $\operatorname{Var} \hat{\mu}_i$.

3. A surveyor makes a single measurement on each of the angles $\beta_1, \beta_2, \beta_3$ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that the measurements have a common but unknown variance σ^2.

(a) Find the least squares estimates of the unknown angles and their variances.

(b) Is it possible to obtain an unbiased estimate for σ^2? If so, find it, and if not, explain why not.

(c) Suppose it is known in advance that $\beta_1 = \beta_2$. Find the least squares estimates of the unknown angles and their variances in this case.

4. Consider the regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, ..., n$$

where the ϵ_i are independent $\mathrm{N}(0, \sigma^2)$ errors, and $\sum_i x_i = 0$, $\sum_i x_i^2 = 1$. Let $r = \sum_i x_i Y_i$. Suppose we estimate β by

$$\tilde{\beta} = \begin{cases} 0, & \text{if } |r| < c \\ r, & \text{if } |r| \geq c \end{cases}$$

where c is a positive constant.

(a) Compute the MSE's of both the usual LSE $\hat{\beta}$ of β and of $\tilde{\beta}$.

(b) Verify that $\mathrm{MSE}(\tilde{\beta}) < \mathrm{MSE}(\hat{\beta})$ when $\beta = 0$.

5. A simple random sample of n elements is selected from a population $\mathcal{U}$ of N elements, and variables x and y are measured on each element in the sample. The regression estimator of the population mean $\bar{y}_{\mathcal{U}}$ is

$$\hat{\bar{y}}_{\text{reg}} = \bar{y}_{\mathcal{S}} + \hat{B}_1(\bar{x}_{\mathcal{U}} - \bar{x}_{\mathcal{S}}),$$

where

$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{S}}(x_i - \bar{x}_{\mathcal{S}})(y_i - \bar{y}_{\mathcal{S}})}{\sum_{i \in \mathcal{S}}(x_i - \bar{x}_{\mathcal{S}})^2}$$

is the sample least squares estimator of the population least squares regression slope

$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{U}}(x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{\sum_{i \in \mathcal{U}}(x_i - \bar{x}_{\mathcal{U}})^2}.$$

Show that the bias of $\hat{\bar{y}}_{\text{reg}}$ is $-\mathrm{Cov}(\hat{B}_1, \bar{x}_{\mathcal{S}})$ and find a large sample approximation for the mean squared error of $\hat{\bar{y}}_{\text{reg}}$.

6. A simple random sample $\mathcal{S}$ of size $n = n_1 + n_2$ is drawn from a finite population $\mathcal{U}$, and a simple random subsample $\mathcal{S}_1$ of size n_1 is drawn from $\mathcal{S}$. Define the sample and subsample means by

$$\bar{y} = \frac{1}{n}\sum_{\mathcal{S}} y_i \quad \text{and} \quad \bar{y}_1 = \frac{1}{n_1}\sum_{\mathcal{S}_1} y_i,$$

and let

$$\bar{y}_2 = \frac{1}{n_2}\sum_{\mathcal{S}\setminus\mathcal{S}_1} y_i = \frac{n\bar{y} - n_1\bar{y}_1}{n_2}$$

be the mean of the elements of the sample not included in the subsample.

(a) Prove that $\operatorname{Var}(\bar{y}_1 - \bar{y}_2) = S^2_{y\mathcal{U}}(1/n_1 + 1/n_2)$.

(b) Prove that $\operatorname{Var}(\bar{y} - \bar{y}_1) = S^2_{y\mathcal{U}}(1/n_1 - 1/n)$.

(c) Prove that $\operatorname{Cov}(\bar{y}, \bar{y}_1 - \bar{y}) = 0$.