Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 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. The model $Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_i) + e_{ij}$ is estimated from data $(x_{ij}, Y_{ij})$, $i = 1, 2, j = 1, \ldots, n_i$ where $n_1, n_2 > 2$ and the $e_{ij}$ are i.i.d. $N(0, \sigma^2)$

(a) Compute the least squares estimates of the mean parameters and the usual unbiased estimator of $\sigma^2$.

(b) Test the hypothesis $H_0: \beta_1 = \beta_2$. What is the distribution of your test statistic under $H_0$?

(c) Assume that it is known that $\beta_1 = \beta_2$. Under this assumption, how would you test $H_0^*: \alpha_1 = \alpha_2$? Give the distribution of your test statistic assuming $H_0^*$ holds and that the slopes $\beta_i$ are known to be equal.
2. The main effect model \( Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}, \ i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K, \) is fitted to a data set with no replication.

(a) What side conditions would you use to derive a system of solutions to the normal equations?

(b) Write out the ANOVA table, including sums of squares, degrees of freedom and expected mean squares.

(c) What is the power of the usual test of \( H_A: \alpha_1 = \cdots = \alpha_I \) against the alternative \( \alpha_1 = \sigma, \ \alpha_2 = -\sigma, \ \alpha_3 = \cdots = \alpha_I = 0. \) Express your answer in terms of an appropriate probability distribution.

3. Consider the linear model \( Y = X\beta + e \) where

\[
X = \begin{bmatrix}
1 & 1 & -1 & -1 \\
1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1
\end{bmatrix}
\]

and \( \beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix} \)

(a) Show that all parameters are estimable. Justify your answer.

(b) Derive the least squares equations, find \( \hat{\beta}_3 \) and compute its variance.

(c) If an additional term \( \beta_1 x_1 x_2 \) is added to the model, are the parameters still estimable?

4. Consider the one way mixed model \( Y_{ij} = \mu + a_i + e_{ij}, \ i = 1, \ldots, I, j = 1, \ldots, n. \) Assume that \( a_i \sim N(0, \sigma_a^2), e_{ij} \sim N(0, \sigma_e^2) \) and all random terms are mutually independent.

(a) Write out the ANOVA table, showing sums of squares, degrees of freedom and expected mean squares.

(b) Derive the variances and covariances of the \( Y_{ij}. \)

(c) Find a \( 1 - \alpha \) confidence interval for \( \sigma_a^2/\sigma_e^2. \)
5. A population $\mathcal{U}$ is divided into strata $\mathcal{U}_h$ of sizes $N_h$, $h = 1, \ldots, H$. A stratified sample $\mathcal{S}$ of size $n$ is selected by independently drawing simple random samples $\mathcal{S}_h$ of $n_h$ units from stratum $h$. The goal is to estimate the population total

$$t_y = \sum_{h=1}^{H} \sum_{i \in \mathcal{U}_h} y_{hi}.$$ 

(a) Show that

$$\hat{t}_{st} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in \mathcal{S}_h} y_{hi}$$

is an unbiased estimator of $t_y$ and derive its variance in terms of the within stratum means and variances

$$\bar{y}_{th} = \frac{1}{N_h} \sum_{i \in \mathcal{U}_h} y_{hi}, \quad S_h^2 = \frac{1}{N_h - 1} \sum_{i \in \mathcal{U}_h} (y_{hi} - \bar{y}_{th})^2.$$ 

You may neglect the finite sample correction.

(b) Assume that $n_h$ is proportional to $N_h$. Show that in general the stratified estimator $\hat{t}_{st}$ is more efficient than the usual unbiased estimate $\hat{t}_{sr}$ of $t_y$ based on a simple random sample of size $n$ which ignores the stratification of $\mathcal{U}$.

6. Let $Y_i = \beta_0 + \beta_1 x_i + e_i$, where the $x_i$ are known constants and the $e_i$ are i.i.d. $N(0, \sigma^2)$

(a) Find a $1 - \alpha$ confidence interval for $E[Y_0]$, where $Y_0$ is the response corresponding to a new predictor $x_0$.

(b) Suppose that $Y_0$ is observed and the goal is to make inference on the corresponding unobserved $x_0$. Starting from the distribution of $Y_0$, solve a quadratic inequality to find a $1 - \alpha$ confidence set for $x_0$. Is your confidence set an interval?