Applied Statistics (Ph.D. and M.A. Version)

Instructions to the Student

a. Answer any 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 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 two way balanced ANOVA model

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} = \mu_{ij} + e_{ijk}, \]

where \( i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K, \) and the \( e_{ijk} \) are i.i.d. \( N(0, \sigma^2) \). There are no side conditions on the parameters.

(a) Show that the contrast \( \psi = \sum c_i \alpha_i, \ \sum c_i = 0 \), is not estimable.

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

(c) What null hypothesis is being tested by the statistic \( F = MSA/MSE? \) Express your answer in terms of the cell means.

(d) Suppose now that the interaction parameters \( \gamma_{ij} \) are replaced by independent random effects \( c_{ij} \sim N(0, \sigma_c^2) \). Find a confidence interval for \( \alpha_1 - \alpha_2 \).
2. In a $2 \times 2$ table with cell frequencies $Y_{ij}$ and cell probabilities $\pi_{ij}$, $i = 1, 2$, $j = 1, 2$, consider the null hypothesis $H_0$: $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1 - \theta)$, $\pi_{22} = (1 - \theta)^2$.

(a) Show that the row and column marginal distributions are identical and that the row and column classifications are independent under $H_0$.

(b) A multinomial sample of size $n$ and probabilities $\pi_{ij}$ is taken. Under $H_0$ show that $\hat{\theta} = (p_{1+} + p_{+1})/2$, where $p_{ij} = Y_{ij}/n$, $p_{1+} = p_{11} + p_{12}$ and $p_{+1} = p_{11} + p_{21}$.

(c) Explain how to test $H_0$. Show that the test statistic has 2 degrees of freedom.

3. A simple random sample $S$ of size $n$ was drawn from a very large population $U$ of size $N$ with the goal of estimating $t_y = \sum_U y_i$. The data consisted of pairs $(x_i, y_i)$, $i \in S$. The population total $t_x$ is known. The sampler planned to estimate $t_y$ by $\hat{t}_R = \hat{B} t_x$, where $\hat{B} = \sum_S y_i / \sum_S x_i$ is the sample ratio.

(a) Find an estimator of $\text{Var}[\hat{t}_R]$ which is approximately unbiased for large $n$.

(b) Using a Taylor expansion, find a large sample approximation for the bias of $\hat{t}_R$.

4. Let $Y_i$, $i = 1, \ldots, n$, have density (or probability mass function)

$$f(y; \theta_i) = \exp(y \theta_i + b(\theta_i) + c(y))$$

and suppose that the $Y_i$ are independent and obey a generalized linear model with link function $g(\mu_i)$, linear predictor $\eta_i = x_i^T \beta$ and variance function $V(\mu_i)$.

(a) Prove that the maximum likelihood equations can be written as

$$\sum_{i=1}^n \frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0.$$

(b) Show that these equations arise from the generalized least squares problem of minimizing $\sum_i [(Y_i - \mu_i)^2 / \text{Var}(Y_i)]$ with the variances treated as known.
5. Based on data \((Y_i, x_{i1}, \ldots, x_{ik}), \ i = 1, \ldots, n\), consider the full rank multiple regression model

\[ Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i = X\beta + e. \]

Define the hat matrix \(H = X(X^TX)^{-1}X^T\) and let \(h_{ij}\) denote the \((i,j)\) entry of \(H\). Let \(\hat{Y}_i\) be the least squares estimate of \(E[Y_i]\).

(a) Prove that \(H\) is an orthogonal projection matrix. Into what space does it project?

(b) Prove that \(0 \leq h_{ii} \leq 1\) and calculate \(\sum_i h_{ii}\).

(c) Suppose that some \(h_{ii}\) is very close to 1. What can be said about \(Y_i\) and the \(i\)-th residual \(Y_i - \hat{Y}_i\)? Justify your answer.

6. Consider the regression model \(Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij} + e_i\) where

\[
X^T = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 0 & 0 \\
-1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 0 & 0 \\
-1 & -1 & -1 & -1 & 1 & 1 & 1 & 0 & 0 & 0
\end{bmatrix}
\]

Assume the \(e_i\) are i.i.d. \(N(0, \sigma^2)\). The index \(j, j = 1, 2, 3, 4\) corresponds to two level experimental factors A, B, C and D, respectively.

(a) Derive the normal equations and calculate \(\hat{\beta}_0\) and \(\hat{\beta}_1\).

(b) How many degrees of freedom are available to estimate \(\sigma^2\)?

(c) Can the experimenter also estimate two factor interactions of A, B, C and/or D? Explain.
7. A population $U$ of size $N$ is divided into strata $U_h$ of sizes $N_h$, $h = 1, \ldots, H$. Let $W_h = N_h/N$ be the stratum weights. It is desired to estimate the population mean

$$\bar{y}_U = \frac{1}{N} \sum_{h=1}^{H} \sum_{i \in U_h} y_{hi}$$

based on a stratified simple random sample $S = \{y_{hi} | h = 1, \ldots, H; i \in S_h\}$, where $S_h$ is a simple random sample of size $n_h = nW_h$ drawn from $U_h$. The statistic

$$\bar{y}_{st} = \sum_{h} W_h \bar{y}_h = \sum_{h} W_h (1/n_h) \sum_{i \in S_h} y_{hi}$$

will be used to estimate $\bar{y}_U$.

(a) Prove that $\bar{y}_{st}$ is unbiased and calculate its variance in terms of the stratum variances $S_h^2 = (N_h - 1)^{-1} \sum_{U_h} (y_{hi} - \bar{y}_U)^2$.

(b) Assuming that $n_h = nW_h$, choose $n$ so that $P[|\bar{y}_{st} - \bar{y}_U| < \varepsilon] \geq 1 - \alpha$.

You may assume $n$, $N_h$ and $N$ are all large and you may neglect finite population corrections.

8. Let $Y_{ij} = \mu_i + e_{ij}, i = 1, \ldots, I; j = 1, \ldots, n_i$. Let the $x_i$ be known covariates observed without error and let the $e_{ij}$ be i.i.d. $N(0, \sigma^2)$. Assuming $n_i > 0$ for some $i$, show how to test $H_0$: $\mu_i = \beta_0 + \beta_1 x_i$ for some unknown constants $\beta_0$ and $\beta_1$ against the general alternative. Give a formula for the test statistic and its distribution under the null hypothesis.