Design of Experiment: Prelim Problems
August 2023

Please Do All Problems. Each of the 15 parts carries an equal weight of 10 points.

1. A Wisconsin family farm contacted a statistician to conduct an experiment to assess the effect of diet on milk production for a new breed of goats they had acquired. Four diets (factor A) were chosen to reflect typical goat diets provided at farms in the area. The response variable is milk production. Assume factor A has fixed effect. There are no interactions. Please write down the statistical models and assumptions for each of the following designs.

   a) 16 goats were used from a flock and marked by a numeric ear tag. Each diet was applied to 4 goats selected at random. Milk production for each goat was measured.

   b) 16 goats were used but they came from 4 different flocks (factor B) of size 4 each. Within each flock, each diet was assigned to one goat selected at random. Milk production for each goat was measured. Assume factor B has random effect.

   c) Four goats (factor B) were selected at random from a flock and marked by a numeric ear tag. Each goat is given a different diet in random order during each of four consecutive lactation periods (factor C), and we will assume that the period between diets is sufficient so previous diets do not affect later milk production. Milk production for each goat after using each diet was measured. Assume factor B and C both have random effects.

2. With increased demand for Chinese medicinal plants in the U.S., a desire for locally produced, high quality plant material is increasing, yet little is known about the feasibility of production of these species outside of China. The purpose of this study was to develop basic agronomic data for cultivation of selected Chinese medicinal plant species in northeastern United States, including the appropriate level of nitrogen for use on these crops. Three plant species, *Agastache rugosa*, *Schizonepeta tenuifolia*, and *Leonurus japonicus*, were grown at a University Research Farm with either 0, 50, or 100 kg/ha of nitrogen supplied as soy bean meal. Six small fields were used in this experiment. Two of the 6 small fields were randomly assigned and entirely planted with one of the three nitrogen levels. Each field was then divided into three equal parts and each part was randomly assigned one of three plant varieties. The 4 center-most plants in each 1/3 field part were harvested at the appropriate time of year. *Agastache rugosa* plants were harvested when plants were in full flower. *Schizonepeta tenuifolia* were harvested when plants were in full bloom. *Leonurus japonicus* were harvested in the fall, prior to frost. The dry weight of each of the 4 harvested plants were measured. Of interest are differences in nitrogen (factor A) and variety (factor B) levels. Denote y_{iujt} as the dry weight for the t^{th} harvested plant from i^{th} nitrogen level, u^{th} assigned small field, j^{th} plant species. Assume there is no interaction, variety has fixed effects, and nitrogen has random effects.

   a) Please write down a model for this experiment. Clearly define all the terms and state all relevant assumptions.

   b) Fill in the ANOVA table below.

Source	d.f.	MS	E(MS)
Nitrogen		0.6135	
Error(w)		0.128	
Total(w)		0.126	
Variety		0.714	
Error(S)		0.0173	
Total(S)			

   c) State and conduct statistical tests for Nitrogen effects at 0.05 significance level, and interpret the test results

   d) State and conduct statistical tests for Variety effects at 0.05 significance level, and interpret the test results.

   e) Calculate E(Mean Square for Nitrogen) and E(MSE(W)).

   f) Construct 95% confidence interval for the variance component for Nitrogen.

   g) For the same experiment, if instead of measuring the dry weight of each of the 4 harvested plants in each 1/3 field part, we measure the average dry weight of the four harvested plants in the 1/3 field part. Will you change your answer to 2f)? Please explain why.

3. An industrial engineer is studying the manual assembly of electronic components on circuit boards. The goal is to improve the speed of the assembly operation. The engineer has designed three assembly fixtures and two workplace layouts that seem promising. Operators are required to manually perform the assembly, and it is decided to randomly select four operators for each fixture and layout combination. Because the workplaces are in different locations within the plant, it is difficult to use the same four operators for each layout. Therefore, the four operators chosen for layout 1 are different individuals than the four operators chosen for layout 2. The eight operators were chosen at random. The data collection was completely randomized and two replicates were obtained. The assembly times were measured in minutes.

   Denote y_{ijkl} as the assembly l^{th} time for the k^{th} operator using j^{th} workplace layout and i^{th} assembly fixtures. Assume assembly fixtures (factor A) and workplace layout (factor B) both have fixed effect. Operator (factor C) has random effect. Assume there are interaction effects between assembly fixtures and workplace layouts and between assembly fixtures and operators. We have the information $\bar{y}_{1...} = 25.25$, $\bar{y}_{2...} = 28.56$, and $\bar{y}_{3...} = 25.063$.

   a) Write a linear model equation appropriate for analyzing the responses Y_{ijkl} from this experiment, clearly state the assumptions and conditions for this model.

   b) Fill in the ANOVA table below.

Source	d.f.	MS	E(MS)
A		61.92	
B		12	
	6	10.042	
AB		3.063	
		3.125	
Error		2.31	
Total			

   c) Use Tukey's method to construct a 95% simultaneous confidence intervals for the true pairwise differences of the mean assembly times for different assembly fixtures.

   d) State and conduct statistical tests for workplace effects at 0.05 significance level, and interpret the test results.

   e) How many replicates are needed to make the width of confidence interval in 3c) at most 1 minute. Please write down the formula and explain your calculation method.

Applied Statistics Preliminary Examination
Theory of Linear Models
August 2023

Instructions:

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. Show all your work, clearly stating any theorem or fact that you use. Each of the 13 parts carries an equal weight of 10 points.
- Abbreviations/Acronyms.
  - IID (independent and identically distributed).
  - SSE (sum of squared errors). Also called *residual sum of squares.*
  - LSE (least squares estimator); BLUE (best linear unbiased estimator). Sometimes the LSE may be designated OLS (ordinary least squares) estimator, in order to differentiate it from the GLS (generalized least squares) estimator.
- Notation.
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\text{tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$: denotes the trace and determinant, respectively, of matrix $\boldsymbol{A}$.
  - $\boldsymbol{I}_n$: the $n \times n$ identity matrix.
  - $\boldsymbol{j}_n = (1, \ldots, 1)^T$ is an n-vector of ones, and $\boldsymbol{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}(\boldsymbol{x})$ and $\mathbb{V}(\boldsymbol{x})$: expectation and variance of random vector $\boldsymbol{x}$.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m-dimensional random vector $\boldsymbol{x}$ has a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - $X \sim t(n, \lambda)$: a t distribution with n degrees of freedom and noncentrality parameter λ. If $\lambda = 0$ we write simply: $X \sim t(n)$.
  - $X \sim F(n_1, n_2, \lambda)$: an F distribution with n_1 and n_2 numerator and denominator degrees of freedom respectively, and noncentrality parameter λ. If $\lambda = 0$ we write simply: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - Note the inverse for the patterned matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix} \qquad \Longrightarrow \qquad \boldsymbol{A}^{-1} = \frac{1}{3} \begin{bmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}.$$

1. An experiment was conducted to compare the yields for three varieties of corn (1, 2, and 3). Three plants of each variety were grown and the yield recorded, for a total of 9 observations, $\boldsymbol{y} = (y_1, \ldots, y_9)^T$, where $\{y_1, y_2, y_3\}$ correspond to variety 1, $\{y_4, y_5, y_6\}$ variety 2, etc. In order to compare the mean yields of each variety, three linear models of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ were proposed, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_9)$. Letting $\{\mu_1, \mu_2, \mu_3\}$ be the true mean yields for varieties 1, 2, and 3, respectively, and defining the 9-dimensional vectors $\boldsymbol{j}_9$ (vector of 9 ones), $\boldsymbol{x}_1 = (\boldsymbol{j}_3, \boldsymbol{0}_6)^T$, $\boldsymbol{x}_2 = (\boldsymbol{0}_3, \boldsymbol{j}_3, \boldsymbol{0}_3)^T$, $\boldsymbol{x}_3 = (\boldsymbol{0}_6, \boldsymbol{j}_3)^T$ the models are as follows.

Model A (cell means): $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$ and $\boldsymbol{\beta} = (\mu_1, \mu_2, \mu_3)^T$, which corresponds to:

$$y_i = \begin{cases} \mu_1 + \epsilon_i, & i = 1, 2, 3 \\ \mu_2 + \epsilon_i, & i = 4, 5, 6 \\ \mu_3 + \epsilon_i, & i = 7, 8, 9 \end{cases}$$

Model B (reference cell mean): $\boldsymbol{X} = [\boldsymbol{j}_9, \boldsymbol{x}_2, \boldsymbol{x}_3]$ and $\boldsymbol{\beta} = (\mu, \alpha_2, \alpha_3)^T$, which corresponds to:

$$y_i = \begin{cases} \mu + \epsilon_i, & i = 1, 2, 3 \\ \mu + \alpha_2 + \epsilon_i, & i = 4, 5, 6 \\ \mu + \alpha_3 + \epsilon_i, & i = 7, 8, 9 \end{cases}$$

Model C (effects): $\boldsymbol{X} = [\boldsymbol{j}_9, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$ and $\boldsymbol{\beta} = (\mu, \beta_1, \beta_2, \beta_3)^T$, which corresponds to:

$$y_i = \begin{cases} \mu + \beta_1 + \epsilon_i, & i = 1, 2, 3 \\ \mu + \beta_2 + \epsilon_i, & i = 4, 5, 6 \\ \mu + \beta_3 + \epsilon_i, & i = 7, 8, 9 \end{cases}$$

The main goal of this Problem is to make inference on the contrasts, $\eta_1 = \mu_1 - \mu_2$ and $\eta_2 = \mu_1 + \mu_2 - 2\mu_3$, and to determine if these inferences depend on which model is fitted. To this end, let $\widehat{\boldsymbol{\beta}}$ denote the LSE of $\boldsymbol{\beta}$ in each respective model, and $\overline{y}_1 = (y_1 + y_2 + y_3)/3$, $\overline{y}_2 = (y_4 + y_5 + y_6)/3$, and $\overline{y}_3 = (y_7 + y_8 + y_9)/3$, be the sample means corresponding to each of varieties 1, 2, and 3.

(a) Explain the meaning of each element of the parameter vector $\boldsymbol{\beta}$ in each of Models A and B. What is the relationship between these elements in each respective Model? (E.g., μ_1 is the mean of variety 1 in Model A; what parameter(s) does it correspond to in Model B?)

(b) Compute the BLUEs of η_1 and η_2. Do the results depend on which of Models A or B is used? Why? (Note the "possibly useful result" on page 1.)

(c) Compute SSE, and verify that it is the same regardless of which of Models A or B is used.

(d) Construct a $(1 - \alpha)100\%$ confidence interval for η_1 in the context of Models A and B, and show that the intervals are identical.

(e) Construct a level α test for $H_0 : \eta_2 = 0$ in the context of Models A and B, and show that the tests are identical.

2. This Problem continues the analysis of Problem 1, but in the context of fitting the over-parametrized Model C to the vector of 9 observations, $\boldsymbol{y}$.

   (a) Characterize *all* the estimable functions of the type $\boldsymbol{\lambda}^T\boldsymbol{\beta}$. Express the contrasts η_1 and η_2 in terms of the Model C parameters, and hence show that they are estimable.

   (b) Are the BLUEs of η_1 and η_2 the same as computed earlier in the context of Models A and B? Either compute them, or give a solid argument for your reasoning.

   (c) Show that SSE is the same as for Models A and B, by either computing it, or by giving a solid reason why this should be the case.

   (d) Construct a level α test for $H_0 : \eta_2 = 0$. Is the test the same as computed earlier in the context of Models A and B?

   (e) Is it possible to remove the rank-defficiency in Model C? Justify your answer, and if so, show explicitly how this can be done using one of the two methods available.

3. Let $\boldsymbol{A}$ be an $m \times n$ non-zero matrix of rank r. Let $\boldsymbol{B}$ and $\boldsymbol{K}$ be nonsingular matrices (of orders m and n, respectively) such that

$$\boldsymbol{A} = \boldsymbol{B} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{K}.$$

   Consider the matrix $\boldsymbol{G}$ defined as:

$$\boldsymbol{G} = \boldsymbol{K}^{-1} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix} \boldsymbol{B}^{-1},$$

   for some $r \times (m-r)$ matrix $\boldsymbol{U}$, $(n-r) \times r$ matrix $\boldsymbol{V}$, and $(n-r) \times (m-r)$ matrix $\boldsymbol{W}$.

   (a) Show that $\boldsymbol{G}$ is a generalized inverse of $\boldsymbol{A}$.

   (b) Show that if $\boldsymbol{H}$ is a generalized inverse of $\boldsymbol{A}$, then it must be of the same form as $\boldsymbol{G}$ for some choice of the matrices $\{\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}\}$.

   (c) Show that different choices for the matrices $\{\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}\}$ lead to distinct generalized inverses of $\boldsymbol{A}$.