Design of Experiment: Prelim Problems

May 2023

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

1. A fisherman performed an experiment to determine the effect of A: line weight (3 levels: 6-lb, 10-lb and 20-lb test lines), B: pole stiffness (2 levels: medium light weight and medium weight poles), and C: lure weight (2 levels: a lightweight lure and heavier lure) upon the distance he could cast. Suppose factors A and B have fixed effects and factor C has random effects. Assume there are no interaction effects. Please write down the statistical models and assumptions for each of the following designs.

   a) Suppose the fisherman ran a completely randomized design with 2 replicates for each combination of the three factors.

   b) Suppose two fishermen each ran a completely randomized design with 1 replicates for each combination of the three factors. (Assume random fisherman effects)

   c) Suppose the 10-lb weight line was randomly chosen to be used for the first four casts. The order of the four combinations of pole stiffness and lure weight were randomized in the first four trials. Next, the 20-lb weight line was randomly selected to be used for the next four trials and again the order of the four combinations of pole stiffness and lure weight were randomized in these trials. Finally, the last four casts were made using the 6-lb weight line. Then the fisherman repeated the whole experiment again as a second block.

2. The effect of cultural background on group decision making was studied by an experiment. Sixteen teams of students were formed and assigned a task. One of response variables was the number of group interactions prior to the final group decision. Eight teams consisted of foreign students, eight of U.S. students. Half of the teams consisted of eight members, the other half of four members. Two foreign observers were used for the foreign teams, and two U.S. observers for the U.S. teams. Thus the design may be represented as follows:

	U.S. Teams (i=1)		Foreign Teams (i=2)	
Size of Team	Observer 1 (k=1)	Observer 2 (k=2)	Observer 3 (k=1)	Observer 4 (k=2)
4 members	replication 1	replication 1	replication 1	replication 1
(j=1)	replication 2	replication 2	replication 2	replication 2
8 members	replication 1	replication 1	replication 1	replication 1
(j=2)	replication 2	replication 2	replication 2	replication 2

Denote y_{ijkt} as the number of group interactions for the t^{th} sample from i^{th} nationality (factor A, assumed with fixed effect), j^{th} team size (factor B, assumed with fixed effect), and k^{th} observer level within nationality(factor C, assumed with random effect). Assume there are interaction effects between factors A and B and between B and C.

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	df	MS	E(MS)
A	________	2.89	________
B	________	8.03	________
C(A)	________	6.54	________
A*B	________	0.59	________
B*C(A)	________	0.78	________
Error	________	2.31	________
total	________		

c) State and conduct statistical tests for observer effect at 0.05 significance level, and interpret the test results.

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

e) Construct 95% confidence interval for the variance component of the observer effect.

f) If we assume team size has random effect, then will you change your test procedure to question 2d)? If you will, please explain how. It is OK to only show the formula without calculation.

3. An experiment was used to compare performance based on a quality characteristic, of a new and old film type and three manufacturing pressures. One film type was selected at random and it was run through the manufacturing process randomly assigning each pressure to $\frac{1}{3}$rd of the roll. Quality measurements were made on each third of the roll. Next, the other film type was run through the process again randomly assigning each pressure to one-third of the roll.This two-step experimental process was repeated eight different times.

   y_{hij} represents the quality characteristic measurement for ith film type jth manufacturing pressure at h th replicate time. Assume film type (factor A) and manufacturing pressure (factor B) both have fixed effect. Replicates (factor C) has random effect. Assume there are interaction effects between film type and manufacturing pressure. We have some summary information about the data as follows: $\sum_i (\bar{y}_{.i.} - \bar{y}_{...})^2 = 0.119$; $\bar{y}_{..1} = 1.667$; $\bar{y}_{..2} = 1.725$; $\bar{y}_{..3} = 1.588$.

   a) Write a linear model equation appropriate for analyzing the responses y_{hij} 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)
Replicate	______	0.08	______
film type	______	2.86	______
Error(W)	______	0.13	______
Total(W)	______		
manufacturing pressure	______	______	______
______	______	0.15	______
Error(S)	______	0.0173	______
Total(S)	______		

   c) State and conduct statistical tests for Type effect at 0.05 significance level, and interpret the test results.

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

   e) Use Tukey's method to construct simultaneous 95% confidence intervals for the pairwise effect difference of the three pressures.

   f) Calculate E(Mean Square for Type).

Applied Statistics Preliminary Examination
Theory of Linear Models
May 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 12 parts carries an equal weight of 10 points.
- Abbreviations/Acronyms.
  - IID (independent and identically distributed).
  - 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 $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ 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.
  - If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given in partitioned form as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

    with $m_1 = \dim(\boldsymbol{x}_1)$, then the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is

$$\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim N_{m_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

1. This question will review covariance and the associated correlation matrices, and demonstrate some techniques for constructing these from combinations of other covariance and correlation matrices, while ensuring the all-important requirement of positive definiteness. Recall that if $\boldsymbol{\Sigma} > 0$ is a (positive definite) covariance matrix, then the associated correlation matrix is $\boldsymbol{R} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}$, where $\boldsymbol{V}$ is a diagonal matrix consisting of the diagonal entries of $\boldsymbol{\Sigma}$.

   (a) Show that a positive linear combination of positive definite matrices is again positive definite. That is, for positive scalars $\alpha_i > 0$ and positive definite matrices $\boldsymbol{A}_i > 0$, $i = 1, \ldots, k$, show that

   $$\sum_{i=1}^{k} \alpha_i \boldsymbol{A}_i$$

   is positive definite.

   (b) Show that if $\boldsymbol{B}$ is any square matrix and $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T$ is positive definite.

   (c) If $\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2$ are nonzero square matrices, $\boldsymbol{a}_1$ is a nonzero vector, and $\boldsymbol{\Sigma}$ is a positive definite covariance matrix, consider the matrix:

   $$\boldsymbol{B} = \boldsymbol{A}_0\boldsymbol{A}_0^T + \boldsymbol{A}_1(\boldsymbol{a}\boldsymbol{a}^T)\boldsymbol{A}_1^T + \boldsymbol{A}_2\boldsymbol{\Sigma}\boldsymbol{A}_2^T.$$

   Determine, with justification, whether or not $\boldsymbol{B}$ is positive definite.

   (d) Let R_0, R_1, R_2 be correlation matrices, and $\theta_1 \geq 0$ and $\theta_2 \geq 0$ be nonnegative scalars such that $\theta_1 + \theta_2 \leq 1$. Consider the matrix:

   $$\boldsymbol{C} = (1 - \theta_1 - \theta_2)R_0 + \theta_1 R_1 + \theta_2 R_2.$$

   Determine, with justification, whether or not $\boldsymbol{C}$ is a correlation matrix.

2. Consider the linear model

   $$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),$$

   where $\boldsymbol{X}$ is a full-rank $(n \times k)$ model matrix, and, as the notation suggests, the elements ϵ_i of the error vector $\boldsymbol{\epsilon}$ are uncorrelated with mean zero and a common variance of σ^2. Recall that $\hat{\boldsymbol{\mu}} = \boldsymbol{P}\boldsymbol{y} = \hat{\boldsymbol{y}}$ is the usual LSE of $\boldsymbol{\mu}$, where $\boldsymbol{P}$ denotes the projection matrix onto $\mathcal{C}(\boldsymbol{X})$. Define the quantity:

   $$r^* = \mathbb{E}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2.$$

   (This is called the *risk* of the model in estimating $\boldsymbol{\mu}$.) We seek unbiased estimators of r^*. Throughout this Problem (except where indicated), assume that $\boldsymbol{\beta}$ is unknown, but σ^2 is <u>known</u>.

   (a) Show that $r^* = \|(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\mu}\|^2 + k\sigma^2$.

   (b) Show that $\mathbb{E}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = r^* + (n - 2k)\sigma^2$, and hence propose an unbiased estimator of r^*.

   (c) If $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, find the distribution of $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2/\sigma^2$.

   (d) If σ^2 is <u>unknown</u>, propose an unbiased estimator of r^*.

3. Consider the following linear model for the observations $\{y_1, y_2, y_3\}$:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \beta_3 + \epsilon_1, \\ y_2 &= \beta_1 + \beta_3 + \epsilon_2, \\ y_3 &= \beta_2 + \epsilon_3, \end{aligned}$$

where $\{\epsilon_1, \epsilon_2, \epsilon_3\} \sim$ IID $\mathrm{N}(0, \sigma^2)$, and $\{\beta_1, \beta_2, \beta_3\}$ are unknown parameters to be estimated.

(a) Write out the model in the usual matrix form, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Determine the rank of the design matrix $\boldsymbol{X}$, and find a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$.

(b) Determine the form of all *estimable* functions $\eta = \boldsymbol{\lambda}^T\boldsymbol{\beta}$. Use this form to decide which of the following are estimable:

$$\eta_1 = \beta_1, \qquad \eta_2 = \beta_2, \qquad \eta_3 = \beta_3, \qquad \eta_4 = \beta_1 - 2\beta_2 + \beta_3.$$

In particular, show that η_4 is estimable.

(c) Find the BLUE of η_4, and find one other linear unbiased estimator (LUE) of η_4 that is different from the BLUE.

(d) Construct a test of the null hypothesis:

$$H_0 : \eta_2 = 0, \qquad \text{and} \qquad \eta_4 = 3.$$

Clearly define the test statistic, and compute its distribution both under H_0 and under the alternative hypothesis, H_1.