Problem Set 4 (three pages)

Statistics 24510-30040 (W19)

New due date: Thurday, February 7 at the beginning of class.

Reminder: Midterm on Thursday, February 7, 5-7 pm Kent 107

Requirements Provide detailed derivations. Select only the relevant part of the output to be inserted. Attach your code or output as an appendix if necessary. Discussions allowed, the assignment should be devised and written by yourself completely.

Problem assignments (Suggested reading: Rice chapter 12 on ANOVA. Review two-sample tests in chapter 11.)

1. (*Two-sample tests with formulas*) (Based on Problem 22 in Rice p. 463)

   Obtain the data at https://www.stat.uchicago.edu/ meiwang/courses/w19stat2/calcium2.txt.
   Read into R with the command

   ```
   calcium = read.table("calcium2.txt", col.names=c("oxalate","flame"))
   ```

   In this exercise, you may use arithmetic operations in R but you may not use the routine `t.test`. Not yet, though you may use the R routine to check your results. So show your work for credits.

   (a) Assuming independence of the two samples, perform a pooled two-sample t-test for the null hypothesis of equal means (assuming equal variances). Comment on the result.

   (b) Perform a paired t-test and comment on the result.

   (c) Make a scatterplot of the data, that is, plot one variable against the other (use `plot` command in R). Compute the sample correlation coefficient $\hat{\rho}$ and construct a 95% confidence interval for the true correlation coefficient ρ.

   (d) Which test is more appropriate for this situation: the paired or the pooled? Explain.

2. (*Two-sample test using R*) (Based on Problem 35 in Rice p. 465)

   Obtain the data at https://www.stat.uchicago.edu/ meiwang/courses/w19stat2/ozonerats.csv.
   (For some browser the data may download immediately automatically.)

   Read into R with the command

   ```
   ozone = read.csv("ozonerats.csv")
   ```

   In this exercise, you may use the R routine `t.test.`

   (a) Assuming independence of the two samples, perform a pooled two-sample t-test for the null hypothesis of equal means (assuming equal variances). Comment on the result.

   (b) Assuming independence of the two samples, perform a two-sample t-test for the null hypothesis of equal means without assuming equal variances. Comment on the result.

   (c) Which test is more appropriate for this situation: equal variance or unequal variance? Justify.

   (d) Perform a paired t-test and comment on the result.

   (e) Which test is more appropriate for this situation: the paired or the two-sample assuming independence? Explain.

3. (*Hands-on one-way ANOVA*)

The leaves of a type of plants fold and unfold in various light conditions. A sample of 15 different leaves were treated with red light for 3 minutes. The leaves were randomly divided into three groups of five. The leaflet angles were then measured 30, 45 and 60 minutes after light exposure in the three groups. The measurements are shown in the following table.

delay(minutes)	Angles (degrees)				
30	140	138	140	138	142
45	140	150	120	128	130
60	118	130	128	118	118

(a) What are the treatments? What are the experimental units? What are the measurement units?

(b) Calculate the treatment sum of squares SS_T (*a.k.a.* SS_B) and the residual sum of squares SS_E(*a.k.a.* SS_W) . Write out every term of the sum SS_T, at least first two and last two terms in SS_E.

(c) Create the ANOVA table for these data. The columns should include variance source, df, SS, MS, F-statistic, and p-value.

(d) Based on the above, test the null hypothesis that delay after exposure does not affect leaflet angle. (What is your alternative hypothesis?) interpret the result.

4. (*Two-way additive model without replication*)

For the two factor observations in the following table,

β_j

α_i

4.4	4.3	4.0	3.6	3.3
4.1	4.1	4.0	4.0	3.9
3.2	3.9	3.8	4.1	4.4
3.9	4.0	3.8	4.1	4.4

consider the model (i indexes row factor, j column factor)

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2), \qquad j = 1.\cdots, 5, \; i = 1, \cdots, 4.$$

(a) As in Section 12.3.1 in Rice, under the constraints $\sum_{i=1}^{4} \alpha_i = 0, \sum_{j=1}^{5} \beta_j = 0$, use simple averaging operations to compute parameter estimates $\hat{\mu}, \hat{\alpha}_i$ and $\hat{\beta}_j$.

(b) Create the 4-by-5 table of residuals $Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. How does the size of the residuals compare to the size of the estimated row and column effects $\hat{\alpha}_i$ and $\hat{\beta}_j$?

5. (*MLE in one-way ANOVA*)

Consider the one-way ANOVA model parameterized as

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2), \qquad j = 1.\cdots, J_i, \; i = 1, \cdots, I.$$

(a) Write the likelihood function of the model.

(b) Find the maximum likelihood estimates of the parameters $\mu_i \in \mathbb{R}$, $i = 1, \cdots, I$.

(c) Find the maximum likelihood estimates of the parameter $\sigma^2 > 0$.

(d) Derive the likelihood ratio test LR for H_o: $\mu_1 = \cdots = \mu_I$ (default H_a: not H_o).

(e) Compare the LR in (d) to the standard F test for one-way ANOVA. Comment.

(f) Reparameterize the model as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \quad i.i.d \ \sim N(0, \sigma^2), \qquad j = 1.\cdots, J_i, \ i = 1, \cdots, I$$

with the constraint $\sum_{i=1}^{I} \alpha_i = 0$. Find the maximum likelihood estimates of the parameters μ, α_i, σ^2.

(g) For the model in (f), find the MLEs for μ, α_i, σ^2 with the constraint $\alpha_1 = 0$, as in R.

6. (*Additive model and design matrix*)

Consider a study of balanced two-way layout with one replicate: The first factor α has 3 levels and the second factor β has 2 levels, with one observation for each of the 6 possible combinations of the two factors.

(a) Write up the linear model for each observation Y_{ij}, in six linear equations.

(b) The model in (a) can be represented in vector-matrix form $Y = X\theta + \varepsilon$. Write down the full design matrix, the X matrix. clearly defining how the rows and columns are ordered.

(c) What is the rank of your design matrix?

(d) Using the constraints $\alpha_1 = 0, \beta_1 = 0$, obtain the corresponding design matrix X_1.

(e) What is the rank of new design matrix X_1?

(f) Write up the linear model for each observation Y_{ij} under the constraints $\alpha_1 = 0, \beta_1 = 0$.

(g) (Optional) Solve the system of equations in (f) to obtain $\hat{\mu}, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_3$.