STA 106: Applied Statistical Methods: Analysis of Variance

Midterm (Due: November 13th, 5pm)

Instructions

This is an open-book exam. You could refer to the textbook and lecture notes/recordings. However, you should answer the questions independently.

You are not allowed to discuss with others (inside or outside the class) or use other's code. Failing to do so will be counted as cheating. The exam will not be graded, and this incidence will be reported to the Student Judicial Affairs.

Questions

The one-way ANOVA model in the questions refers to the following model:

$$Y_{ij} = \mu_i + \varepsilon_{ij};\ i = 1, \ldots, r; j = 1, \ldots, n_i,$$

with ε_{ij} independent and follow the $N(0, \sigma^2)$ distribution.

Part I: (6 points) Tell true or false of the following statements and justify your answer.

1. In a study of assessing the effectiveness of two surgical procedures, A and B, in treating a disease, 1000 patients were recruited. Among them, 620 patients chose surgical procedure A, and the rest 380 patients chose surgical procedure B. The success rate for surgical procedure A and B are 0.92 and 0.75, respectively. Since the sample sizes are quite large, we conduct the two-sample t-test. The p-value is extremely small, 2.2×10^{-11}. Hence, we could conclude that procedure A is a better procedure in treating the disease.

2. In the one-way ANOVA model, the 99% confidence interval of $\mu_1 - \mu_2$ based on the Tukey procedure is [-6.3, -2.3]. Then μ_1 must be less than μ_2.

3. In the one-way ANOVA model, since ε_{11} and ε_{12} are independent, the residuals e_{11} and e_{12} are independent.

Part II: (16 points) Data analysis.

A company owing a large fleet of trucks wishes to determine whether or not four different brands of snow tires have the same mean tread life (in thousands of miles). The data file can be downloaded here. In this problem, you may use the following existing functions in R (read.table, rep, length, levels, as.factor, which, which.max, which.min, data.frame, c, sum, sqrt, mean, median, plot, abline, axis, lines, hist, qqnorm, qqline, pt, qt, pf, qf, ptukey, qtukey, points, list) but not others.

1. Write functions: List the functions that will be used for #2 here, and explain these functions for potential users of your code. This question will **not** be graded by the instructors, but will

be considered when nominating outstanding reports. You can also choose not to write functions but answer the questions in #2 using codes directly.

2. Data analysis.

   **For each of the following hypothesis test, you need to state the null and alternative hypotheses, state what testing procedure you decide to use if there are more than one choice and explain why the particular procedure is chosen, compute the test statistic, state the decision rule and make your conclusion.*

   (a) *Data visualization and model diagnostics.*

      (i) (1 pt) Plot the aligned dot plot and comment on what you see.

      (ii) (2 pt) Plot the residual plot and perform a proper test to check the equal variance assumption for the one-way ANOVA model. What's your conclusion?

      (iii) (1 pt) Choose an appropriate plot to check the normality assumption. What's your conclusion?

   Assume the one-way ANOVA model for the following questions.

   (b) (3 pt) Test whether the mean tread life is the same for all 4 brands at 0.05 significance level.

   (c) (3 pt) Brands 1 and 2 are national brands and brands 3 and 4 are local brands. Suppose before seeing the results, you only plan to test all the pairwise comparisons between a local brand and a national brand and whether three more contrasts ($L_1 = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$, $L_2 = \mu_3 - \frac{\mu_1+\mu_2}{2}$, $L_3 = \mu_4 - \frac{\mu_1+\mu_2+\mu_3}{3}$) equal to 0 or not. Conduct the test with the familywise error rate controlled at 0.1.

   (d) (3 pt) You feel the testing results in (c) is not enough and would like to go ahead to construct confidence intervals for these pairwise comparisons and contrasts. However, after seeing the results, you are no longer interested in L_1, L_2 and L_3. So you are going to only construct confidence intervals for all the pairwise comparisons between a local brand and a national brand. Construct the simultaneous confidence intervals for these pairwise comparisons with family confidence coefficient 0.95. You need to explain clearly why a particular procedure is chosen for constructing the confidence intervals and compute out these intervals. Briefly comment on the confidence intervals.

   (e) (3 pt) In a follow-up study, there are other three brands of snow tires of interest. It is important to conclude that the three brands of tires have different mean tread lives when the difference between the means of the best and worst brand is 1 (thousand miles) or more. We use σ estimated from the current study. How would you plan for the sample sizes so that the type I error controlled at 0.05 and type II error controlled at 0.2?