MA409: Statistical Data Analysis (SAS)

Assignment 4

Note: **Please work on problem 2 by hand calculation** and 3-4 by SAS procedures. Please provide the SAS code in a separate ***.sas file*** and the outputs from SAS (using screenshots) together with problems 1, 2 in a ***PDF file***.

1. Under the two-way ANOVA model, $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, the variation between groups is defined as:

$$SSM = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\bar{Y}_{ij} - \bar{Y}\right)^2.$$

The variation between groups due to factor A, B, and interaction of A and B are defined as:

$$SSA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{Y}_{i\cdot} - \bar{Y})^2, SSB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\bar{Y}_{\cdot j} - \bar{Y}\right)^2,$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}\right)^2.$$

Show that $SSM = SSA + SSB + SSAB$ when the n_{ij}'s are all equal. (10 points)

2. An introductory statistics course has four tutorial classes. All classes used the same textbook, taught by the same lecturer and had the same examination paper but taught by different teaching assistants. The examination scores for random samples of students independently draw from the four classes are given in the following table:

Teaching Assistants			
Amber	Bell	Chris	Daniel
80	80	50	70
40	60	60	80
50	80	90	85
35	90	40	75
55	75	70	90
70	55	80	60
	50	60	100
		70	

Let $(\bar{Y}_A, S_A^2, \mu_A, \sigma_A^2)$, $(\bar{Y}_B, S_B^2, \mu_B, \sigma_B^2)$, $(\bar{Y}_C, S_C^2, \mu_C, \sigma_C^2)$, $(\bar{Y}_D, S_D^2, \mu_D, \sigma_D^2)$ denote the sample mean, sample variance, population mean, population variance of the examination scores of the students from the four classes, respectively.

(1) Compute $\bar{Y}_A$，$\bar{Y}_B$，$\bar{Y}_C$，$\bar{Y}_D$，S_A^2，S_B^2，S_C^2，S_D^2，and the overall mean $\bar{Y}$, variation between groups SSB, variation within groups SSW. Construct the ANOVA table based on the data. Test $H_0: \mu_A = \mu_B = \mu_C = \mu_D$ vs. $H_1: \mu_A, \mu_B, \mu_C, \mu_D$ are not all equal at significance level $\alpha = 0.1$. (10 points)

(2) Apply Levene's test with $Z_{ij} = \left(Y_{ij} - \bar{Y}_i\right)^2$ to test equality of group variances at $\alpha = 0.1$:

$$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 \text{ vs. } H_1: \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2 \text{ are not all equal.}$$

Please provide the detailed steps of the test, only providing the final result is not acceptable. (10 points)

(3) Apply the Kruskal-Wallis test to test the equality of group medians at $\alpha = 0.05$. Please provide the detailed steps of the test, only providing the final result is not acceptable. (10 points)

3. A production manager who supervises an assembly operation wants to investigate the effect of the arrival rate of components (parts per minute) X_1 and room temperature X_2 (°F) on the productivity (number of items produced per minute) Y. It is thought that an increase in the arrival rate of components has a positive effect on the production rate, up to a point, after which increases may annoy the workers and reduce productivity. Similarly, it is believed that lowering the room temperature is beneficial to a point, after which reductions may reduce productivity. 39 workers were randomly selected and assigned to each of the 16 factor level combinations of a 4×4 factorial experiment. The response Y (productivity averaged over a 5-minute period) are shown in the following table. Use $\alpha = 0.05$.

X_2	X_1			
	40	**50**	**60**	**70**
65°F	24.0 23.8 23.6	25.6 25.4	29.2 29.4 29.0	28.4 27.6
70°F	25.0 26.0	28.8 28.8 28.4	31.6 32.0 32.2	30.2 30.0
75°F	25.6 25.0	27.6 28.0 27.8	29.8 28.6	28.0 27.0 27.4
80°F	24.0 24.2 24.6	27.6 26.2	27.6 28.6	26.0 24.4

(1) Do the data provide sufficient evidence to indicate differences in worker productivity under the 4 levels of component arrival rate? (5 points)

(2) Do the data provide sufficient evidence to indicate differences in worker productivity under the 4 levels of room temperature? (5 points)

(3) Let $\mu_A, \mu_B, \mu_C, \mu_D$ be the mean worker productivity under room temperature 65°F, 70°F, 75°F, and 80°F, respectively. Compute the simultaneous confidence interval of $\mu_A - \mu_B$，$\mu_A - \mu_C$，$\mu_A - \mu_D$，$\mu_B - \mu_C$，$\mu_B - \mu_D$，$\mu_C - \mu_D$ using the Tukey-Kramer method. Do the results indicate significant difference between any pair of room temperature? (5 points)

(4) Do the data provide sufficient evidence to indicate an interaction effect between component arrival rate X_1 and room temperature X_2 on worker productivity? (5 points)

(5) For the two-way ANOVA model with the interaction effect, test whether the residuals follow a normal distribution. (5 points)

(6) Treating X_1 and X_2 as continuous variables, perform a regression analysis. Do you need to add quadratic or cubic terms of X_1 or X_2 into the model? (10 points)

(7) Comparing the two-way ANOVA model and the regression model in (6), which is better? Briefly state your justification. (5 points)

4. This is a game to convince you that unequal variances (heteroscedasticity) would cause problems when using the traditional F-test in one-way ANOVA to compare group means.

(1) Generate 1000 datasets (put all datasets in a single SAS dataset, with a column *Ind* to specify the dataset ID), each dataset contain data generated for three groups A, B, C: $n_A = n_B = n_C = 7$, and $Y_A \sim N(1,1)$, $Y_B \sim N(1,1)$, $Y_C \sim N(1,1)$ (this is a setting with equal variances). Then perform the one-way ANOVA on the 1000 datasets and store the resulting p-values to a SAS dataset. Compute the Type I error rate based on the 1000 p-values ($\alpha = 0.05$) and provide your result. Note: **use seed 12345** to make sure everyone gets the same result. **Hint:** the "*BY Ind*" statement in *PROC GLM* allows you to perform one-way ANOVA for each dataset separately; the "*OUTSTAT =* " option in *PROC GLM* allow you to store the results to a SAS dataset (check here); remember to add the "*NOPRINT*" option in *PROC GLM* to suppress outputs in the *RESULTS* tab of SAS Studio to avoid memory problem. (10 points)

(2) Generate another 1000 datasets with similar procedure as above, but with: $n_A = n_B = n_C = 7$, and $Y_A \sim N(1, 0.5^2)$, $Y_B \sim N(1,1)$, $Y_C \sim N(1, 2^2)$ (this is a setting with unequal variances). Then also perform the one-way ANOVA on the 1000 datasets and compute the Type I error rate based on the resulting 1000 p-values ($\alpha = 0.05$). Compare the result with that in (1) and briefly state your conclusion. (10 points)