ASSIGNMENT 2

ONE-WAY ANALYSIS OF VARIANCE

Analysis of variance is widely used in medical and biological sciences for testing hypotheses about population means and estimating the differences between the means. When the variable of interest is examined for differences by a single factor, the analysis is said to be one-way. Researchers often want to go beyond the conclusion that groups are different, wanting to also know which means differ. This analysis is conducted using a priori contrasts and post-hoc tests. In this lab, you will use analysis of variance tools in SPSS to evaluate the reliability of lipid-bound sialic acid as a marker in breast cancer.

Breast Cancer Detection and Monitoring

In order to determine if the measurement of serum lipid-bound sialic acid (LSA) might be used in detecting and monitoring breast cancer, 1050 cancer patients were recruited from different clinics. The patients were classified into one of three groups using established standards, depending on the stage of the disease:

Benign – Patients with benign breast disease
Primary – Patients with primary breast cancer, and
Recurrent – Patients with recurrent metastatic breast cancer.

Each of the three cancer groups consisted of 350 patients. In addition, the researchers recruited 350 healthy individuals as controls. The LSA measurements (mg/dl) were taken for all the subjects. This dataset is available in the Data link located in the Lab 2 tab display in the Labs section on eClass. The data are not to be printed in your submission.

The following is the description of the variables in the data file:

Column	**Variable Name**	**Description of Variable**
1	ID	Patient's ID,
2	GROUP	Cancer (Benign (2), Primary (3), Recurrent (4)) or Control (1),
3	LSA	LSA value for the patient (in mg/dl).

Use the data set to answer the following questions:

1. The following questions refer to the design of the study:

   (a) Describe the study design briefly. In particular, is this an observational study or experimental study? Why is the healthy control group important for this study?

   (b) Given that the ID numbers assigned to the subjects represent the order in which the measurements were taken, obtain a scatterplot of LSA versus ID number with different markers for each group.

      (i) Does the plot indicate any time trend which would mean the presence of measurement bias?

      (ii) Given the pattern of the plot, suppose some patients with recurrent metastatic breast cancer were wrongly classified as benign or primary, what effect would the misclassification have on the study?

2. Use the *Explore* procedure to obtain the descriptive statistics, the side-by-side boxplots, and the normality plots of the LSA measurements for the four groups:

   (a) Obtain and paste the descriptive statistics for each group into your report. Compare the means and standard deviations of the four groups. Is there any indication that might suggest that individuals with breast cancer have higher LSA measurements than healthy individuals?

   (b) From the SPSS output of part (a), report the 95% confidence intervals for the population means of the four groups. Comment on the precision of the sample means as estimates of the population means. Are any of the intervals overlapping?

   (c) Obtain and paste the side-by-side boxplots into your report. Compare the centers and spreads of the four distributions. Comment on the shape of each distribution. Do you recognize any outliers?

   (d) Obtain and paste the normality plots for the four groups into your report. Are any of the plots suggesting a clear departure from normality?

3. Now you will use the *One-Way ANOVA* tool in SPSS to determine whether LSA measurements for the four groups are different.

   (a) Paste the ANOVA output into your report. Define the appropriate null and alternative hypotheses. Report the sums of squares of residuals from fitting the 4-mean and 1-mean models, the pooled estimate of the variance, the distribution of the test statistic under the null hypothesis, the value of the test statistic, and the *p*-value. State your conclusion using $\alpha = 0.01$.

   (b) Paste the output of Levene's Test into your report. Define the appropriate null and alternative hypotheses. Use the test ("Based on Mean" row of output) to comment on the validity of the F-test in part (a), as well as referring to the boxplots and normality plots of Question 2.

4. Which groups differ in their mean LSA levels? Answer this question by carrying out the Tukey's and Scheffe's multiple-comparison procedures at the 5% significance level. Paste the output of both procedures into your report. Make appropriate conclusion from the results for each of these procedures. Are the results from the two procedures consistent? Or do the results not agree for certain pairs? Summarize which groups differ with a means comparison diagram.

5. Now you will set up contrasts to address specific questions.

   (a) Suppose the question of interest is to determine whether cancer patients have different LSA from healthy individuals. Answer this question by setting up appropriate contrasts in SPSS. State the appropriate hypotheses for the contrast. Paste the related SPSS output into your report and make relevant inference given the output. Use $\alpha = 0.01$.

   (b) Suppose the researchers had been interested in the question "Does LSA measurements in metastatic breast cancer cases differ from LSA levels in benign and primary breast cancer cases?" Answer this question by setting up appropriate contrasts in SPSS. State the appropriate hypotheses for the contrast. Paste the related SPSS output into your report and make relevant inference given the output. Use $\alpha = 0.01$.

6. Summarize your findings from the various analyses you have conducted on this data set. Why would LSA levels not be a very effective breast cancer detection tool? Explain briefly.

LAB ASSIGNMENT 2 MARKING SCHEMA

Question 1 (10)

(a) Description of study and study design: 2 points
Importance of healthy control group: 2 points

(b) Scatterplot of LSA versus ID: 2 points
(i) Comment on measurement bias: 2 points
(ii) Effect of misclassification: 2 points

Question 2 (46)

(a) Descriptive statistics: 2 points each (8 points total)
Comparison of mean and standard deviations: 2 points each (4 points total)
Comparison of LSA levels between cancer patients and healthy control group: 1 point

(b) Confidence intervals: 2 points each (8 points total)
Comment on precision of sample means: 2 points
Overlap: 2 points

(c) Side-by-side boxplots: 4 points
Comparisons of centers and spreads: 2 points each (4 points total)
Shape of distributions: 2 points
Outliers: 1

(d) Normal probability plots: 2 points each (8 points total)
Description of patterns: 2 points

Question 3 (18)

(a) ANOVA output: 2 points
Hypotheses: 2 points
Sum of squares residuals: 1 point each (2 points total)
Pooled estimate of variance: 1 point
Null distribution: 1 point
Test statistic: 1 point
P-value: 1 point
Conclusion: 2 points

(b) Levene's Test output: 2 points
Hypotheses: 2 points
Comment on validity of F-test: 2 points

Question 4 (12)

Output of procedures: 3 points per procedure (6 points total)
Conclusions: 2 points per procedure (4 points total)
Diagram and summary: 2 points

Question 5 (12)

(a) Correct contrast expression and hypotheses: 4 points
P-value and conclusion: 2 points
(b) Correct contrast expression and hypotheses: 4 points
P-value and conclusion: 2 points

Question 6 (4)

Summary of findings: 2 points
Reason why LSA levels would not be effective for breast cancer detection: 2 points

TOTAL = 102