LAB ASSIGNMENT 2

ONE-WAY ANALYSIS OF VARIANCE

The One-Way Analysis procedure in SPSS produces a one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable. Analysis of variance is used to test the hypothesis that several means are equal. In addition to determining that differences exist among the means, you may want to know which means differ. This analysis can be conducted by priori contrasts (set up before running the experiment) and post-hoc tests run after the experiment. The ANOVA inferences are valid if the assumption of normality is not seriously violated for each of the groups. Moreover, the groups should come from populations with equal variances.

In this assignment you will analyze the results of an experiment to compare survival times of patients with different primary cancers. You will use analysis of variance tools in SPSS to see whether patient survival differed with respect to the organ affected by cancer. In case when the ANOVA will establish that patients with certain cancers have a greater mean survival time than patients with other cancers, you will estimate the differences.

Comparison of Cancer Survival

In Canada cancer is the 2nd top cause of death, second only to heart disease. One-half of cancer deaths in Canada in 2004 were due to four kinds of malignancies: lung, colorectal, female breast and male prostate. It is widely believed that the average survival times of cancer patients might depend on the type and organ of origin of the cancer. In order to verify this hypothesis, some researchers conducted a study to compare the survival times of patients who had various types of terminal cancers. In the study, 100 terminal cancer patients were obtained by random selection from the alphabetical index of cancer patients in a hospital. Survival times were measured from the date the cancer was established to be untreatable.

In this assignment, we will use a subset of the data for 64 patients, who had advanced cancers of the breast, bronchus, colon, ovary or stomach, to determine if patient survival differed with respect to the organ affected by the cancer. The data were originally obtained from an online Data and Story Library (DASL). We have modified the data for the special purpose of this assignment.

The data are available in the SPSS file *lab2.sav* located on the *STAT 252* Laboratories web site at http://www.stat.ualberta.ca/statslabs/index.htm (click *Stat 252* link). In order to download the data for the lab, click on the link *Data* for lab 2 and follow the instructions. The data are not to be printed in your submission.

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

Column	Variable Name	Description of Variable
1	SURVIVAL	survival time, in days (since the cancer was established to be untreatable),
2	ORGAN	organ affected by the cancer (breast, bronchus, colon, ovary or stomach).

Answer the following questions using the data:

1. First you will analyze the study design.

   (a) What is the purpose of the study? What kind of inferences can be made from the study? Can we conclude that any observed differences among survival times for the five cancer groups are due to the cancer type? Can the results of the study be extended to the populations of all patients with one of the five cancers discussed in the study in all hospitals?

   (b) Except for types of cancers, what are other possible variables that might have affected the survival times of the cancer patients? How can you control some of the variables? Provide brief explanations.

   (c) Notice that the survival time is measured since the cancer was established to be "untreatable". The cancer can be determined to be untreatable when a substantial part of the organ has been attacked by the malignancy, the cancer has begun to spread outside the organ, the level of aggressiveness of the cancer is high, etc. Explain how developing consistent criteria for untreatability across the five types of cancer is important for the outcome of the study.

2. Use the *Explore* procedure to obtain the descriptive statistics, the side-by-side boxplots, and the normality plots of survival time for the five different cancer groups. In particular:

   (a) Obtain and paste the descriptive statistics for each group into your report. Compare the means and standard deviations of the five distributions.

   (b) Obtain and paste the side-by-side boxplots into your report. Comment about the center, spread, and shape (symmetric, skewed) of each distribution. Do the plots indicate any differences in the centers and spreads among the five groups? Are there any outliers?

   (c) Obtain and paste the normality plots for each distribution into your report. Do any of the plots indicate any clear deviations from normality?

3. Use the *Compute* command in the *Transform* menu to form a new variable *LNSURV* defined as follows: *LNSURV = LN(SURVIVAL)*, that is, apply the natural logarithm transformation to the *SURIVIAL* variable. Then use the *Explore* procedure to carry out the analysis in Question 2 for the log-transformed data. In particular:

   (a) Obtain and paste the descriptive statistics for each group for the log-transformed data into your report. Compare the means and standard deviations of the five distributions.

   (b) Obtain and paste the side-by-side boxplots of the log-transformed data into your report. Comment about the shape (symmetric, skewed) of each distribution. Compare the spreads of the five distributions. Are there any outliers?

   (c) Obtain and paste the normality plots for each distribution into your report. Do any of the plots indicate any clear deviations from normality?

4. The ANOVA inferences are valid if the assumption of normality is not seriously violated for each of the groups. Moreover, the groups should come from populations with equal variances.

   (a) For log-transformed data, is there any evidence that the assumption of equal variance might be violated? Refer to the side-by-side boxplots obtained in part (b) of Question 3 to verify the assumption. Moreover, use the Levene's test to check the equal variance assumption. To run the Levene's test, you will need to recode the five categories of *ORGAN* (breast, bronchus, colon, ovary or stomach) into numerical values from 1 to 5 respectively. The recoding is also useful in Questions 5-8.

(b) For log-transformed data, is the normality assumption reasonable? Comment referring to the normality plots in part (c) of Question 3.

The statistical analyses in Questions 5-7 should be carried out on the log-scale if justified by the analysis in Question 4.

5. Is there any evidence that patients affected by certain cancers survived longer than patients affected by other cancers, on the average? Answer the question by running the one-way ANOVA test in SPSS.

   (a) Define the null the alternative hypotheses in terms of the population parameters of interest that correspond to the question asked.

   (b) Paste the ANOVA output into your report. What are the sums of squared residuals from fitting the full and reduced model? What is the pooled estimate of the variance? What are the value of the F statistic, the distribution of the test statistic under the null hypothesis, and the p-value of the test? State your conclusions.

   (c) By hand, demonstrate how to obtain the value of the F-statistic given the sums of squared residuals from fitting the full and reduced model. Moreover, use the summaries in part (a) of Question 3 to demonstrate (for the cases 1 and 12 in the data file) how the residuals for the full and reduced model can be obtained.

6. Which cancer patients differ in their mean survival times from others? Answer the question by carrying out the Scheffe's multiple-comparison procedure at the level of significance 0.05. Copy and paste the outputs to your report and state your conclusions.

7. Suppose that before conducting the experiment, the researchers decided to compare the mean survival time of patients with breast, colon or ovary cancers to that of patients with bronchus or stomach cancers. Do patients with advanced cancer of breast, colon or ovary tend to live longer than patients with bronchus or stomach cancers? Answer the question by setting up an appropriate contrast in SPSS, pasting the output into your report and interpreting the result.

8. The Kruskal-Wallis test (*Analyze, Non-parametric Tests, Independent Samples…*) is the nonparametric equivalent of the analysis of variance F-test. Use the Kruskal-Wallis test to test the null hypothesis that all cancer populations possess the same survival time distribution against the alternative hypothesis that the distributions differ in location. Report the p-value of the test. What do you conclude? Are the conclusions consistent with the outcome of the F-test in part (b) of Question 5? Explain briefly.

LAB ASSIGNMENT 2 MARKING SCHEMA

Proper Header: 10 points

Question 1

(a) Purpose of the study: 2 points
Causal inferences: 2 points
Inferences to the populations: 2 points
(b) Confounding variables: 2 points
How to control the confounding variables: 2 points
(c) Developing consistent criteria for untreatability: 2 point

Question 2

(a) Descriptive Statistics for each group: 5 points
Comparison of means, and standard deviations: 2 points
(b) Side-by-side boxplots: 4 points
Center, spread, and shape: 3 points
Outliers: 1 point
(c) Normality plots: 5 points
Comments about normality: 2 points

Question 3

(a) Descriptive Statistics for each group: 5 points
Comparison of means, and standard deviations: 2 points
(b) Side-by-side boxplots: 4 points
Center, spread, and shape: 3 points
Outliers: 1 point
(c) Normality plots: 5 points
Comments about normality: 2 points

Question 4

(a) Comment on side-by-side boxplots: 2 point
Levene't test: 2 point
(b) Normality assumption: 2 points

Question 5

(a) Null and alternative hypotheses: 3 points
(b) ANOVA output: 3 points
Sums of squares of residuals from fitting the full and reduced model: 2 points
Pooled estimate of the variance: 2 points
Value of the F-statistic and p-value: 2 points
Null distribution: 2 points
Conclusion in plain language: 2 points
(c) Calculating the value of F by hand: 4 points
Calculating the residuals for full and reduced model: 4 points

Question 6

Scheffe's procedure output: 3 points
Comments: 2 points

Question 7

Defining appropriate contrast: 3 points
Contrast output: 2 points
Conclusion: 2 points

Question 8

P-value: 1 point
Conclusions: 2 points
Consistent: 1 point

TOTAL= 112

Lab developed by Yuan Yuan Liang
Grant: G227150109 – Acc1Time Science
Revised by Henryk Kolacz