STAT 337 LAB EXAM

Instructions

1. This exam consists of two problems. For each problem, carry out the appropriate analysis using SPSS and give the answer in the space provided. When a text box is provided below the question, give brief, concise answers in the format provided by the box. All numerical values should be rounded to four digits. The exam is out of 128. The exam consists of 7 pages.

2. For each of the two problems you will have to download the appropriate data from STAT 337 Labs web site (*Lab Exam Data* in *Exams and Tests* panel). Once you have saved the exam data files to your desktop, close all programs including your authentication. At this point you are only allowed to use SPSS.

3. You are allowed to use the *Statistical Sleuth* text in the exam. You are not to communicate with any other individual, in any manner, with the exception of the proctor.

4. Complete the following (please print):

Lab Section Number: ___________ Name ___________________________________

Problems

1. In order to compare the strength qualities of 5 new alloys at extremely high temperatures, random samples of specimens from each alloy were obtained and their tensile strengths were measured. The related data are saved in the file *alloys.sav* available on STAT 337 Labs web site (*Exam Data* link). The following is a description of the variables contained in the data file:

Variable Name	**Description of Variable**
Strength	Tensile strength (pounds per square inch, often abbreviated to psi),
Type	Alloy type (an integer from 1 to 5).

The five different types of alloys considered in the experiment are:

Type 1: Nickel-based alloy with aluminum added (NA),
Type 2: Nickel-based alloy with chromium added (NC),
Type 3: Nickel-based alloy with titanium added (NT),
Type 4: Iron-based alloy with aluminum added (IA),
Type 5: Iron-based alloy with chromium added (IC).

Is there any evidence that some alloys are stronger than others? Answer the question and other related questions by running the one-way ANOVA test in SPSS.

(a) Define the null and alternative hypotheses of the ANOVA model in terms of the group means $\mu_{NA}, \mu_{NC}, \mu_{NT}, \mu_{IA}, \mu_{IC}$.

(2) Null hypothesis: $H_0 : \mu_{NA} = \mu_{NC} = \mu_{NT} = \mu_{IA} = \mu_{IC}$

(2) Alternative hypothesis: At least one mean is different from others.

(b) What are the sums of squared residuals (SSR) from fitting the full (five-mean) and reduced (one-mean) model? What is the pooled estimate of the variance?

(2) SSR(full model): 5423.017

(2) SSR(reduced model): 15077.943

(2) Estimate of variance: 83.431

(c) What is the value of the F-statistic, the distribution of the F statistic under the null hypothesis, and the p-value of the test? Express in plain language what the output says about the differences in the tensile strength of the six alloys.

(2) F-statistic value: 28.931

(2) Distribution: F distribution with 4 DF (numerator) and 65 DF (denominator)

(1) P-value: 0

(1) Conclusion: At least one pair of means is different.

(d) Consider the following two-mean model: the first three nickel-based groups have the same mean, possibly different from the mean of the two iron-based groups. Does the five-mean model discussed in parts (a)-(c) provide a significantly better fit than the two-mean model? Calculate the value of the appropriate test statistic to answer the question (show your calculations). Then specify the distribution of the test statistic and estimate the p-value of the test with the attached table. What is your conclusion?

(2) SSR(two-mean model): 5983.829

(4) Value of the test statistic (show the calculations):

$$F = \frac{(5983.829 - 5423.017)/(68-65)}{83.431} = 2.2406.$$

(2) Distribution: F distribution with 3 DF (numerator) and 65 DF (denominator)

(2) P-value: between 1-0.95 and 1-0.9

(1) Conclusion: The five-mean model does not provide a significantly better fit.

(e) Which alloys do not differ in their tensile strength from the others? Answer the question by carrying out the Tukey's (HSD) range tests at the level of significance 0.05. Use the abbreviations NA, NC, NT, IA, and IC in your answer.

(3) Groups of alloys, which are not different:

1. NA, NC, NT
2. IA, IC

(f) How strong is the effect of the components added (aluminum, titanium, and chromium) on the tensile strength of the five alloys? Refer to part (e) to answer the question (3)

The components added do not change significantly the tensile strength. The main component (nickel or iron) makes the difference.

(g) Do alloys with aluminum added tend to be stronger in their tensile strength than alloys with chromium added? Answer the following questions by setting up an appropriate contrast in SPSS, and interpreting the result.

(3) Contrast: $\chi = \frac{\mu_1 + \mu_4}{2} - \frac{\mu_2 + \mu_5}{2}$

(3) Hypotheses: $H_0 : \chi = 0$ vs. $H_A : \chi > 0.$

(2) Estimate g of the contrast: g=5.5125

(2) p-value of the test: 0.025/2 or 0.028/2

(1) Conclusion: aluminum alloys tend to be stronger than chromium alloys.

(h) Obtain a side-by-side boxplots and Q-Q plots of tensile strength for the five alloys. Which ANOVA assumptions may be violated? Specify the alloy(s) the assumptions may be violated. Which of the assumptions is crucial? How can the problem be corrected? (3)

The assumption of equal variances may be violated. The first two alloys exhibit clearly larger variation than the three other. The assumption of normality may be violated for type-3 alloy (systematic departure from a straight-line pattern). The equal variance assumption is crucial. The log-transformation may be used to correct the problem.

(i) How would the value of the F statistic and the corresponding p-value be affected if it turned out that the measurements for the iron-based alloys (type 4 and 5 alloys) were seriously deflated due to measurement errors?

(2) Effect on F value: stays the same, increase, **decrease**

(1) Effect on the p-value: stays the same, **increase,** decrease

2. Some red spruce forests in the Appalachian Mountains show signs of decline, with many dead or dying trees. Environmental stress may contribute to this decline; there is evidence of heavy deposition of airborne pollutants such as metals or acids in the area. The related data from 61 Appalachian sites is saved in the file *spruce.sav* available on STAT 337 Labs web site (*Exam Data* link). The following is a description of the variables contained in the data file:

Variable Name	**Description of Variable**
LOC	1 if North, 0 if South,
ELEV	Elevation in meters,
DEAD	Percentage of damaged or dead trees.

You will compare the mean percentage of dead or damaged trees in the two locations (North, South) first with the t-tools, a then use linear regression compare the percentage of damaged or dead trees in the two locations.

(a) Is there any difference between the mean percentage of damaged or dead trees in the two locations (North, South)? Use the appropriate t-tools on the **natural logarithm** scale to make the comparison.

(3) Null and alternative hypotheses:

$H_0 : \mu_{NORTH} = \mu_{SOUTH}$ vs. $H_A : \mu_{NORTH} \neq \mu_{SOUTH}$,

where μ_{NORTH}, μ_{SOUTH} are the mean percentage of damaged or dead trees on the natural logarithm scale (or equivalently in terms of the medians).

(2) Name of the t-test in SPSS: Independent Samples T test

(2) t-statistic value: t=4.515 (DF=62).

(1) p-value of the test: reported as 0.

(1) Conclusion: There are differences between the percentage of damaged or dead trees in North and South.

(b) Use the output in part (a) to estimate the ratio of the median percentage of damaged or dead trees in the North to the median percentage of damaged or dead trees in the South? What is a 95% confidence interval for the ratio?

(2) Estimate: exp(1.09397)=2.986105.

(2) 95% confidence interval: [exp(0.60964), exp(1.57830)]=[1.839769, 4.846709].

(c) Would the test in part (a) be valid on the original scale? Explain briefly referring to the appropriate plots.(3)

The spread in the two distributions (North, South) is very different. The test would not be valid.

(d) Apply the appropriate non-parametric test to answer the question in part (a).

(2) Name of the test: Mann-Whitney U Test (to compare the distributions of the percentage of damaged or dead trees for the two locations) or the test to compare the medians for the two distributions.

(2) Value of the test statistic: Not reported by SPSS 19

(1) P-value: 0.001 for Mann-Whitney or 0.002 for the medians test.

(1) Conclusion: Strong evidence of differences in the distributions of percentage of damaged or dead trees for the two locations (strong evidence that the medians of percentage of damaged or dead trees are different for the two locations).

Now you will use linear regression to compare the percentage of damaged or dead trees in the two locations. You will take into account another variable, elevation in the comparison.

(e) Now apply the natural logarithm transformation to the variable ELEV. Obtain a scatterplot of percentage of trees damaged or dead vs. log-elevation with different marking symbol for each location (north or south). Describe the relationship between percentage and elevation for each location (linear? positive or negative? weak, moderate, or strong?)

(2) Relationship for North: moderate in strength, positive linear relationship

(2) Relationship for South: moderate in strength, negative linear relationship. Separate scatterplot for the location should be obtained to evaluate properly the relationship.

(f) Calculate the correlation coefficient between percentage damaged or dead trees and log-elevation for each location (North, South).

(2) Correlation for North: 0.6742

(2) Correlation for South: -0.5377

(g) Consider the following regression model with percentage of damaged or dead trees as the response variable:

$$DEAD = \beta_0 + \beta_1 \cdot LN(ELEV) + \beta_2 \cdot LOC + ERROR,$$

where ERROR follows a normal distribution with mean 0 and standard deviation σ. Write the estimated regression equation for the model. (4)

$$\mu\{DEAD \mid ELEV, LOC\} = -438.245 + 63.228 \cdot LN(ELEV) + 49.762 \cdot LOC.$$

(h) Is the regression model in part (g) suitable given the scatterplot in part (e)? Explain.(3)

The regression equation in (g) defines two parallel lines. However, the scatterplot in (e) shows two intersecting bands. Thus the model with an interaction would be more appropriate.

(i) What is an estimate of the standard deviation σ? (2)

$\hat{\sigma} = 16.0336$

(j) What percent of the variation in percentage damaged is explained by log-elevation and location? (2)

44.1% or adjusted 42.3%

(k) Is the regression model useful, i.e. at least one explanatory variable is an useful predictor? Report the value of the appropriate test statistic and the p-value of the test.

(2) Test statistic value: F=24.048, F distribution (2, 61)
(1) p-value reported as 0,
(1) Conclusion: model is useful.

(l) Use the estimated regression equation in part (g) to estimate the difference in mean percentage of dead or damaged trees between the northern and southern locations at any elevation.(2)

49.762

(m) What is a 95% confidence interval for the difference in part (l)? (3)

[34.849, 64.675]

(n) Use the estimated regression equation in part (g) to estimate the change in percentage of damaged or dead trees as elevation increases by 10%? Show your calcualtions.(3)

Additive change of $\beta_1 \cdot LN(1.1) = 63.228 \cdot 0.09531 = 6.026272.$

(o) Is there any evidence that the percentage of dead trees increases with elevation regardless of the location? State this question as null and alternative hypotheses about a regression coefficient in the above model, obtain the test statistic and its p-value from the output, and give your conclusion.

(2) Null hypothesis: $\beta_1 = 0$

(2) Alternative hypothesis: $\beta_1 > 0.$

(2) Test statistic value: 5.425

(2) Distribution of the test statistic: t distribution with 60 DF

(1) P-value: reported as 0

(1) Conclusion: Strong evidence that percentage of damaged or dead trees increases with log-elevation (so also with elevation).

(p) What is the predicted percentage of damaged or dead trees in North at the elevation of 1,000 meters? What is the value of the residual for this case?

(2) Predicted percentage: 48.28048
(1) Residual: -6.28048

(q) Obtain a 95% confidence interval for the mean percentage of dead or damaged trees in North at elevation of 1,000. Obtain also a 95% prediction interval for the percentage of dead or damaged trees in a northern location, at elevation of 1,000. Use the theory to calculate the elevation at which the two intervals are narrowest?

(2) 95% confidence interval: [43.18639, 53.37456]

(2) 95% prediction interval: [15.81715, 80.74380]

(2) The intervals are narrowest at the average elevation for North of 6.7895 (log-scale) or 888.4692 on the original scale.

(r) Obtain the normal probability plot of standardized residuals for the regression model in part (g). Is there any evidence that the assumption of normality may be violated? (2)

No evidence that the assumption of normality may be violated.

(s) Obtain the plot of standardized residuals versus standardized predicted values for the regression model in part (g). What assumptions may be examined using the plot? Is there any evidence that any of the assumptions may be violated?

(2) Assumptions tested: linearity, equal variance.

(2) Assumptions: The assumption of equal variances may be violated.