22S:30/105, Statistical Methods and Computing
Spring 2013, Instructor: Cowles
Final Exam

Name: Solutions _____ Course no. (30 or 105) _____

1. Research by Singh et al. (1999) as reported in the journal *Clinical Immunology and Immunopathology* is concerned with immune abnormalities in autistic children. As part of their research, they took measurements on the serum concentration of an antigen in three samples of children, autistic children, normal children, and mentally-handicapped children (non-Down's-syndrome). All children were 10 years old or younger.

   This problem uses data from only two of the samples, autistic children and mentally-handicapped children. This dataset contains two variables:

   ```
   concentration of the antigen (in units per milliliter of serum)
   group,     coded A for autistic
                    M for mentally handicapped
   ```

   I wish to use these data to infer whether the center of the population distribution of antigen concentration is the same in the population of all autistic children and the population of all mentally handicapped children.

   (a) I will use a two-independent-sample method. Why is that the right thing to do for this example? (Circle all that apply.)

   i. There are outliers in the boxplots.
   ii. The standard deviations are unequal in the two samples.
   iii. The population is less than 10 times as large as the sample.
   iv. There are fewer than 10 successes and 10 failures.
   v. The description gives no evidence that the samples were paired in any way.
   vi. There are at least 15 observations in each sample.
   vii. The sample sizes in the two samples are unequal.

   (b) I have provided SAS results from a two-independent-sample t test and from the Wilcoxon rank sum test. Which test would be the better one to use for these data? Briefly justify your answer, referring to various SAS output.

We will not have a question about Wilcoxon on the 2014 test. It would be the better choice here because of the outliers in the autistic group.

   (c) For the test that you chose in the preceding question, write the null and alternative hypotheses being tested. Define any statistical symbols that you use.

$H_0: M_a = M_m$
$H_0: M_a \neq M_m$

where M_a is median of antigen concentration in autistic population
M_m is median in mentally handicapped population

(d) At the 0.05 significance level, should I reject the null hypothesis? Justify your answer, referring to specific SAS output.

No, $p = 0.0730$ for two-sided Wilcoxon test

$0.0730 > 0.05$

(e) Explain what your answer to the previous question means for antigen concentrations in the two populations of children.

There is weak evidence of a difference between the median antigen concentration in the two populations, but the difference is not statistically significant.

(f) Circle all of the statements below that are true.

F

F

F

i. The Wilcoxon rank sum test is a parametric test.

ii. The two-independent-sample t-test requires the standard deviations in the two populations to be equal.

iii. An assumption of the Wilcoxon rank sum test is that the medians in the two populations are equal.

2. The dataset for this problem is described as follows (in the textbook *Applied Regression* by Fox):

```
Data on Vocabulary and Education from the 1989 General Social Survey

[1] Observation Index

[2] Education, in years

[3] Vocabulary Test Score, 10-Item Test with possible scores 0 - 10.

Source: 1989 General Social Survey, National Opinion Research Center.
Distributed by the Inter-University Consortium for Political and Social
Research.
```

Here are the first 10 rows of the dataset:

Obs	obsno	yrsed	score
1	1	0	5
2	2	1	1
3	3	3	1
4	4	3	3
5	5	4	1
6	6	4	0
7	7	4	1
8	8	4	2
9	9	4	4
10	10	4	5

I wish to use these data to infer about the relationship between education and vocabulary.

(a) The scatterplot possibly indicates a weak linear relationship, but it looks wierd. Why are all the data points laid out in rows and columns?

Because all the values of both the response variable and the predictor are integers.

(b) From the regression output, give the point estimate and 95% confidence interval for the slope (numeric answer):

$b = 0.374$ 95% c.i. for β is $(0.333, 0.415)$

(c) In one sentence, interpret the point estimate of the slope in terms of the years of education and vocabulary.

For each 1 year increase in yrs of education we expect on average a 0.374 point increase in vocabulary score.

(d) What quantity are you 95% confident lies in the confidence interval that you cited?

population slope β

(e) At the 0.05 significance level, is the slope different from 0? Cite two different parts of the regression output to support your answer.

Yes. p-value for $H_0: \beta = 0$ is < 0.0001 and the 95% c.i. does not contain 0.

(f) Based on this regression model, what vocabulary score would you predict for a person with 8 years of education? (Numeric answer; show your work.)

$\widehat{score}_i = 1.1354 + 0.374(8)$

$=$

3. Researchers wished to compare the effectiveness of four different treatment regimens for chronic renal failure in dogs. Twenty-eight dogs in stage 3 renal failure were randomly assigned to four different treatment groups (7 dogs in each group). The outcome variable was change in serum creatinine from pretreatment to 8 weeks on treatment. A negative change represented a decrease in serum creatinine, which is the hoped-for outcome.

Refer to the SAS output below in answering the following questions.

(a) ANOVA was used for the analysis. Explain briefly why ANOVA was the correct method instead of the Chi square test.

ANOVA is used to compare population means of a quantitative variable such as serum creatinine. The Chi square test is for comparing proportions.

(b) Do the data appear to satisfy the assumptions of ANOVA? Explain briefly, referring to specific SAS output.

Yes. There are no outliers or extreme skewness in any of the boxplots, so the normality assumption is ok. The smallest sample standard deviation (2.319) is more than half as big as the largest, so equality of population standard deviations probably is OK. No way to tell about SRS.

(c) Write the null and alternative hypotheses being tested by the global F test. Use standard statistical symbols. Briefly define the symbols that you use.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Any one or more pairs of population means are not equal.

μ_1 = mean change in serum creatinine in population of all dogs treated with treatment 1

(d) At the 0.05 significance level, is the null hypothesis rejected? Justify your answer with specific SAS output.

Yes. p-value = 0.0296 < 0.05.

(e) Which population means, if any, are significantly different from each other? Justify your answer with specific SAS output.

$\mu_1 \neq \mu_4$ These are the only two that do not share a letter in the "Box grouping" section

(f) Does this study indicate that any one of the 4 treatments is clearly the best? Explain briefly.

No. Although treatment 1 has the best sample mean, the t-tests show that the differences between treatment 1 and treatments 2 and 3 are not statistically significant.

4. Biologists are planning to study body temperature in a particular species of finches (a kind of bird). They want to construct a confidence interval for the mean body temperature in adults of this species.

(a) The biologists are convinced that body temperatures in this species follow a normal distribution with standard deviation 1.5 degrees Fahrenheit. How many finches will they need to capture and measure in order to construct a 90% confidence interval width no greather than 2 degrees? (Numeric answer; show your work.)

if width is 2, margin of error is 1

$$n = \left(\frac{z^* \sigma}{m}\right)^2 = \left(\frac{1.645(1.5)}{1}\right)^2 = 6.08$$

Round up to 7.

(b) Without doing any calculations, state whether the biologists would need a larger sample or a smaller sample if they wanted a 95% c.i. instead of a 90% c.i.

Larger sample would be needed for an interval with higher confidence level.

5. Write the data type for each of the following variables. Choose from binary, nominal, ordinal, quantitative discreet, or quantitative continuous.

(a) winning scores in college football games quantitative discrete

(b) whether or not the home team wins each college football game binary

(c) patients' conditions as described by the American Hospital Association: good, fair, serious, critical ordinal

SAS for problem 1

The UNIVARIATE Procedure
Variable: conc

Schematic Plots