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

`Name: ________________________________________ 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.

   (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.

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

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

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

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?

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

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

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

(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.

(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.)

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.

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

   (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.

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

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

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

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.)

   (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.

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

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

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

SAS for problem 1

```
        The UNIVARIATE Procedure
            Variable:  conc

            Schematic Plots

      |
 1000 +
      |
      |            *
      |
  800 +            *
      |            *
      |                        0
      |
  600 +
      |                        |
      |                        |
      |        +-----+         |
  400 +        *--+--*     +-----+
      |        +-----+     |  +  |
      |           |        *-----*
      |           |        |     |
  200 +           |        +-----+
      |           0            |
      |                        |
      |
    0 +
       ------------+-----------+-----------
group                  A           M
```

The TTEST Procedure

Variable: conc

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	23	419.9	178.0	37.1210	170.0	900.0
M	15	329.3	171.0	44.1396	105.0	715.0
Diff (1-2)		90.5797	175.3	58.1817		

group	Method	Mean	95% CL Mean		Std Dev
A		419.9	342.9	496.9	178.0
M		329.3	234.7	424.0	171.0
Diff (1-2)	Pooled	90.5797	-27.4183	208.6	175.3
Diff (1-2)	Satterthwaite	90.5797	-27.0541	208.2	

group	Method	95% CL Std Dev	
A		137.7	252.0
M		125.2	269.6
Diff (1-2)	Pooled	142.6	227.7
Diff (1-2)	Satterthwaite		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	36	1.56	0.1283
Satterthwaite	Unequal	30.953	1.57	0.1265

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	14	1.08	0.8975

NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable conc
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	23	509.0	448.50	33.470412	22.130435
M	15	232.0	292.50	33.470412	15.466667

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	232.0000
Normal Approximation	
Z	-1.7926
One-Sided Pr < Z	0.0365
Two-Sided Pr > |Z|	0.0730

Kruskal-Wallis Test

Chi-Square	3.2673
DF	1
Pr > Chi-Square	0.0707

SAS for problem 2

```
    Plot of score*yrsed.  Legend: A = 1 obs, B = 2 obs, etc.

score |
      |
   10 +               A       D B C BM G G F F
      |
    9 +                   A   S E E CQ E J B D
      |
    8 +             B B   A C Z I W LS C E B B
      |
    7 +             A E E E H Z T V IM C E A A
      |
    6 +               E D I M Z Q Z MU G C B
      |
    5 +A       A  B B K H D G Z Q Q GI D B
      |
    4 +        A AB B D D H L Z E M AB B B   C
      |
    3 +      A    A A G F D D N B G  C   B
      |
    2 +        A C  A B D D H P A   B    A
      |
    1 +  A   A B  B   C C D B B
      |
    0 +        A    B D   A A C   B
      |
      -+------------------+------------------+-
       0                 10                 20
                        yrsed
```

The REG Procedure
Model: MODEL1
Dependent Variable: score

Number of Observations Read	968
Number of Observations Used	968

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1175.11129	1175.11129	318.92	<.0001
Error	966	3559.41351	3.68469		
Corrected Total	967	4734.52479			

Root MSE	1.91956	R-Square	0.2482
Dependent Mean	5.94008	Adj R-Sq	0.2474
Coeff Var	32.31530		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.13481	0.27606	4.11	<.0001
yrsed	1	0.37413	0.02095	17.86	<.0001

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	0.59306	1.67656
yrsed	1	0.33301	0.41524

SAS for problem 3

```
                        The UNIVARIATE Procedure
                              Variable:  y

                            Schematic Plots

      |
   10 +
      |
      |                                            |
      |                                            |
    8 +                                            |
      |           |                                |
      |           |          |                  +-----+
      |           |       +-----+               |     |
    6 +           |       |     |       |       |     |
      |           |       |     |       |       |     |
      |           |       |     |    +-----+    *-----*
      |           |       |     |    |     |    |  +  |
    4 +           |       |  +  |    |     |    |     |
      |           |       |     |    |     |    |     |
      |           |       *-----*    |     |    |     |
      |           |       |     |    |     |    |     |
    2 +           |       +-----+    *--+--*    |     |
      |           |          |       |     |    +-----+
      |           |          |       |     |       |
      |           |                  +-----+
    0 +        +-----+                  |
      |        *--+--*                  |
      |        |     |                  |
      |        |     |                  |
   -2 +        |     |                  |
      |        |     |                  |
      |        |     |                  |
      |        |     |                  |
   -4 +        |     |                  |
      |        +-----+                  |
      |           |                     |
      |
   -6 +
       -----------+----------+----------+----------+----------
group                1          2          3          4
```

The MEANS Procedure

Analysis Variable : y

group	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	7	7	-0.6900000	4.1492931	-5.1900000	7.2900000
2	7	7	3.9514286	2.3187743	1.1300000	6.8100000
3	7	7	1.9714286	3.6674125	-4.9700000	6.0700000
4	7	7	4.5342857	2.8108472	1.0400000	8.8500000

The ANOVA Procedure

Class Level Information

Class	Levels	Values
group	4	1 2 3 4

Number of Observations Read	28
Number of Observations Used	28

The ANOVA Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	116.8082679	38.9360893	3.54	0.0296
Error	24	263.6647429	10.9860310		
Corrected Total	27	380.4730107			

R-Square	Coeff Var	Root MSE	y Mean
0.307008	135.7416	3.314518	2.441786

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	3	116.8082679	38.9360893	3.54	0.0296

The ANOVA Procedure

Bonferroni (Dunn) t Tests for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	24
Error Mean Square	10.98603
Critical Value of t	2.87509
Minimum Significant Difference	5.0938

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	group
	A	4.534	7	4
	A			
B	A	3.951	7	2
B	A			
B	A	1.971	7	3
B				
B		-0.690	7	1