22S:30/105, Statistical Methods and Computing
Spring 2014, Instructor: Cowles
Final Exam

`Name: _________________________________________ Course no. (30 or 105) _____`

1. Briefly explain the statistical error in each of the following statements.
   (a) The researchers found a strong positive correlation between employment status (working, unemployed, retired) and happiness.
   (b) The researchers found a strong positive correlation between annual income and score on a happiness test. This proves that money makes people happy.
   (c) The researchers reported a p-value of 0.032 for their hypothesis test. This means that there is 0.032 probability that the null hypothesis is true.

2. This problem uses data from collected in West Africa for the years 1971 - 1999. In each year two variables were recorded:
   - `fish` – the fish supply in kilograms per person
   - `biomchg` – the change in total weight of animals in nature reserves

   The expectation was that, in years in which the fish supply was smaller, declines in animals would be greater because the local populace would turn to such animal meat when other sources of protein were reduced.

   We will use linear regression to see whether the data confirm this expectation.

   The attached SAS output is needed to answer some of the following questions.

   (a) In the scatterplot, which variable – `fish` or `biomchg` – is treated as the response variable?
   (b) Does the scatterplot indicate a somewhat linear relationship? (yes no) If so, is the relationship positive or negative?

(c) SAS `proc corr` gives the sample correlation coefficient between `fish` and `biomchg` as 0.67. Would this correlation get larger or smaller if the outlier were removed from the dataset. Explain why in one or two sentences.

(d) Does the residual plot give evidence of any violations of the assumptions of linear regression? Explain briefly.

(e) I claim that the outlier will not be very influential on the regression analysis. Why is that the case?

(f) Give the least squares regression line for the regression of `biomchg` on `fish`.

(g) Write the null hypothesis that means that there is no linear relationship between `biomchg` and `fish`. Use conventional statistical symbols.

(h) Write the alternative hypothesis that expresses the expectation stated in the problem description.

(i) Write the test statistic value and p-value (numbers from SAS output) for carrying out the test of the hypotheses that you stated.

(j) What is your conclusion? Does the data contain evidence in favor of the expectation stated? Explain briefly.

3. Investigators are interested in whether men or women have better perception of color. They design a test of color perception, in which higher numeric scores are better. They know that age affects vision in general and may well affect color perception. To balance out the effects of age, they recruit the participants in their study in the following way. They first recruit 10 men into their study. Then for each male subject, they recruit one female subject of the same age. Thus, they end up with a sample of 10 men and a sample of 10 women. Which statistical procedure is most likely to be best for their study? (Circle one.)

   (a) ANOVA
   (b) chi square test
   (c) paired t-test
   (d) two-independent-sample t-test

4. It is expected that rainfall in California will increase over the coming decades, but it is not known whether the increase will occur only during the winter wet season or whether there also will be more rain in the spring and summer, which historically have been dry.

   Researcher Kenwyn Suttle of the University of California at Berkeley did an agricultural study to investigate the effect of three possible conditions on the plant biomass in open grassland:

   - Control (no added water)
   - Winter (add water equal to 20% of annual rainfall during January to March)
   - Spring (add water equal to 20% of annual rainfall during April to June)

   They randomly assigned 6 plots of open grassland to each of the three conditions. This problem concerns their results obtained during the 2001 growing season. SAS output is attached.

   (a) The researchers wish to use ANOVA to determine whether the mean biomass produced per acre is different under the three conditions. Write the null and alternative hypotheses that they will test. Use standard statistical symbols, and briefly define the symbols that you use.

   (b) What assumptions must be met in order for the results of ANOVA to be trustworthy? State each one, and then comment whether the assumption appears to met with these data. Refer to specific SAS output when appropriate.

(c) Suppose that the researchers proceeded to do an ANOVA analysis and will do the overall hypothesis test at significance level alpha = 0.05.

   i. Should they reject the null hypothesis that you stated above? Justify your answer referring to specific SAS output.

   ii. What does the result you gave in the previous question mean regarding biomass and extra water in different seasons?

(d) Which means (if any) are significantly different from each other? Refer to specific SAS output.

(e) Are you convinced by these results? Why or why not?

5. Sociologists want to determine the average number of hours per day that teenagers spend using social media such as Facebook. Suppose that they believe that the number of hours that individual teenagers spend daily on social media follows a normal distribution with known standard deviation 1.5 hours.

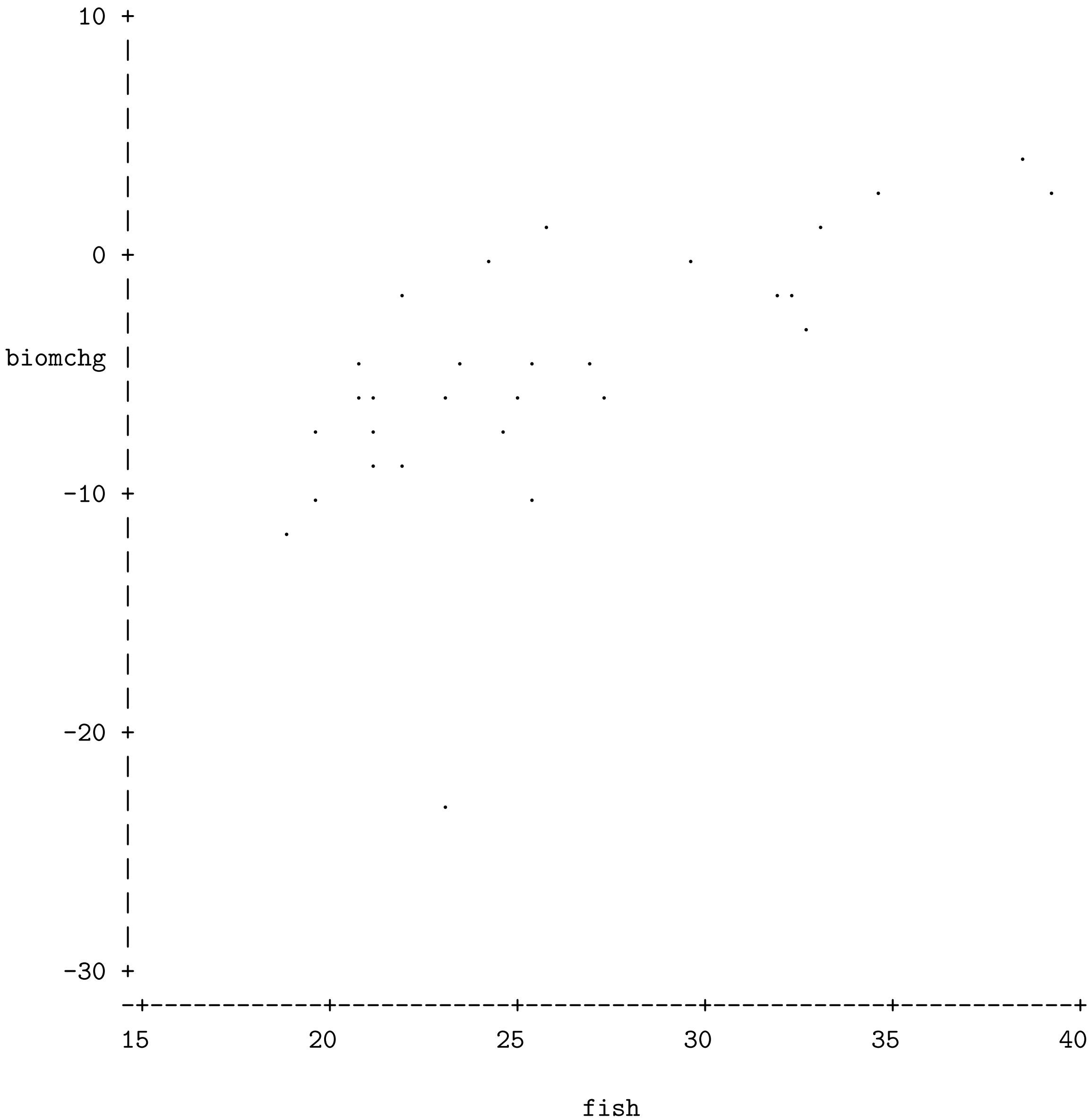
   (a) How many teenagers will they need in their sample in order to obtain a 90% confidence interval of width no greater than 1? (Numeric answer. Show your work.)

   (b) What quantity will the sociologists be 90% confident lies in their interval? (Circle one)

      i. μ
      ii. p
      iii. $\hat{p}$
      iv. s
      v. σ
      vi. $\bar{x}$

SAS output for problem 2

Plot of biomchg*fish. Symbol used is '.'.



The CORR Procedure

2 Variables: biomchg fish

Simple Statistics

Variable	N	Mean	Std Dev	Sum
biomchg	29	-4.60690	5.40271	-133.60000
fish	29	25.97931	5.72522	753.40000

Simple Statistics

Variable	Minimum	Maximum
biomchg	-22.90000	3.70000
fish	18.90000	39.30000

Pearson Correlation Coefficients, N = 29
Prob > |r| under H0: Rho=0

	biomchg	fish
biomchg	1.00000	0.67242 <.0001
fish	0.67242 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: biomchg

Number of Observations Read	29
Number of Observations Used	29

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	369.54264	369.54264	22.28	<.0001
Error	27	447.75598	16.58355		
Corrected Total	28	817.29862			

Root MSE	4.07229	R-Square	0.4522
Dependent Mean	-4.60690	Adj R-Sq	0.4319
Coeff Var	-88.39554		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-21.09189	3.57311	-5.90	<.0001
fish	1	0.63454	0.13442	4.72	<.0001

```
                             The REG Procedure
                               Model: MODEL1
                       Dependent Variable: biomchg

           -------+------+------+------+------+------+------+------+-------
  RESIDUAL |                                                              |
        10 +                                                              +
           |                                                              |
           |                                                              |
           |                                                              |
         5 +                .    .   .                                    +
           |             .                                                |
           |             ..                  .       .  .                 |
           |           .       .   .                                      |
R        0 +              .   .        .          .             .         +
e          |           .  . .     ..    .          .              .       |
s          |         .                              .                     |
i          |                                                              |
d       -5 +                                                              +
u          |                       .                                      |
a          |                                                              |
l          |                                                              |
       -10 +                                                              +
           |                                                              |
           |                                                              |
           |                                                              |
       -15 +                                                              +
           |                  .                                           |
           |                                                              |
           |                                                              |
       -20 +                                                              +
           -------+------+------+------+------+------+------+------+-------
                 -10    -8     -6     -4     -2      0      2      4

                         Predicted Value of biomchg     PRED
```

```
            The UNIVARIATE Procedure
               Variable:  biomchg

Stem Leaf                     #             Boxplot
   0 12334                    5                |
  -0 444321110                9             +-----+
  -0 998777666655            12             *--+--*
  -1 11                       2                |
  -1
  -2 3                        1                0
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

```
            The UNIVARIATE Procedure
                 Variable:  fish

Stem Leaf                     #             Boxplot
  38 43                       2                |
  36                                           |
  34 7                        1                |
  32 482                      3                |
  30 8                        1                |
  28 7                        1             +-----+
  26 14                       2             |  +  |
  24 350239                   6             *-----*
  22 0004                     4             |     |
  20 880138                   6             +-----+
  18 967                      3                |
     ----+----+----+----+
```

SAS output for problem 4

```
            The UNIVARIATE Procedure
                Variable:  x2001

                 Schematic Plots

    |
350 +
    |                        |
    |                        |
    |                        |
300 +                        |
    |                     +-----+
    |                     |     |
    |                     |  +  |
250 +                     |     |
    |                     *-----*
    |                        |
    |                        |
200 +                        |
    |
    |
    |
150 +                                    |
    |                                 +--+--+
    |                                 *-----*
    |            |
100 +         +-----+
    |         *--+--*
    |         +-----+
    |
 50 +
    |
    |            0
    |
  0 +
     ------------+-----------+-----------+-----------
trt           control      spring      winter

               The MEANS Procedure

           Analysis Variable : x2001
```

trt	N Obs	N	Mean	Std Dev	Minimum	Maximum
control	6	6	81.6696300	28.0673133	30.5886100	110.6306000
spring	6	6	257.6862667	49.5874144	197.5830000	338.1301000
winter	6	6	132.5802667	11.2219402	121.6323000	151.4154000

The ANOVA Procedure

Class Level Information

Class	Levels	Values
trt	3	control spring winter

Number of Observations Read	18
Number of Observations Used	18

The ANOVA Procedure

Dependent Variable: x2001

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	98450.5211	49225.2605	43.79	<.0001
Error	15	16863.0885	1124.2059		
Corrected Total	17	115313.6095			

R-Square	Coeff Var	Root MSE	x2001 Mean
0.853763	21.31380	33.52918	157.3121

Source	DF	Anova SS	Mean Square	F Value	Pr > F
trt	2	98450.52109	49225.26055	43.79	<.0001

The ANOVA Procedure

Bonferroni (Dunn) t Tests for x2001

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	15

Error Mean Square	1124.206
Critical Value of t	2.69374
Minimum Significant Difference	52.146

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	trt
A	257.69	6	spring
B	132.58	6	winter
B			
B	81.67	6	control