STAT:2010/4200, Statistical Methods and Computing
Spring 2016, Instructor: Cowles
Final Exam
May 12, 2016

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

1. Compute the 5-number summary of the following numbers. (Numeric answer.)

```
6 17 17 29 38 39 47 63  65 77 84
```

2. This problem is based on a dataset found at `http://www.stat.ufl.edu/~winner/datasets.html`. Here is a description of the data.

```
Dataset:  brainhead.dat

Source: R.J. Gladstone (1905). "A Study of the Relations of the Brain to
to the Size of the Head", Biometrika, Vol. 4, pp105-123

Description: Brain weight (grams) and head size (cubic cm) for 237
adults classified by gender and age group.

Variables/Columns
Gender   8   /* 1=Male, 2=Female  */
Age Range  16   /* 1=20-46, 2=46+  */
Head size (cm^3)  21-24
Brain weight (grams)  29-32
```

We can use the last two variables to assess the relationship between head volume in cubic centimeters and brain weight in grams.

(a) From the scatterplot, does the relationship between head volume and brain weight appear to be roughly linear? (yes/no)

(b) List the two characteristics that make a data point influential in linear regression.

(c) Do you see any influential points in this dataset? (yes/no)

(d) What proportion of the variability in brain weight is explained by head volume? (numeric answer taken from SAS output)

(e) Your answer to part (d) indicates that head volume is (circle the best answer)

i. useless as a predictor of brain weight

ii. a moderately good predictor of brain weight

iii. a nearly perfect predictor of brain weight

(f) The sample slope $b = 0.263$. Explain to someone who knows no statistics what this means in terms of the relationship between head volume and brain weight.

(g) Suppose you wished to test the alternative hypothesis that there is a positive linear relationship between head volume and brain weight, versus the null hypothesis that there is no linear relationship between these two variables. Write these hypotheses formally, using conventional statistical symbols.

(h) At the .05 significance level, what conclusion do you draw regarding the hypothesis test? Justify your answer using SAS output.

(i) In the fit plot for the linear regression, we can see dashed lines around the fitted regression line. These dashed lines represent (Circle the one best answer).

i. The endpoints of 95% confidence intervals for the means of brain weight at each of the values of head volume.

ii. The endpoints of 95% prediction intervals for individual new observations of brain weight.

iii. 95% confidence intervals for the intercept and slope parameters.

iv. None of the above.

3. The same dataset could be used to compare the population mean of head size in men and in women.

   Note that in the data, men are coded as 1 and women as 2. There are 134 men and 103 women.

   (a) Write the null and alternative hypotheses that would be appropriate if we expect to show that the population mean of head size is larger in men than in women. Use conventional statistical symbols.

   (b) I used a two-independent-sample t-test to analyze the data. Do you see any evidence in the SAS output provided (graphical or numeric) that the assumptions of the test you selected are not met? Briefly explain.

   (c) The following 95% confidence interval is given in the SAS output: (298.0, 457.9). What quantity are we 95% confident lies in that interval? (Use conventional statistical symbols.)

   (d) Based on the confidence interval, what conclusion do you draw regarding the hypothesis test? Explain briefly.

4. A study in Cameroon found that the wing length of males of a species of finches varies according to a normal distribution with mean 61.2 mm and standard deviation 1.8 mm.

   (a) What is the probability that the wing length of an individual male finch of this species would be greater than 63.5 mm? (Numeric answer; show your work.)

(b) Consider random samples of size 10 from the population of male finches of this species. What is the value such that only 2.5% of simple random samples of 10 male finches have a sample mean $\bar{x}$ larger than this value? (Numeric answer; show your work.)

5. In a recent SurveyUSA poll of 826 California voters likely to participate in the California Democratic primary, 471 stated that they would vote for Hilary Clinton.

   (a) Compute a 95% confidence interval for the proportion of all likely voters in the California Democratic primary who will vote for Clinton. (Numeric answer; show your work.)

   (b) Reports of polling results usually include a margin of error. What margin of error should be reported for this result? (Numeric answer.)

   (c) Would the margin of error have been larger or smaller if 400 likely voters had been surveyed instead of 826 (and if the sample proportion of Clinton supporters had been the same)?

6. Circle *all* of the following statements that are true.

   (a) The p-value is the probability that the null hypothesis is true.

   (b) In linear regression, a residual is the difference between an observed value of the response variable and the predicted value for the same observation.

   (c) Correlation measures the strength of the linear relationship between two quantitative variables measured on the same subjects.

   (d) The power of a statistical test is the probability of rejecting the null hypothesis when it is false.

   (e) All of the statistical methods that we studied this semester require that the data be a simple random sample from the population.

The REG Procedure
Model: MODEL1
Dependent Variable: brain

Number of Observations Read	237
Number of Observations Used	237

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2184982	2184982	416.53	<.0001
Error	235	1232728	5245.65113		
Corrected Total	236	3417710			

Root MSE	72.42687	R-Square	0.6393
Dependent Mean	1282.87342	Adj R-Sq	0.6378
Coeff Var	5.64568		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	325.57342	47.14085	6.91	<.0001
head	1	0.26343	0.01291	20.41	<.0001

The TTEST Procedure

Variable: head

gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	134	3798.3	327.8	28.3146	3095.0	4747.0
2	103	3420.3	295.0	29.0693	2720.0	4204.0
Diff (1-2)		378.0	314.0	41.1429		

gender	Method	Mean	95% CL Mean		Std Dev
1		3798.3	3742.3	3854.3	327.8
2		3420.3	3362.6	3477.9	295.0
Diff (1-2)	Pooled	378.0	296.9	459.0	314.0
Diff (1-2)	Satterthwaite	378.0	298.0	457.9	

gender	Method	95% CL Std Dev	
1		292.7	372.5
2		259.5	341.9
Diff (1-2)	Pooled	288.0	345.2
Diff (1-2)	Satterthwaite		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	235	9.19	<.0001
Satterthwaite	Unequal	229.16	9.31	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	133	102	1.23	0.2652

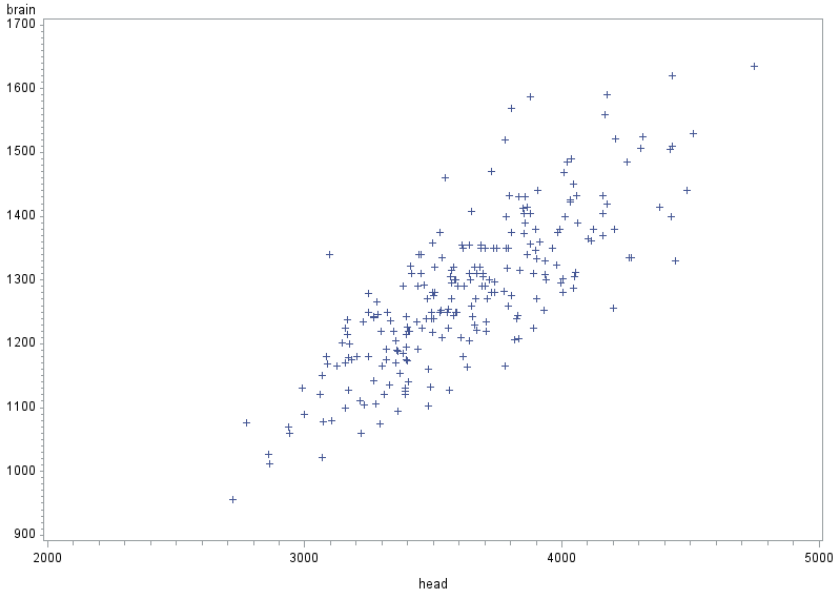

Figure 1: Scatterplot

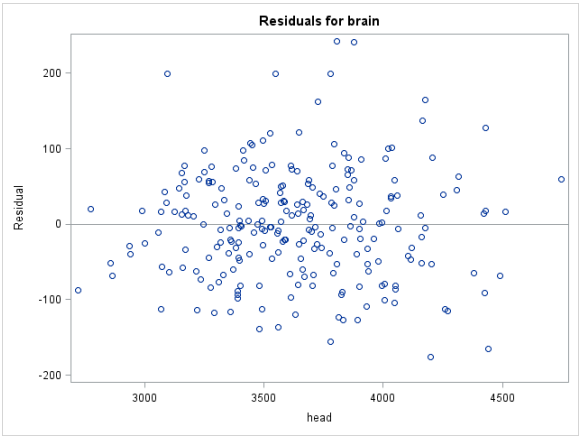

Figure 2: Residual plot

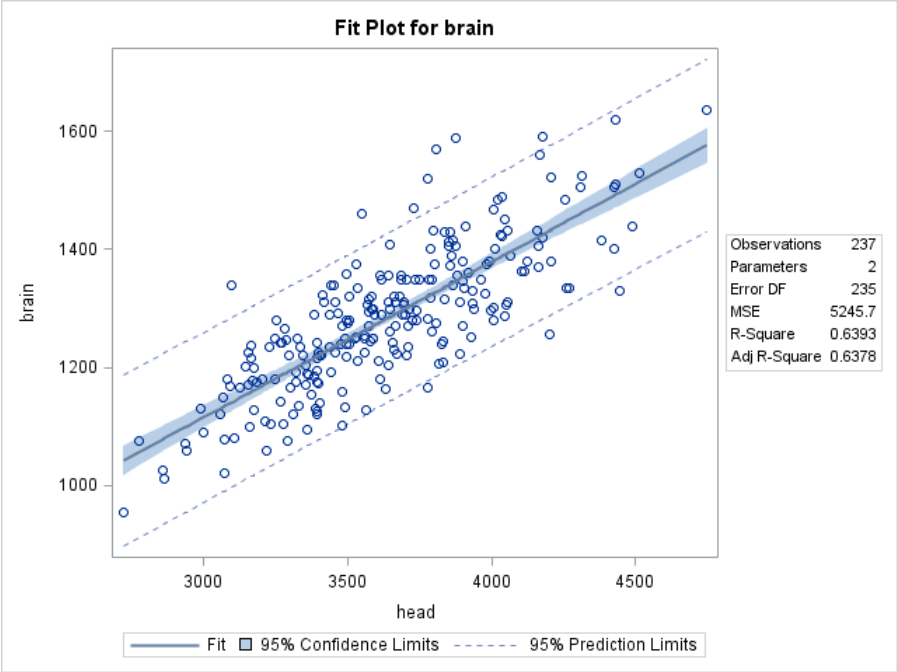

Figure 3: Fit plot

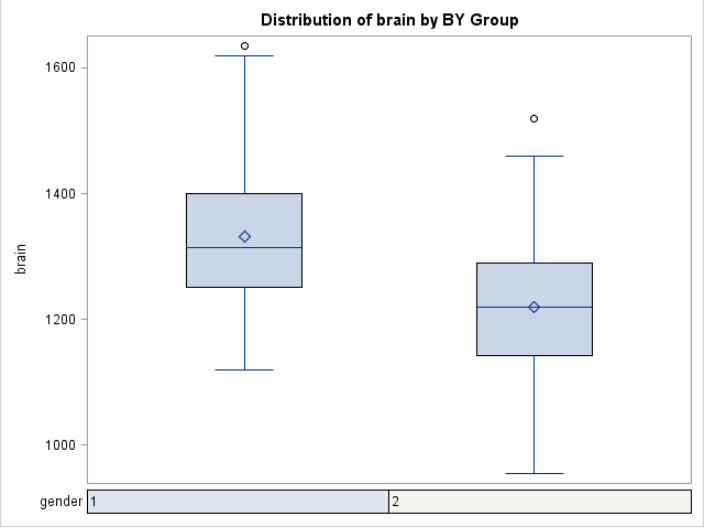

Figure 4: Boxplots of brain weight by gender (Male = 1, Female = 2)