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

Name: Solutions 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

11 observations
median is 6th

min 1st quartile median 3rd quartile max

min 6
Q1 17
Med 39
Q3 65
max 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.

outlier in vertical direction
outlier in horizontal direction (extreme predictor value)

(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)

$R^2 = 0.6393$

(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.

For each 1 cubic centimeter increase in head volume, we expect on average a 0.263 gram increase in 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.

$H_0: \beta = 0$
$H_A: \beta > 0$

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

Reject H_0. The sample slope ($b = 0.263$) is positive, so the data supports H_A. The p-value for the 2-sided test is < 0.0001, so p-value for 1-sided is $< 0.00005 < \alpha = .05$.

(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.

   $H_0: \mu_m \leq \mu_w$

   $H_A: \mu_m > \mu_w$

   (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.

   1 outlier in each sample (from boxplots). Not severe outliers and sample sizes are large enough that we may be ok.

   (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.)

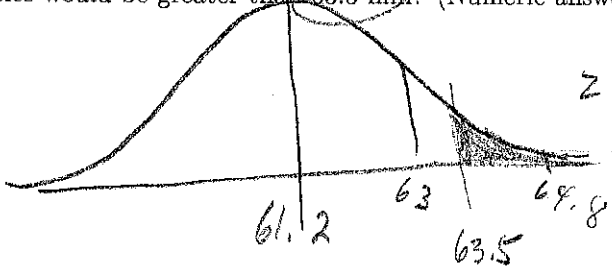
   $\mu_m - \mu_w$

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

   There is strong evidence that the mean head size is larger in men than in women.

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.)

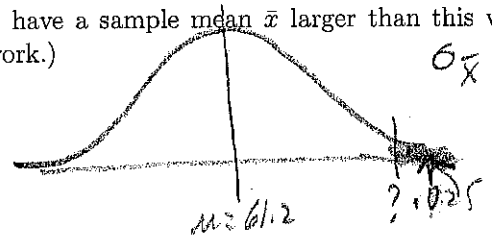


$$z = \frac{63.5 - 61.2}{1.8} = 1.278$$

$Pr(Z > 1.278)$

$= .10$

(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.)



$\sigma_{\bar{x}} = \frac{1.8}{\sqrt{10}} =$

$\bar{x} = \mu + z\sigma_{\bar{x}}$

$= 61.2 + 1.96\,(0.57)$

$= 62.32$

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.)

$\hat{p} = \frac{471}{826} = 0.57$

$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$.57 \pm 1.96\,(.0172)$

$(0.536,\ 0.604)$

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

$1.96\,(.0172) = .034$

(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.