STAT:2010/4200, Statistical Methods and Computing
Spring 2017, Instructor: Cowles
Final Exam

Name: _____

1. A dentist gathers data to determine whether people who floss their teeth at least three times a week get fewer new cavities than people who floss less frequently. During a one-month period, she asks each patient who comes in for a dental appointment whether they floss their teeth at least three times a week. She also counts how many new cavities they have developed since their last appointment.

   The dentist's investigation is (circle one):

   (a) an observational study `this one`
   (b) an experiment
   (c) neither of the above

2. For each of the following variables, state which data type it is (binary, nominal, ordinal, quantitative continuous, or quantitative discreet).

   (a) eye color (evaluated on a sample of human beings) `nominal`
   (b) boiling temperature of water (evaluated at a number of different elevations in the mountains) `quantitative continuous`
   (c) number of pigs owned (evaluated on a sample of farms in Iowa) `quantitative discreet`

3. Corn snakes are a popular pet. We wish to estimate the population mean of length in feet of adult corn snakes. Based on a sample of 15 corn snakes, we obtain a sample mean of 4.9 feet and a 95% confidence interval of (4.65, 5.15) .

   Circle **all** of the following statements that are true:

   (a) We are 95% confident that the sample mean lies in the interval.
   (b) We should reject the null hypothesis because the sample mean lies in the interval.
   (c) We are 95% confident that the population mean lies in the interval. `True`
   (d) We are 95% confident that the length of any randomly selected corn snake will lie in the interval.

4. In this class, we have used the Bonferroni correction in conjunction with `proc anova` to (circle one):

   (a) approximate the degrees of freedom in pairwise t-tests when we cannot assume equal population standard deviations

   (b) control the probability of type I error when several pairwise t-tests are conducted as a group `this one`

   (c) adjust for poor sampling designs

   (d) none of the above

5. A marketing researcher was interested in people's recall of the brand names of products advertised on television and on the Internet. He recruited 100 participants into his study. He randomly assigned 50 of them to the TV group and the other 50 to the Internet group. People in the TV group were assigned to watch a particular TV station for a particular 30-minute period. People in the Internet group were assigned to surf the web for 30 minutes, beginning from a particular web site. After completing their TV or web viewing, each subject was asked to write down the names of all products for which they had seen advertisements while watching TV or surfing the Internet. The sampling design of this study was:

   (a) one sample

   (b) two independent sample `this one`

   (c) paired sample

   (d) 3 or more independent samples

   (e) none of the above

Name: _____

6. Below are 11 quiz scores. Find the interquartile range (IQR) of these values. (Numeric answer; show your work.)

   85 79 92 98 67 81 88 71 95 77 91

   Begin by sorting the values.

   67 71 77 79 81 85 88 91 92 95 98
   Q1 M Q3

   IQR = Q3 - Q1 = 92 - 77 = 15

7. Researchers wish to estimate the population proportion of U.S. families who have at least one child in daycare. How large a simple random of families must they obtain in order to get a 99% confidence interval that is no wider than 0.1? Assume that they have no prior information about how large the proportion is. (Numeric answer; show your work.)

   z* = 2.58

   Since the desired width of the interval is 0.1, margin of error is 0.05

   m = 0.05

   n = (z*/m) ^2 (p*))1-p*)

   = (2.58/0.05)^2 * .5 * .5 = 665.64

   Round UP to 666

8. Blood alcohol level (BAC) is measured at the time a person is arrested for drunk driving. The age in years and BAC were reported for 8 people who have been convicted and jailed for drunk driving.

   SAS output follows for fitting a regression model with age as the predictor variable and BAC as the response variable.

   (a) Suppose that you believed that there was a linear relationship between age and BAC at time of arrest, and that older drunk drivers were likely to have lower BAC than younger drivers. In regression, the corresponding formal null and alternative hypotheses are statements about:

      i. a population intercept

ii. a sample intercept
iii. a population slope `this one`
iv. a sample slope
v. a population standard deviation
vi. a sample standard deviation
vii. none of the above.

(b) From the SAS output, copy the following values that relate directly to your hypothesis test:

i. test statistic -0.17

ii. p-value 0.8707

iii. confidence interval for the relevant parameter (-0.00281, 0.00244)

(c) At the .05 significance level, should the null hypothesis be rejected? (yes/no) NO

(d) Write the estimated regression equation (using values from the SAS output).

`BAC-hat = 0.21380  -0.00018218 age`

(e) What value of BAC would be predicted for a person 33 years of age? Show your work. 0.21380 -0.00018218 * 33 = 0.208

(f) Does knowing `age` help in predicting `BAC`? Explain.

`No, it helps very little.  The proportion of variability in BAC explained by age is only 0.0048 .`

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00003707	0.00003707	0.03	0.8707
Error	6	0.00771	0.00129		
Corrected Total	7	0.00775			

Root MSE	0.03585	R-Square	0.0048
Dependent Mean	0.20750	Adj R-Sq	-0.1611
Coeff Var	17.27889		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.21380	0.03920	5.45	0.0016
age	1	-0.00018218	0.00107	-0.17	0.8707

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	0.11789	0.30971
age	1	-0.00281	0.00244

9. Suppose that the weight of adult llamas follows a normal distribution with population mean 365 pounds and population standard deviation 37 pounds.

   If we get a simple random sample of 16 adult llamas, what is the probability that the sample mean of their weights will be greater than 375 pounds? (Numeric answer; show your work).

   ```
   The sampling distribution of xbar will be normal with mean 365 pounds
   and standard deviation 37 / sqrt(16) = 9.25.

   z = (375 - 365) / 9.25 =  1.08.

   Pr(Z > 1.08) = 1 - 0.8599 = 0.1401
   ```

10. An automotive magazine editor was interested in characteristics of car purchasers. In particular, he was interested in whether there were differences among the mean ages of purchasers of U.S.-made cars, purchasers of Asian-made cars, and purchasers of European-made cars. He obtained simple random samples of 13 purchasers of U.S. made cars, 10 purchasers of Asian-made cars, and 7 purchasers of European-made cars. He learned the ages of all of the purchasers.

    SAS output for an analysis of the editor's dataset is attached. You will need to refer to it in answering some of the following questions. Codes are D for U.S.-made (or domestic) cars, A for Asian, and E for European.

    (a) In this study, the populations of interest are (circle one):
        i. the 3 groups of purchasers in the dataset
        ii. all purchasers of U.S.-made cars, all purchasers of Asian-made cars, all purchasers of European- made cars `this one`
        iii. the mean age of all purchasers of U.S.-made cars, the mean age of all purchasers of Asian-made cars, the mean age of all purchasers of European-made cars
        iv. the mean ages in each group in the dataset
        v. none of the above

    (b) The SAS output includes results of the ANOVA procedure. Why was ANOVA used instead of a chi-square test or a t-test?

    ```
    The parameters of interest or population means of a
    quantitative variable, so ANOVA fits.  The chi-square test
    is for population proportions of a binary variable.  There
    are three population means to compare.  T-tests are for at
    most two.
    ```

    (c) One assumption required for ANOVA is that the data are independent simple random samples from the populations of interest. Is there any way to assess this assumption by examining the data values? Briefly explain.

No. We have to know how the data was collected in order to assess independence. We suspect that the samples are independent here because of the unequal sample sizes.

(d) The SAS output includes efforts to assess whether two other assumptions required for ANOVA have been met. Briefly state each assumption, and tell what the SAS output says as far as whether it has been met.

Assumption: Age follows a normal distribution in all three populations. The outliers in the A and D samples suggest that this may not be the case.

Assumption: The standard deviations of age are the same in all three populations. The rule of thumb that the largest sample standard deviation is not more than twice as large as the smallest sample standard deviation is met here. 17.3 < 2 * 13.5.

(e) State the null hypothesis being tested by the ANOVA procedure. Use conventional statistical symbols.

$$H_0: \quad \mu_A = \mu_D = \mu_E$$

(f) At the .05 significance level, should we reject the null hypothesis? (yes/no) Why or why not?

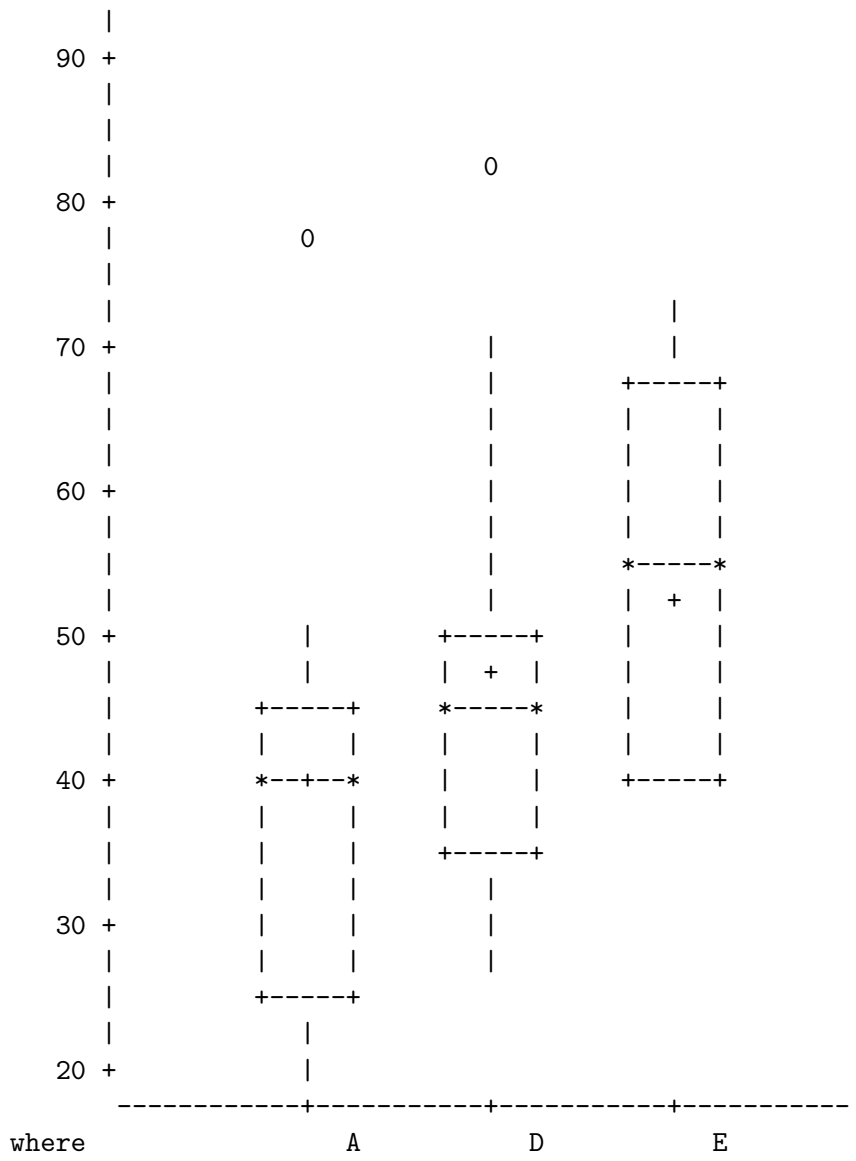
No. The p-value for the F test is 0.2248 > 0.05.

(g) Briefly explain what this result means about the average ages of buyers of the three different classifications of cars.

The data does not provide evidence that the population means of age are different among buyers of domestic, Asian, and European cars.

The UNIVARIATE Procedure
Variable: age

Schematic Plots



Analysis Variable : age

where	N Obs	N	Mean	Std Dev	Minimum
A	10	10	39.6000000	17.3089572	21.0000000
D	13	13	47.3846154	15.7402799	27.0000000
E	7	7	53.1428571	13.5330284	39.0000000

Analysis Variable : age

where	N Obs	Maximum
A	10	78.0000000
D	13	83.0000000
E	7	72.0000000

The ANOVA Procedure

Class Level Information

Class	Levels	Values
where	3	A D E

Number of Observations Read	30
Number of Observations Used	30

Dependent Variable: age

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	791.132601	395.566300	1.58	0.2248
Error	27	6768.334066	250.679039		
Corrected Total	29	7559.466667			

R-Square	Coeff Var	Root MSE	age Mean
0.104655	34.31975	15.83285	46.13333

Source	DF	Anova SS	Mean Square	F Value	Pr > F
where	2	791.1326007	395.5663004	1.58	0.2248