STAT:2010/4200, Statistical Methods and Computing
Spring 2017, Instructor: Cowles
Midterm 2

Name: ____ Course no. (2010 or 4200) _____

Show your work on any problems that involve calculations. If your answer to a multiple choice or true-false question would vary under different conditions, write an explanation. I will grade on a curve and will give partial credit wherever possible.

1. Researchers wish to test the null hypothesis that the population mean daily caffeine consumption among U.S. adults is less than or equal to 300 milligrams per day.

   (a) Write the null and alternative hypotheses using the conventional statistical symbols taught in class.

   (b) The researchers set the significance level of their test at $\alpha = 0.01$. This means that they will carry out their test in such a way that (circle the one best answer):

      i. they will have only one chance in 100 of getting a sample that causes them to reject H_0 when it is actually true
      ii. they will have only one chance in 100 of getting a sample that causes them to reject H_0 when it is false
      iii. they will have only one chance in 100 of getting a sample that causes them to fail to reject H_0 when it is actually true
      iv. the will have only one chance in 100 of getting a sample that causes them to fail to reject H_0 when it is false

   (c) A statistician helps them design the study so that they will have 90% power against the alternative that the population mean actually is 325 milligrams per day. This means that (circle the one best answer):

      i. they will have a 90% chance of getting a sample that causes them to reject H_0 if the population mean actually is 325
      ii. they will have a 90% chance of getting a sample that causes them to fail to reject H_0 when the population mean actually is 325
      iii. none of the above

   (d) The researchers believe that values in the population follow a normal distribution. What realistic procedure should they use for their hypothesis test? Circle one.

      i. one-sample z test
      ii. paired sample z test
      iii. one-sample t-test
      iv. paired sample t-test

(e) The researchers collect their data and carry out their test. They get a p-value of 0.027. This indicates (circle the one best answer):

   i. There is 0.027 probability that H_0 is true.
   ii. There is 0.027 probability that H_0 is false.
   iii. There would be 0.027 probability of getting a dataset with at least as much evidence against H_0 as the data we have, if H_0 were true.
   iv. None of the above.

2. The dataset used for this problem is described as follows:

```
Investigators suspected that Benzo(a)pyrene, or BaP, from a pipe foundry
in Phillipsburg, NJ, might be contaminating household air.  This dataset
presents data from 14 different days on samples of indoor air from
a house near the foundry and samples of outdoor air collected at the same
times.  The measures are concentrations of BaP-containing particles no
larger than 10 micrograms.

The two variables are:
   indoor air BaP
   outdoor air BaP

Reference: Lioy, PL, Walman, JM, Greenberg, A, Harkov, R and Pietarninen, C
(1988).  The total human environmental exposure study (THEES) to
Benzo(a)pyrene:  Comparison of the inhalation and food pathways.  Archives
of Environmental Health, 43: 304-312.
```

SAS output relating to both variables is attached.

(a) We wish to use these data to do a test of the null hypothesis that the population mean of indoor BaP measurements is 40. Is there anything about `indoor` data that suggests that we should *not* use a t-test to do this? Briefly explain, referring to specific SAS output.

(b) I did a t-test of the null hypothesis that the population mean of *outdoor* BaP measurements is 40. What is the p-value of my test? (numeric answer from SAS output).

(c) Should I reject H_0 at significance level $\alpha = 0.1$? (If you couldn't find the p-value, pretend it was 0.25.) Briefly explain why or why not.

(d) Briefly explain to a nonstatistician what your decision in the previous question means about outdoor BaP measurements.

(e) Suppose I had asked SAS to compute a 90% confidence interval for the population mean. Would the value 40 have been in the interval? (yes/no) Briefly explain.

3. Researchers want to study how long it takes for food to pass through the digestive systems of dogs. Specifically, they want a point estimate and a 95% confidence interval for the population mean passage time in minutes.

   (a) Suppose the researchers believe that the variable passage time follows a normal distribution in the population, and that the population standard deviation is 20 minutes. How large a sample will they need in order to obtain a 95% confidence interval that has a margin of error no larger than 2 minutes? (Numeric answer; show your work.)

   (b) The conventional symbol for the value that the researchers should use as their point estimate is (circle the one best answer):

      i. $\bar{x}$
      ii. s
      iii. μ
      iv. σ
      v. none of the above

   (c) The conventional symbol for the quantity that the researchers will be 95% confident is contained in their interval is (circle the one best answer):

      i. $\bar{x}$
      ii. s
      iii. μ
      iv. σ
      v. none of the above

4. Fitness trainers wish to investigate the effects of type of diet (high protein versus high carbohydrate) and type of exercise (weight training versus running) on muscle development in teenage girls aged 16 to 18. Sixty girls between the ages of 16 and 18 will be recruited into the study. Each girl's lean body mass will be measured at the beginning of the study. Then girls will be randomly assigned to each of four groups: high protein diet plus weight training, high protein diet plus running, high carb diet plus weight training, and high carb diet plus running. After 60 days, each girl's lean body mass will be measured again. The outcome of interest is the change in lean body mass from study entry to the end of the study.

   (a) What are the treatments in this study (circle the one best answer)?

      i. type of diet and type of exercise
      ii. high protein diet plus weight training, high protein diet plus running, high carb diet plus weight training, and high carb diet plus running
      iii. change in lean body mass
      iv. the 60 girls
      v. the 15 girls in each group
      vi. all girls aged 16 to 18

   (b) What is the population of interest (circle one)?

      i. type of diet and type of exercise
      ii. high protein diet plus weight training, high protein diet plus running, high carb diet plus weight training, and high carb diet plus running
      iii. change in lean body mass
      iv. the 60 girls
      v. the 15 girls in each group
      vi. all girls aged 16 to 18

   (c) What are the factors (circle one)?

      i. type of diet and type of exercise
      ii. high protein diet plus weight training, high protein diet plus running, high carb diet plus weight training, and high carb diet plus running
      iii. change in lean body mass
      iv. the 60 girls
      v. the 15 girls in each group
      vi. all girls aged 16 to 18

5. IQ tests are designed so that the population mean score is 100 and the population standard deviation is 15. If a simple random sample of 20 scores is drawn, what is the probability that the sample mean $\bar{x}$ will be greater than 120? (Numeric answer; show your work.)

Variable: indoor

Moments

N	14	Sum Weights	14
Mean	76.0714286	Sum Observations	1065
Std Deviation	79.109918	Variance	6258.37912
Skewness	2.0340219	Kurtosis	3.61146118
Uncorrected SS	162375	Corrected SS	81358.9286
Coeff Variation	103.994258	Std Error Mean	21.1430149

Basic Statistical Measures

Location		Variability	
Mean	76.07143	Std Deviation	79.10992
Median	50.00000	Variance	6258
Mode	10.00000	Range	275.00000
		Interquartile Range	35.00000

Note: The mode displayed is the smallest of 4 modes with a count of 2.

Tests for Location: Mu0=40

Test	-Statistic-		-----p Value------	
Student's t	t	1.706068	Pr > |t|	0.1118
Sign	M	3	Pr >= |M|	0.1460
Signed Rank	S	21	Pr >= |S|	0.1069

```
Stem Leaf                     #             Boxplot
   2 8                        1                *
   2 2                        1                *
   1
   1
   0 66789                    5             +--+--+
   0 1124444                  7             +-----+
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+2
```

Variable: outdoor

Moments

N	14	Sum Weights	14
Mean	42.2142857	Sum Observations	591
Std Deviation	18.3771538	Variance	337.71978
Skewness	0.64758432	Kurtosis	-0.7157978
Uncorrected SS	29339	Corrected SS	4390.35714
Coeff Variation	43.5330207	Std Error Mean	4.91150094

Basic Statistical Measures

Location		Variability	
Mean	42.21429	Std Deviation	18.37715
Median	39.00000	Variance	337.71978
Mode	25.00000	Range	58.00000
		Interquartile Range	31.00000

Tests for Location: Mu0=40

Test	-Statistic-		-----p Value------	
Student's t	t	0.450837	Pr > |t|	0.6595
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	4	Pr >= |S|	0.8000

```
Stem Leaf                     #             Boxplot
   7 8                        1                |
   6 57                       2                |
   5 06                       2             +-----+
   4 01                       2             |  +  |
   3 58                       2             *-----*
   2 04557                    5             +-----+
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```