Name: _____

Biostatistics 1st year Comprehensive Examination: Applied in-class exam

Thursday May 31st, 2018: 9am to 1pm

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. There are five questions and 6 pages.
3. Answer to the best of your ability. Read each question carefully.
4. Be as specific as possible and write as clearly as possible.
5. This is a closed-book in-class examination. **NO BOOKS, NO NOTES, NO INTERNET DEVICES, NO CALCULATORS, NO OUTSIDE ASSISTANCE.**
6. You may leave the examination room to use the restroom or to step out into the hallway for a short breather. **HOWEVER, YOU MUST LEAVE YOUR CELL PHONE AND ALL EXAM MATERIALS IN THE EXAMINATION ROOM.** If there is an emergency, please discuss this with the exam proctor.
7. ***Vanderbilt's academic honor code applies; adhere to the spirit of this code.***

Question	Points	Score	Comments
1	50		
2	50		
3	50		
4	50		
5	50		
Total	**250**		

1. Researchers have developed a new measurement of lung expiratory function, which they call the **BRETH** test. The BRETH test scores are known to be normally distributed, and the true mean BRETH score for a patient is of clinical importance. However, the BRETH test suffers from high measurement error, i.e. the variance of multiple BRETH tests from the same patient can be non-trivial. Hence, researchers created a protocol for taking the mean of four BRETH tests administered far enough apart to be treated as independent measurements. The **BRETH4** score is the average of these four BRETH tests, which are administered at least one hour apart from each other. Patient John Doe was tested and had a BRETH4 score of 50. He belongs to a group of patients who are known to all have a standard deviation in their individual BRETH scores equal to 12.

    a. Propose an interval method that, when used repeatedly on patients like John Doe, would capture a patient's true mean BRETH score with a 68% chance. Write a formula filling in numbers where appropriate.

    b. For Mr. Doe, calculate the interval you proposed in part (a).

    c. Can you say there is a 0.68 probability that the interval reported in part (b) includes Mr. Doe's true mean BRETH score? Answer yes or no and briefly explain your answer.

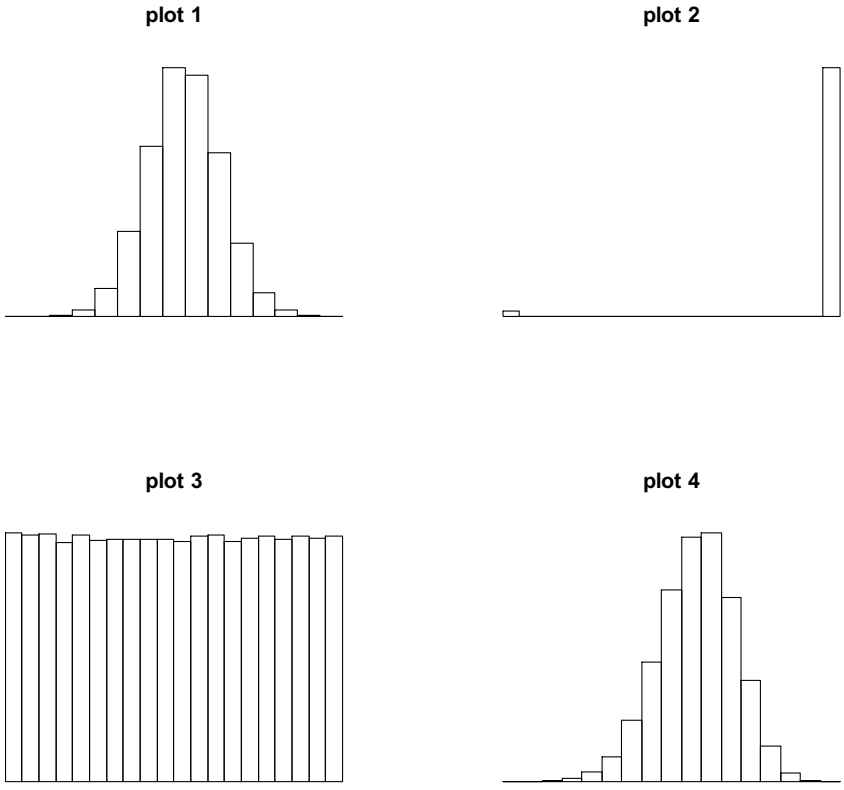
    d. Using the BRETH test, suppose you wanted to have an interval *with the same width as the interval in part (a)*, however it would have a 99.7% chance of capturing a patient's true BRETH score. Explain how you could achieve this or why it is impossible.

    e. Suppose patient Jane Smith comes from a different group of patients where a patient's BRETH scores are known to be normally distributed but their variance is unknown. Propose an interval method that, when used repeatedly on patients like Ms. Smith, would capture a patient's true mean value for the BRETH4 test with a 95% chance. Write a formula filling in numbers where appropriate.

    f. What conditions must hold for the interval you proposed in part e to have *exactly* a 95% a chance of including the true mean BRETH score or explain why this interval can only have an *approximately* 95% chance of including the true mean.

2. Consider the following R code and graphs. The code creates large samples of four different distributions. The four samples are named *a*, *b*, *c*, and *d*.

```
set.seed(1)
a <- runif(10^5, -10, 10)
b <- rbinom(10^5, 1, 0.98)
sum777 <- function(x){ sum( sample(x, 777, replace=T) ) }
c <- rep(NA,10^5)
for(i in 1:10^5){ c[i] <- sum777(a) }
d <- rep(NA,10^5)
for(i in 1:10^5){ d[i] <- sum777(b) }
```



Question 2 is continued on the next page.

Question 2 continued:

a1-a4. Answer questions 1-4 below for **sample *a***.

b1-b4. Answer questions 1-4 below for **sample *b***.

c1-c4. Answer questions 1-4 below for **sample *c***.

d1-d4. Answer questions 1-4 below for **sample *d***.

Questions 1-4 to be answered for each of the four samples

1. State **which plot** corresponds to this sample.

2. **Name the underlying distribution** of this sample. If you cannot name the exact distribution, suggest an approximate distribution and explain why that approximate distribution is a good choice.

3. **Estimate the mean** of the underlying distribution or explain why it can't be estimated from the information given. Where appropriate, insert numbers into formulas, but don't try to solve for the final answer.

4. **Estimate the standard deviation** of the underlying distribution or explain why it can't be estimated from the information given. Where appropriate, insert numbers into formulas, but don't try to solve for the final answer.

3. A leading cough drop company is developing a new lozenge, which has the development codename LX200. The goal is to create a lozenge that is more effective than their bestselling lozenge, codenamed GOLD1. They perform a randomized controlled trial of patients with a recently diagnosed upper respiratory tract infection. Patients are given a lozenge and observed how many times they cough during a 20-minute period. The observed number of coughs per patient range from 9 to 61 and their distribution has a slight positive skew. The mean number of coughs and standard deviation for the LX200 and GOLD1 lozenges are summarized below.

Treatment Group	**GOLD1**	**LX200**
Sample Size	221	193
Mean (coughs/20 min)	19	27
SD (coughs/20 min)	6	5

   a. The research team wants a hypothesis test of no difference in true means with a 5% significance level. Is an equal variance t-test appropriate in this situation? Answer yes/no and explain why.

   b. Using standard mathematical notation, write down the null hypothesis for the equal variance t-test and the equal variance t-test test statistic.

   c. What is the sampling distribution and degrees for freedom for the equal variance t-test, in this setting, when the null hypothesis is true?

   d. Using standard mathematical notation, write down the formula for the test statistic for an *unequal* variance t-test. Fill in the numbers from this study, but don't reduce/calculate it.

   e. What sampling distribution, or approximation thereof, would you use for the *unequal* variance t-test when the null hypothesis is true? Justify your answer.

   f. Name a nonparametric test that would be applicable to this data, and using standard mathematical notation, write down that test's null hypothesis.

   g. For this dataset, the *p*-values for the equal variance t-test, unequal variance t-test, and an appropriate nonparametric test are all <0.0001. Based on this study, what conclusion and recommendation should you make to the company regarding LX200?

   h. Set up the equation you would solve to create a 1/8th likelihood support interval for the difference in sample means. Hint: Create an approximate solution by treating the sample standard deviations as if they were the true standard deviations.

4. National health, welfare, and education statistics for 199 places, mostly UN members, were collected. Measured social and health variables included *fertility* (number of children per woman), *ppgdp* (per capita gross domestic product in US dollars), and *lifeExpF* (female life expectancy in years). Additionally, for the resident population of each place, the percent urban (*pcturban*), versus rural, was collected. The results of fitting the following model are provided below.

$$\log(fertility) = \beta_0 + \beta_1 \log(\text{ppgdp}) + \beta_2 \text{lifeExpF} + \beta_3 \text{pcturban} + e$$

```
. regress lfert lppg lifeexpf pcturban

      Source |       SS           df       MS      Number of obs   =       199
-------------+----------------------------------   F(3, 195)       =    147.24
       Model |  27.1936584         3  9.06455279   Prob > F        =    0.0000
    Residual |  12.0045527       195  .061561809   R-squared       =    0.6937
-------------+----------------------------------   Adj R-squared   =    0.6890
       Total |   39.198211       198  .197970763   Root MSE        =    .24812

------------------------------------------------------------------------------
       lfert |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lppg |  -.0758004   .0214944    -3.53   0.001    -.1181917    -.033409
    lifeexpf |  -.0283677   .0027458   -10.33   0.000     -.033783   -.0229524
    pcturban |   .0009794   .0011363     0.86   0.390    -.0012617    .0032205
       _cons |   3.547848   .1355562    26.17   0.000     3.280503    3.815192
------------------------------------------------------------------------------
```

a. What are the usual assumptions made when fitting a linear regression such as this?

b. Provide an interpretation for $\hat{\beta}_1$ in terms of *fertility* and *ppgdp*. [Hint: Consider a proportional change in *ppgdp*, say, from *ppgdp* to *ppgdp*× $\boldsymbol{a}$.]

c. Provide an interpretation of *R-squared*.

d. What is *Root MSE* and what does it estimate? Explain.

e. Using the correlation table below, discuss the relationship between log-*fertility* and each of log-*ppgdp*, female life expectancy, and percent of population that is urban.

```
. corr lfert lppg lifeexpf pcturban
(obs=199)

             |    lfert     lppg lifeexpf pcturban
-------------+------------------------------------
       lfert |   1.0000
        lppg |  -0.7252   1.0000
    lifeexpf |  -0.8194   0.7723   1.0000
    pcturban |  -0.5352   0.7483   0.6013   1.0000
```

f. Note that the interval estimate of the regression coefficient on *pcturban*, β_3, includes zero. Using the table below, determine the value of R^2 for the fitted model that excludes *pcturban*.

```
. pcorr lfert lppg lifeexpf pcturban
(obs=199)

Partial and semipartial correlations of lfert with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable |    Corr.         Corr.      Corr.^2       Corr.^2          Value
------------+------------------------------------------------------------------
       lppg |  -0.2449       -0.1398       0.0600        0.0195         0.0005
   lifeexpf |  -0.5948       -0.4094       0.3537        0.1676         0.0000
   pcturban |   0.0616        0.0342       0.0038        0.0012         0.3898
```

5. Suppose we have a multiple regression model with 4 regressors. Call this Model 1. We have a 5th regressor that can be expressed as a linear combination of the other regressors plus an error with mean 0 and constant variance.

   a. Suppose the correlation between the new variable and its regression on the other regressors is 0.999. How are R^2 and R^2_{adj} affected by adding the new variable to Model 1? Explain.

   b. Now suppose the correlation between the new variable and its regression on the other regressors is 1. How is R^2 affected by adding the new variable to Model 1? Can we estimate all 5 regression coefficients? Why or why not? Explain.

   c. Lastly, suppose the correlation between the new variable and its regression on the other regressors is 0. How are R^2 and R^2_{adj} affected by adding the new variable to Model 1? Explain.