Name: ____

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

June 7th, 2017: 9am to 1pm

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. There are six questions and 8 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	40		
2	40		
3	40		
4	40		
5	40		
6	40		
Total	**240**		

1. These are *True or False* questions. Use a separate sheet of paper to indicate which option (*True or False*) you are choosing for each answer. **Write a brief justification for each answer (1-3 sentences).**

   a. **True** or **False**: As a general convention in randomized trials, a p-value < 0.05 can be interpreted as meaning the observed effect is meaningful.

   b. **True** or **False**: Out of 1,000 randomized controlled trials that tested two biologically equivalent therapies against each other, we would expect 50 of the trials to yield p-values < 0.05.

   c. **True** or **False**: When comparing two randomized trials that tested the same clinical question, the trial with a p-value < 0.001 provides stronger evidence for a differential effect than the trial with a p-value < 0.05.

   d. **True** or **False**: Among classical hypothesis testing, likelihood inference, and Bayesian inference, only Bayesian inference allows the analyst to provide an estimate of the probability that treatment A is more effective than treatment B.

   e. **True** or **False**: If a p-value requires adjustments for multiple comparisons or multiple looks to be considered "valid" in the traditional frequentist sense, then the corresponding confidence interval must also be adjusted to remain "valid" in the same sense.

   f. **True** or **False**: The most powerful rejection region for a hypothesis test is always in the tails of the test statistic distribution under H_0.

   g. **True** or **False**: Given a model, null hypothesis, data, and the absence of adjustments for early looks or multiple comparisons, it remains possible for the *p-value* from a significance test to differ from the *p-value* from a hypothesis test.

   h. **Agree** or **Disagree**: In a 1978 JASA paper, George E. P. Box said "All models are wrong but some are useful." (Explain why you agree or disagree with this statement in 5 or fewer sentences).

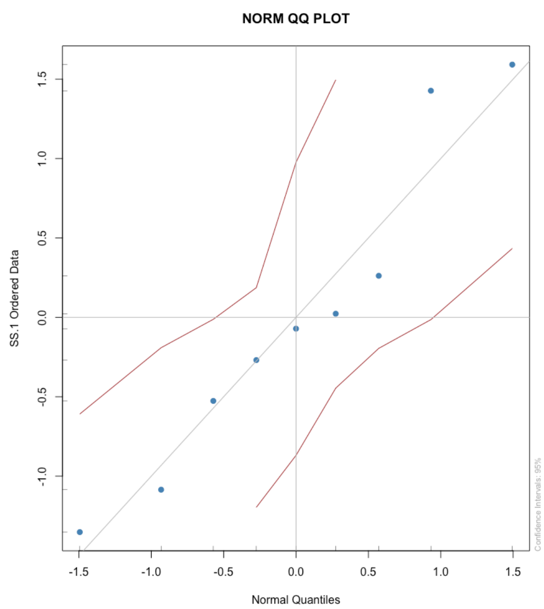
2. The Harrellet drug company is testing a new antihypertensive drug (labeled A) against standard therapy (labeled B). They conducted a study where systolic blood pressure was measured at the time of randomization (SBP0) and again 12 months later (SBP12). The following output is from a simple analysis performed in R.

Output from an R session

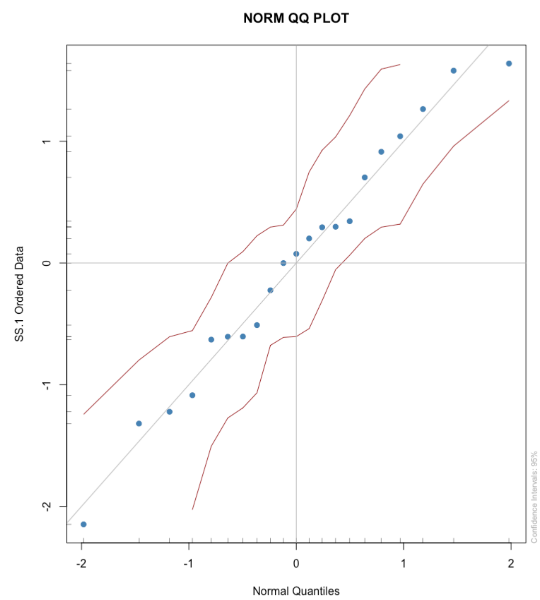
```
> library(fBasics)

> round( c( nrow(xA), mean(xA$SBP12), sd(xA$SBP12) ), 1 )
[1]   9.0 132.5  10.6

> round( c( nrow(xB), mean(xB$SBP12), sd(xB$SBP12) ), 1 )
[1]  21.0 141.4  16.8

> t.test(xA$SBP12,xB$SBP12,var.equal=F)

      Welch Two Sample t-test

data:  xA$SBP12 and xB$SBP12
t = -1.7374, df = 23.523, p-value = 0.09539
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.409950   1.677457
sample estimates:
mean of x mean of y
 132.4902  141.3564

> t.test(xA$SBP12,xB$SBP12,var.equal=T)

      Two Sample t-test

data:  xA$SBP12 and xB$SBP12
t = -1.4537, df = 28, p-value = 0.1572
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.359996   3.627503
sample estimates:
mean of x mean of y
 132.4902  141.3564
```

Q-Qplots from R session

```
> qqnormPlot(xA$SBP12)
```




```
> qqnormPlot(xB$SBP12)
```



a. The analysis above conducts two statistical tests. Name them and list their assumptions. Which assumptions appear to be satisfactorily met in these data? If certain assumptions cannot be assessed from this analysis, note this and explain how they should be assessed (if at all).

b. The two tests are often said to have the same null and alternative hypotheses. Write down each null hypothesis using the appropriate mathematical notation. Be as specific as possible. Are they identical? Explain, if no.

c. On the basis of this output, which statistical test would you use? Explain your reasoning. Be sure to discuss the parts of this output that are generally considered acceptable to use for selecting a test and what parts of this output are generally not considered acceptable to use for selecting a test.

d. Write a brief conclusion for this study on the basis of this analysis. Be sure to refer to the *p*-value, the effect size, and the confidence interval.

3. Using the same data from question (2), the researchers categorized the outcomes as improved (*defined as SBP12 dropping by more than 2 mmHg from baseline*), stayed the same (*defined as SBP12 remaining within 2 mmHg from baseline*), and got worse (*defined as SBP12 increasing by more than 2 mmHg from baseline*). A different analysis was conducted on the categorized data, and that analysis is shown below. (Note: answer all questions using the below analysis on the categorized outcome).

<u>Output from another R session</u>

```
> counts <- c(sum( xA$SBP12-xA$SBP0 < -2 ),
+             sum( xA$SBP12-xA$SBP0 >= -2 & xA$SBP12-xA$SBP0 <= 2 ),
+             sum( xA$SBP12-xA$SBP0 > 2 ),
+             sum( xB$SBP12-xB$SBP0 < -2 ),
+             sum( xB$SBP12-xB$SBP0 >= -2 & xB$SBP12-xB$SBP0 <= 2 ),
+             sum( xB$SBP12-xB$SBP0 > 2 )
+             )
> counts <- matrix(counts, nrow=3)
> counts
     [,1] [,2]
[1,]    9    9
[2,]    0    4
[3,]    0    8

> chisq.test(counts)

      Pearson's Chi-squared test

data:  counts
X-squared = 8.5714, df = 2, p-value = 0.01376

Warning message:
In chisq.test(counts) : Chi-squared approximation may be incorrect

> fisher.test(counts)

      Fisher's Exact Test for Count Data

data:  counts
p-value = 0.01442
alternative hypothesis: two.sided
```

Question 3 (continued)

a. The analysis above conducts two statistical tests. Name them and list their assumptions. Which assumptions appear to be satisfactorily met in these data? If certain assumptions cannot be assessed from this analysis, note this and explain how they should be assessed (if at all).

b. The two tests are often said to have the same null and alternative hypotheses. Write down each null hypothesis using the appropriate mathematical notation. Be as specific as possible. Are they identical? Explain, if no.

c. On the basis of this output, which statistical test would you use? Explain your reasoning. Be sure to discuss the parts of this output that are generally considered acceptable to use for selecting a test and what parts of this output are generally not considered acceptable to use for selecting a test.

d. The analysis does not provide an estimate of the effect size. Using the counts table, propose a point estimate for the effect that is appropriate for this dataset.

e. Explain how to calculate a 95% confidence interval for the point estimate you proposed in part (d). If there is a formula or algorithm to use, provide it. (You are not required to solve this by hand; just define your solution using clear notation).

f. Write a brief conclusion for this study on the basis of this analysis. Be sure to refer to the *p*-value, the effect size, and the confidence interval from parts (d) and (e).

g. It is often said that one should never categorize a continuous variable. What is sacrificed by categorizing? What can be gained? Comment on whether the analysis given in this question is valid and if/how it could outperform the analysis given in question (2). (Keep your answer to 2 paragraphs or less).

4. The conditional expectation of continuous measure Y (cm) given another continuous measure X (cm) is modeled as a linear relation. The 25 observed values of each variable were standardized to have mean 0 and standard deviation 1. These standardized variables are denoted as ZX and ZY. The regression of Y on X and ZY on ZX is:

```
. regress y x

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =    7.17
       Model |   6.4777658     1   6.4777658           Prob > F      =  0.0134
    Residual |  20.7835533    23  .903632754           R-squared     =  0.2376
-------------+------------------------------           Adj R-squared =  0.2045
       Total |  27.2613191    24   1.1358883           Root MSE      =   .9506

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.4693033   .1752819    -2.68   0.013    -.8319015   -.1067052
       _cons |   .1282957   .1901395     0.67   0.507    -.2650378    .5216293
------------------------------------------------------------------------------

. regress zy zx

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =    7.17
       Model |   5.7028193     1   5.7028193           Prob > F      =  0.0134
    Residual |   18.297181    23  .795529608           R-squared     =  0.2376
-------------+------------------------------           Adj R-squared =  0.2045
       Total |  24.0000003    24  1.00000001           Root MSE      =  .89192

------------------------------------------------------------------------------
          zy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          zx |  -.4874602   .1820634    -2.68   0.013     -.864087   -.1108335
       _cons |  -1.98e-09   .1783849    -0.00   1.000    -.3690174    .3690173
------------------------------------------------------------------------------
```

a. Interpret the estimated coefficient for X.

b. Interpret the estimated coefficient for ZX.

c. Find the conditional mean of Y given $X = 1$.

d. What is the (Pearson-product) correlation between Y and X? How is this related to the coefficient of X and the coefficient of ZX?

e. Is there evidence of a linear relationship between X and Y? Can you conclude the relationship is linear? Explain.

f. Explain why the R-squared does not change even though the estimated coefficients do change.

5. The true mean function corresponding to the regression of Y on X_1 and X_2 is

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

However, an analyst fits a simple linear regression model of Y on X_1, ignoring X_2.

For this problem, consider the following three cases:

i. X_1 and X_2 are independent.
ii. $E[X_2|X_1] = \gamma_0 + \gamma_1 X_1$ where $\gamma_1 \neq 0$.
iii. X_2 given X_1 has an Exponential distribution with parameter λX_1 (assume $X_1 > 0$). That is, $E[X_2|X_1] = (\lambda X_1)^{-1}$.

a. State at least 3 key assumptions of the analyst's reduced model. Explain how each assumption could be checked or if it is possible to do so.

b. Show that for cases (i) and (ii) the correct form of the mean function $E[Y|X_1]$ is $\alpha_0 + \alpha_1 X_1$. For each of cases (i) and (ii), find α_0 and α_1 in terms of $\beta_0, \beta_1, \beta_2, \gamma_0$, and γ_1.

c. What is the form of the mean function for case (iii)? Is this a linear model? Is it possible to find a unique least squares solution for each of the 4 parameters? Explain.

d. Suppose for case (ii) that (X_1, X_2) has a bivariate normal distribution. What are α_0 and α_1 as functions of the bivariate normal parameters? Use the following notation for the bivariate normal parameters: means μ_1, μ_2 and Variances σ_1^2, σ_2^2 and correlation ρ.

6. Address each question in less than 3 paragraphs.

   a. Assume that X, the independent variable in a simple regression, can be transformed in a way that dramatically improves the fit of a regression model. Furthermore, assume that a transformation of the dependent variable – leaving X unchanged – produces the same improvement. Would it make any difference which transformation is used? Explain.

   b. Zou, Tuncall, and Silverman (Radiology 2003, 227(3), 617-628) provide a regression example where they consider the effect of radiation dose received by a patient on the total CT flouroscopy procedure time. They transform the total time measures to "make the data appear normal, for more efficient analysis...". They go on to state "However, normality is not necessary in the subsequent regression analysis". They go on to conclude that the "Effects of both the intercept and slope are statistically significant ($P < .005$)". Discuss the importance of their transformation and normality assumption in this context.

   c. Zhou, Stroupe, and Tierney (JRSS, Series C: Appl Stat 2001, 50:303-312) state: "When the dependent variable has been log-transformed, a regression coefficient can no longer be interpreted as the change in the dependent variable given a 1-unit change in the corresponding independent variable." There are at least two things wrong with this statement. Identify what is wrong with this statement. Discuss each error and clarify.