Name: _____

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

June 8th, 2016: 9am to 1pm

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. There are four questions and 9 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	42		
2	42		
3	42		
4	84		
Total	**210**		

*** Note: Every sub-question is worth 6 points. There are 35 sub questions for 210 points.*

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 new blood pressure medication is tested against a placebo. A Wilcoxon-Mann-Whitney test on systolic blood pressure (SBP) has a p-value = 0.001.

a. **True** or **False**: We can conclude at a 1% significance level that the true medians of the drug and placebo exposed populations are different.

A new blood pressure medication is tested against a placebo. An unequal variance two-sample t-test on systolic blood pressure (SBP) has a p-value = 0.001.

b. **True** or **False**: We should conclude at a 1% significance level that the sample means of the drug and placebo groups are different.

A new blood pressure medication is tested against a placebo. The mean and a BCA bootstrapped 95% confidence interval are 120 (110, 129).

c. **True** or **False**: We can conclude at a 5% significance level that the true mean SBP of the drug exposed populations are different.

A new blood pressure medication is tested against a placebo in a randomized controlled trial. The number of patients achieving SBP < 130 for each exposure will be used in a Chi-squared test, which will be evaluated at a 5% level.

d. **True** or **False**: The Type I error rate for this experiment is exactly 5%.

A new blood pressure medication is tested against a placebo. The number of patients achieving SBP < 130 for the drug exposure will be used to find an Exact Binomial 95% confidence interval to estimate the true percentage achieving controlled BP.

e. **True** or **False**: The coverage rate for the confidence interval being used here can be assumed to be $\geq$95%.

A new blood pressure medication is tested against a placebo. The number of patients achieving SBP < 130 for the drug exposure will be used with a non-informative prior to find a 95% credible interval to estimate the true percentage achieving controlled BP.

f. **True** or **False**: The coverage rate for the credible interval being used here can be assumed to be $\geq$95%.

g. **True** or **False**: When two studies yield the exact same *p*-value, both studies have generated equivalent amounts of statistical evidence.

2. A large "new-user" propensity score matched study using electronic health records data compared a dual therapy regimen of an antihypertensive medication plus a diuretic administered as individual pills versus as one combination pill (two pills vs one pill). Systolic blood pressure (SBP) was observed approximately six months after randomly assigned therapy was begun. A table summarizing key data from this study follows; STATA output for these data are on the following page.

		Systolic Blood Pressure (SBP)	
Treatment	**N**	**Mean**	**Standard Deviation**
Two Pills	400,000	125	15
One Pill	400,000	124	13

a. Using standard notation, write out the null and alternative hypotheses for a two-sample equal variance t-test of the mean difference in SBP for two pills vs one pill.

b. Write out a test statistic that can be used to test the hypothesis from part (a) and insert the appropriate numbers from the table above (do not solve it).

c. Interpret the STATA output using a *formal hypothesis test* with a pre-specified size of 5%. Provide a correct interpretation that is also suitable for a non-statistician.

d. Interpret the STATA output using a *formal significance test*. Provide a correct interpretation that is also suitable for a non-statistician.

e. Interpret the STATA output using an approach other than classical testing. Provide a correct interpretation that is also suitable for a non-statistician. If your ideal statistics are not reported here, define those missing statistics and provide an example to illustrate how they would be interpreted.

f. The sample standard deviations are very close in this example. What would be a potential advantage of using an equal-variance t-test in this case?

g. Histograms of SBP in both arms show the distributions are positively skewed. What concerns, if any, do you have about using a two-sample unequal variance t-test in this case?

STATA Output for Question #2

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |  400000         124    .0205548          13    123.9597    124.0403
       2 |  400000         125    .0237171          15    124.9535    125.0465
---------+--------------------------------------------------------------------
combined |  800000       124.5    .0157023    14.04456    124.4692    124.5308
---------+--------------------------------------------------------------------
    diff |                  -1    .0313847               -1.061513    -.938487
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t = -31.8626
Ho: diff = 0                                     degrees of freedom =   799998
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000


Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |  400000         124    .0205548          13    123.9597    124.0403
       2 |  400000         125    .0237171          15    124.9535    125.0465
---------+--------------------------------------------------------------------
combined |  800000       124.5    .0157023    14.04456    124.4692    124.5308
---------+--------------------------------------------------------------------
    diff |                  -1    .0313847               -1.061513    -.938487
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t = -31.8626
Ho: diff = 0                     Satterthwaite's degrees of freedom =   784157
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

3. Consider the following R code:

```
# initialize variables
reps <- 10^5
n <- 30
p1 <- 0.20
p2 <- 0.20
r.s <- rep(NA,reps)
r.t <- rep(NA,reps)
d.st <- rep(NA,reps)

# run simulation study
for(i in 1:reps){
  x1 <- rbinom(n,1,p1)
  x2 <- rbinom(n,1,p2)

  p <- mean( c(x1,x2) )
  a <- (sum(x1)-n*p)^2 / (n*p)
  b <- (sum(x2)-n*p)^2 / (n*p)
  c <- (sum(1-x1)-n*(1-p))^2 / (n*(1-p))
  d <- (sum(1-x2)-n*(1-p))^2 / (n*(1-p))

  s <- a+b+c+d

  v <- (var(x1)+var(x2))/2
  t <- ( (mean(x1)-mean(x2)) / sqrt((1/n+1/n)*v) )^2

  r.s[i] <- ( s > qnorm(0.975)^2 )
  r.t[i] <- ( t > qnorm(0.975)^2 )

  d.st[i] <- abs(s-t)
}

# calculate results
mean(r.s)
mean(r.t)
mean(d.st)
```

Question 3 (parts e through g continue on the next page):

a. Describe the values `s` will take as explicitly as possible.

b. Describe the values `t` will take as explicitly as possible.

c. Make an educated guess for the value of `mean(r.s)`. Explain your guess or explain why no reasonable guess can be made.

d. Make an educated guess for the value of `mean(r.t)`. Explain your guess or explain why no reasonable guess can be made.

Question 3 continued:

e. Make an educated guess for the value of `mean(d.st)`. Explain your guess or explain why no reasonable guess can be made.

f. Set `n <- 10^9` and make an educated guess for the value of `mean(d.st)` with this change. Explain your guess or explain why no reasonable guess can be made.

g. Set `n <- 10^9` and `p2 <- 0.21` and make an educated guess for the value of `mean(r.s)` with these changes. Explain your guess or explain why no reasonable guess can be made.

4. A prediction model is developed for outcome Y using predictors X_1 and X_2. Both predictors are ratio scale measures: X_1 is continuous, but X_2 is discrete and only takes the values 1, 2, 3. Consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

   a. State at least 3 key assumptions that would made for a typical multiple regression model of this sort. Explain how each assumption could be checked with a given dataset, if it is possible to do so.

   b. Suppose the predictor X_1 is replaced with $X_1^* = X_1 - 1.5$ and the original model is refit. Denote the new coefficients as β_k^* for $k = 0,1,2,3$. How do these new coefficients relate to the original coefficients β_k? For each $k = 0,1,2,3$, find an expression for β_k^* as a function of β_k.

   c. Collect the X_1 terms and rewrite the original model so that it looks like a simple linear regression of Y on X_1 where X_2 is treated as a constant. What are the intercept and slope parameters in this model? Interpret the coefficients of this model and explain why the model expressed in this form might be useful.

   For parts d through g, please refer to Table 1.

   d. Provide an interpretation of the estimated coefficient on X_1. Also interpret the corresponding confidence interval.

   e. Is the predictor X_2 important to the model? Explain.

   f. Refer to rewritten model in part (c). Using the estimated coefficients, sketch the estimated mean function for each value of X_2. What is the role of the interaction coefficient in how these lines are related? Explain.

   g. What is correlation between Y and its predicted value? Explain.

Question 4 continued on next page (parts h through n).

Question 4 continued.

For parts h through n, please use Tables 1 &2 and Figures 1 & 2.

h. Suppose we were to regress Y on X_1 alone. How would the *R-squared* for this simple regression model compare to the proposed model? Of these two regression models, which do you recommend? Explain. (See Table 2.)

i. A colleague recommends that X_2 be treated as a categorical variable. How would this affect the regression results? Do you agree with your colleague's recommendation? Explain. (See Figures 1& 2)

j. Compare and discuss the graph you constructed in part (f) with Figure 2. How are they different? How are they similar?

For parts k through m, refer to Table 3 and Figure 3. A new variable, X_4, is to be considered for the model. It's correlation with Y is 0.89 and it is correlated with both X_1 and X_2: 0.49 and 0.69, respectively.

k. Table 3 shows the partial and semi-partial correlations. Interpret these correlations and discuss their influence on you when building a parsimonious prediction model.

l. An *added variable plot* related to adding X_4 to the proposed model is shown in Figure 3. Explain how this plot is derived. Should you add X_4 to the model?

m. The fitted model with X_4 being the only prediction variable yields an *adjusted R-squared* of 0.7789. Adding X_1 yields an *adjusted R-squared* of 0.9131. Adding X_2 yields an *adjusted R-squared* of 0.9162. What is the *adjusted R-squared* after adding the interaction between X_1 and X_2? [Hint: The information in Table 3 will be useful.]

n. Using the information available to you, which model would you choose as your final prediction model? Explain.

Tables

Table 1: Regression table

```
. regress Y c.X1##c.X2
```

Source	SS	df	MS		
				Number of obs	= 30
				F(3, 26)	= 34.67
Model	1114.13129	3	371.377095	Prob > F	= 0.0000
Residual	278.532822	26	10.7128008	R-squared	= 0.8000
				Adj R-squared	= 0.7769
Total	1392.66411	29	48.0229003	Root MSE	= 3.273

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X1	-1	.6498699	-1.54	0.136	-2.335827	.3358266
X2	1.5	1.254622	1.20	0.243	-1.078913	4.078913
c.X1#c.X2	1	.3148409	3.18	0.004	.3528352	1.647165
_cons	8.88e-16	2.050264	0.00	1.000	-4.214378	4.214378

Table 2: Correlation matrix

```
. cor Y X1 X2
(obs=30)
```

	Y	X1	X2
Y	1.0000		
X1	0.7394	1.0000	
X2	0.7694	0.5780	1.0000

Table 3: The variable X1X2 is defined as $X_1 \times X_2$.

```
.         pcorr Y X4 X1 X2 X1X2
(obs=30)

Partial and semipartial correlations of Y with
```

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
X4	0.7911	0.3538	0.6258	0.1252	0.0000
X1	0.2598	0.0736	0.0675	0.0054	0.1906
X2	0.2060	0.0576	0.0424	0.0033	0.3025
X1X2	0.0656	0.0180	0.0043	0.0003	0.7451

Figures

Figure 1: Schematic boxplots of Y over X_2.

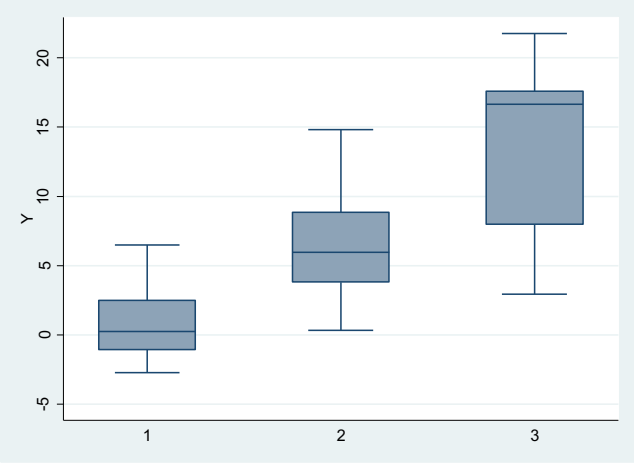

Figure 2: Each line is the least squares regression over the points of the same color.

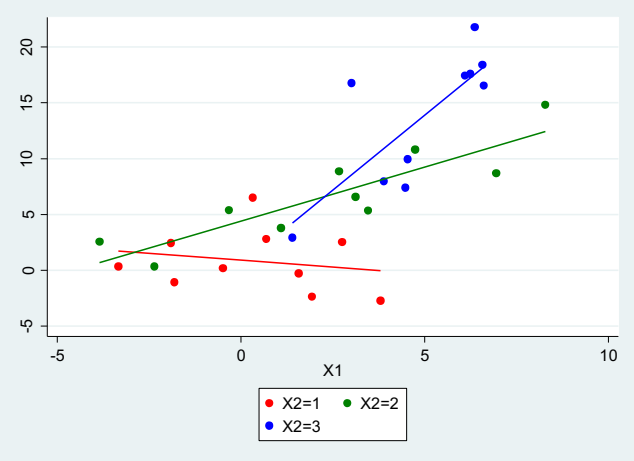

Figure 3: The avplot for *X*_4 after fitting the proposed model.