NAME: _____ **EXAM NUMBER:** _____

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 19th, 2016

Instructions:

This three hour examination consists of five questions numbered 1-5. Answer all five questions. Each question is worth the same number of points.

This is a closed book, closed notes exam. You may use a calculator and the formula sheets and tables that have been provided.

You may begin your answer on the sheet on which the question is typed and continue on additional pages if necessary. Do not write the answer to more than one question on a single page. Do not write on the back of any sheet. Be sure to write your exam number on each sheet you turn in, and if you use additional sheets be certain to mark them carefully with the question number as well as your exam number.

It is your responsibility to check that you have the whole exam. There are 7 question pages.

Question 1

A pharmaceutical company developed a new medication (called Drug A) to increase levels of a certain enzyme in the blood compared to the current standard medication (called Drug S) for persons with enzyme deficiency. The company performed a randomized clinical trial to compare Drug A to Drug S. A total of $2,400$ persons with the enzyme deficiency between the ages of 20 and 79 were randomized to receive either Drug A or S. The randomization was stratified on 5 year age groups to ensure that at the end of the study in each of the 12 age ranges $20-24$, $25-29$, $30-34, \ldots, 75-79$ there were 200 persons, of whom exactly 100 got drug A and 100 got drug S. Enzyme levels were measured both at baseline and 6 weeks later on each individual. The main endpoint of the trial (Y) was the change in enzyme levels ($y = 6$ week measurement- baseline measurement). The company hopes that drug A will increase the expected value of Y **by at least 1.5σ** where σ is the standard deviation of Y with drug S; anything less is not considered an important clinical benefit by the clinical community and the Food and Drug Administration.

First the company compares the mean values of Y with Drug A and Drug S in each of the 5-year age groups by performing 12 2-sample t-tests, using two sided tests with significance level $\alpha = 0.10$.

(a) Suppose Y is affected by neither age nor drug type. Under that assumption, what is the probability the company will find **at least one of the 12 age groups** with a statistically significant difference between the drugs?

(b) Continue to assume that Y is affected neither by age nor by drug. What is the **expected number of the 12 age groups** where the sample mean of Y for drug A will exceed the sample mean of Y for drug S **by 0.1σ**?

(c) Now assume the expected value of Y for the ith age group is μ_i, and that the effect of Drug A is to increase μ_i by 0.1σ, but Drug A does not change the standard deviation (σ) of Y. What is the probability the company will find **at least two of the 12 age groups** with a statistically significant difference by performing 2-sample t tests, using 2-sided tests with significance level $\alpha = 0.10$?

(d) The company performed the first analysis described above. They reported to the Food and Drug Administration that they found a significant drug effect in two of the age groups and believe their drug should be approved for use in these age groups. You are the statistical consultant for the FDA and are asked to write a paragraph describing your view of the matter. Specifically, say whether you agree with the drug company report ('yes', 'no' or 'gray zone') and whether there are any cautionary warnings, briefly explaining your reasoning. If you answer 'no' or 'gray zone,' also say whether there any other analyses you would recommend performing. (You need to provide enough information so your analyses can be carried out by the company statistician.)

(e) The company conducted a second analysis and submitted the following summary report to the FDA:

"We now compared all 1200 persons randomized to Drug A to all 1200 persons randomized to Drug S by performing a single 2-sample t-test, two sided with significance level $\alpha = 0.10$. The test statistic was $T = 3.89$ with 2398 degrees of freedom. The result was highly statistically significant, with $p = .0002$. The sample standard deviations of Y were 2.60 and 2.62 with Drugs A and S respectively, and we found no statistically significant difference between these standard deviations using an F test ($p > 0.7$). Based on all of these analyses we conclude that the Drug A is extremely effective in increasing enzyme levels compared to Drug S and we request that Drug A be labeled as clinically superior to Drug S."

As in part (d) explain whether you agree with the drug company report ('yes', 'no' or 'gray zone') and why and whether there are any cautionary warnings. If you answer 'no' or 'gray zone,' say whether there are any other analyses you would recommend, providing enough information so they can be understood and carried out by the company statistician.

Question 2

A study collected data from 194 women who had been diagnosed with breast cancer. Variables of interest included the CES-D score, a measure of depressive symptoms ranging from 0 to 60 points, with higher scores indicating worse symptoms, and the Physical Component Scale-12 (PCS-12), a measure of physical health ranging from 0 to 100, with higher scores indicating better health.

(a) A linear regression was conducted with CES-D score as the outcome and PCS-12 as the predictor. The following output was obtained:

```
      Source |       SS       df       MS              Number of obs =     194
-------------+------------------------------           F(  1,   192) =   15.51
       Model |  1488.30903     1  1488.30903           Prob > F      =  0.0001
    Residual |  18418.5982   192  95.9301989           R-squared     =  0.0748
-------------+------------------------------           Adj R-squared =  0.0699
       Total |  19906.9072   193  103.144597           Root MSE      =  9.7944
------------------------------------------------------------------------------
        cesd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       PCS12 |  -.2595472   .0658942    -3.94   0.000    -.3895168   -.1295776
       _cons |   28.23496   3.379126     8.36   0.000     21.56999    34.89994
------------------------------------------------------------------------------
```

Provide an interpretation of the coefficient for PCS12, an interpretation of its 95% confidence interval, and an interpretation of the R^2 value.

(b) CES-D and PCS-12 were standardized by subtracting their means and dividing by their standard deviations. The standardized CES-D score was regressed on the standardized PCS-12 and the following output was obtained:

```
      Source |       SS       df       MS              Number of obs =     195
-------------+------------------------------           F(  1,   193) =   13.99
       Model |  13.1162384     1  13.1162384           Prob > F      =  0.0002
    Residual |   180.88364   193  .937220931           R-squared     =  0.0676
-------------+------------------------------           Adj R-squared =  0.0628
       Total |  193.999878   194  .999999372           Root MSE      =   .9681
------------------------------------------------------------------------------
     cesdStd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    PCS12Std |  -.2600181   .0695056    -3.74   0.000    -.3971063     -.12293
       _cons |  -3.69e-07   .0693272    -0.00   1.000    -.1367366    .1367359
------------------------------------------------------------------------------
```

Provide an interpretation of the coefficient for PCS12Std.

(c) The researchers want to test the hypothesis that mean CES-D scores increase more rapidly with age for women who are not married compared to those who are married, controlling for PCS-12. Describe how you could use linear regression to test this hypothesis. State the model(s) you would fit and explain specifically how you would test the hypothesis.

(d) Suppose the data set is randomly split into two equally-sized subsets of size $n = 97$, and the linear regression of CES-D on PCS-12 is conducted separately on each of the two smaller data sets. What do you expect will be the standard errors of the coefficients for PCS-12 that you will observe in these analyses? Provide a calculation to support your answer.

Question 3

Let the random variable X follow a Geometric distribution with probability mass function

$$P(X = x; p) = p(1 - p)^x,$$

where $x = 1, 2, \ldots$ and $0 < p < 1$. Suppose that we are interested in estimating $\theta = p/(1 + p)$.

(a) Show that $T = (1/2)^X$ is an unbiased estimator of θ based on a single observation.

(b) Is T the only unbiased estimator of θ based on a single observation? Justify your answer.

(c) Suppose you have a random sample $X_1, X_2, \ldots, X_n$ of size n from the above Geometric distribution. Derive the MLE of θ based on the entire sample. Is your estimator UMVUE? Justify your answer.

(d) Using the aforementioned random sample $X_1, X_2, \ldots, X_n$, derive a 95% asymptotic confidence interval for θ. Justify the steps in your derivation.

Question 4

Cards in a deck are labeled $1, 2, \ldots, n$. We shuffle the deck and next draw n cards without replacement. A match occurs on the j^{th} draw if the card drawn at that time is labeled j.

Let S_n be the number of matches in the n draws and let

$$p_n(k) = P(S_n = k)$$

for $k = 0, 1 \ldots, n$.

(a) Show that the probability of no match in n draws is equal to

$$p_n(0) = \sum_{j=0}^{n} \frac{(-1)^j}{j!}$$

(b) Show that the probability of having exactly k matches in n draws is

$$p_n(k) = \frac{1}{k!} p_{n-k}(0)$$

(c) Show that S_n has approximately Poisson distribution with mean 1.

Question 5

Suppose Y_1, Y_2, and Y_3 are independent random variables, each distributed Poisson with means $E(Y_i) = \lambda_i$ for $i = 1, 2, 3$.

(a)You are interested in testing $H_0 : \lambda_1 = \lambda_2 = \lambda_3$ against a general alternative that the means are not all equal. Suppose you observe $Y_1 = y_1, Y_2 = y_2, Y_3 = y_3$ where the y_i's are non-negative integers.

(i) Write the likelihood function, $L(\lambda_1, \lambda_2, \lambda_3; y_1, y_2, y_3)$ for $\lambda_1, \lambda_2, \lambda_3$ given the observed data y_1, y_2, y_3. Show that the maximum likelihood estimators are $\hat{\lambda}_i = y_i$.

(ii) Determine the likelihood ratio statistic, Λ, for testing your null hypothesis. State the 5% critical region for the asymptotic test of this hypothesis.

(b) An alternative method for testing for equal means is to use the fact that the conditional distribution of Y_1, Y_2, Y_3 given $\sum_1^3 Y_j = n$ is multinomial $M(n; \theta_1, \theta_2, \theta_3)$:

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 | \sum_1^3 Y_j = n) = \frac{n!}{y_1! y_2! y_3!} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3}$$

for y_1, y_2, y_3 non-negative integers summing to n.

(i) Prove this and relate the θ_i's to the λ_i's. What does your null hypothesis in part (a) become when stated in terms of the θ_i's?

(ii) Write the likelihood function, $L(\theta_1, \theta_2, \theta_3; y_1, y_2, y_3)$ for the multinomial. Show that the maximum likelihood estimators are $\hat{\theta}_i = y_i/n$.

(iii) Determine the likelihood ratio statistic, Λ, for testing your null hypothesis as stated in part (b)(i). Show that the asymptotic test of H_0 is identical to the test you found in part (a) (ii).

(c) Given data values $y_1 = 30, y_2 = 17, y_3 = 13$, evaluate your test statistic from either (a) (ii) or (b)(iii) and perform your test.

(d) One could also do a Pearson's χ^2 test of the null hypothesis using the multinomial distribution in part (b). Will this be identical to your test in (c)? Explain why or why not.