NAME: _____ **EXAM NUMBER:** _____

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 21st, 2015

Instructions:

This three hour examination consists of five (5) questions numbered 1-5. Answer all 5 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 6 question pages.

Question 1

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ and σ^2 are unknown.

(a) Derive the maximum likelihood estimate $\hat{\sigma}^2$ for σ^2.

(b) Show that $\hat{\sigma}^2$ is a biased estimate of σ^2. Justify your answer.

(c) Derive the asymptotic distribution of $\hat{\sigma}^2$, appropriately normalized.

(d) Derive the exact distribution of $\hat{\sigma}^2$. (Hint: First demonstrate the independence of the sample mean and sample variance. Then use this to construct your derivation of the exact distribution of $\hat{\sigma}^2$.)

Question 2

Hepatitis C virus (HCV) is a virus that can cause severe liver disease. A study is conducted to determine the proportion of elderly persons (over the age of 65) who are truly infected with HCV. Suppose there is a blood test that can help determine whether a person is infected with HCV but it is not completely accurate. In particular, suppose it is known that the sensitivity and the specificity of the blood test for detecting HCV infection among elderly persons are exactly 0.92 and 0.85, respectively. (Recall that the *sensitivity* or *true positive rate* is the proportion of people who test positive among those who are actually infected while the *specificity* or *true negative rate* is the proportion of people who test negative among those who are truly not infected.) The blood test is given to a random sample of 2000 elderly persons and 421 of them test positive for HCV infection.

(a) Write down a general expression for the probability that a randomly selected elderly person in the population tests positive for HCV as a function of the true probability they are infected with HCV and the sensitivity and specificity of the test.

(b) Provide a point estimate and a 95% confidence interval for the probability an elderly person is truly infected with HCV based on the observed data, briefly justifying any assumptions you make.

(c) In a second study the investigators examined young intravenous drug users (less than age 40). Suppose it is known that among intravenous drug users the sensitivity of the blood test is also exactly 0.92, but that the specificity is exactly 0.65. In a random sample of 1000 intravenous drug users, 430 are found positive for HCV infection based on the blood test. Provide a point estimate for the probability that a young intravenous drug user is truly infected with HCV.

(d) The investigators now wish to test the null hypothesis that the probability of being truly HCV infected among elderly persons is the same as that among young intravenous drug users versus the alternative hypothesis that the probabilities are different. Develop a procedure to perform that hypothesis test, carry out the calculations for the observed data and report a p-value. If you require assumptions to justify your procedure, state them.

(e) Write a couple of complete and grammatically correct sentences summarizing your findings. Your description should include estimates of the HCV prevalence rates (that is, the proportions who are truly HCV infected) among the elderly and young intravenous drug users and whether you have found any statistically significant differences between those rates.

Question 3

A scientific study on drug abuse aims to compare the drug modafinil to a common antidepressant and a placebo for reducing methamphetamine (meth) use. Treatment-seeking meth-dependent participants were randomly assigned to modafinil, antidepressant or placebo for 2 months. Each subject was required to report the number of times he or she made use of meth. The observed data are as follows.

Group	Number of subjects	Average frequency of meth use
(A) Modafinil	$n_A = 10$	$\bar{Y}_A = 4$
(B) Antidepressant	$n_B = 10$	$\bar{Y}_B = 5$
(C) Placebo	$n_C = 15$	$\bar{Y}_C = 7$

Assume that an individual subject's methamphetamine use episodes follow a Poisson process and let λ_A, λ_B and λ_C, respectively, be the average monthly rates of methamphetamine use in the three treatment arms.

(a) Write down the general expression for the distribution of the total number of episodes of meth use in the placebo group and use this to provide the form of an exact 95% C.I. for λ_C.

(b) You are interested in testing whether or not the rates of methamphetamine use differ across the three treatment arms, namely

$$H_0 : \lambda_A = \lambda_B = \lambda_C \quad \text{vs. } H_1 : \text{Not all rates are equal}$$

Write down specific details of an appropriate testing strategy, clearly stating the null distribution of the test statistic and the form of the p-value. (You do not need to do the actual calculations.)

(c) You are asked to design a followup study to determine whether a combined treatment (antidepressant + modafinil) is more effective than modafinil alone. Describe how to conduct a power analysis for sample size determination in this new study. Highlight how you would use available data to make appropriate choices about effect size and variability for the calculations.

Question 4

Consider a slowly-developing disease which has two stages, infection and full-blown illness. Suppose that the probability of a healthy person becoming infected in any given year is 0.1, independent of previous years. Moreover, once a person becomes infected they either stay in that state or progress to having the full-blown disease; they can not go back to being disease free. Assume that if a person is infected, the chance of developing the full-blown disease is 0.01 per year including the year in which the person became infected.

(a) Write down a general expression for the probability that a person will **not** become infected with the disease for k consecutive years and give the value of the probability for $k = 3$.

(b) Find the expected number of years it takes for a healthy person to become infected with the disease.

(c) What is the probability that a person who is uninfected in a given year will develop the full-blown disease in the following year?

(d) Find the expected number of years it takes for a person to contract the full-blown disease. Practically speaking, how concerned are you about getting this illness? Explain briefly.

(e) Write down a general expression for the probability that a healthy person will develop the full-blown disease in k years and then calculate that probability for $k = 3$.

Question 5

(a) A standard assumption of the ordinary linear regression model is constant error variance, $Var(\epsilon_i) = \sigma^2$ for all i. Suppose that the constant error variance assumption is violated. Give brief answers (no more than a few sentences) to each of the following questions.

(i) What is the consequence of violation of this assumption in terms of impact on statistical inference?

(ii) Describe how you would check for nonconstant error variance when conducting a data analysis.

(iii) Describe two different remedies that could be used in a data analysis to lessen the effects of nonconstant error variance.

(b) Data are collected from 267 recently diagnosed breast cancer patients. CESD score, a measure of depressive symptoms (high scores mean more severe symptoms), is regressed on the following variables:

Stress: Perceived stress score, centered at its mean of 15

Married: Marital status (coded as 1 for married and 0 for not)

Edu2, Edu3: Highest education level, coded using two indicator variables (Edu2=1 if college graduate, 0 otherwise; Edu3=1 if post-graduate degree, 0 otherwise).

StressByMarried: Interaction between Stress and Married.

The following output is obtained:

```
      Source |       SS       df       MS              Number of obs =     267
-------------+------------------------------           F(  5,   261) =   77.58
       Model |  17007.8638     5  3401.57276           Prob > F      =  0.0000
    Residual |  11443.3347   261  43.8441943           R-squared     =  0.5978
-------------+------------------------------           Adj R-squared =  0.5901
       Total |  28451.1985   266  106.959393           Root MSE      =  6.6215

                 Coef     Std. Err.
Intercept        12.702   0.909
Stress            1.343   0.102
Edu2             -0.170   1.005
Edu3             -1.854   1.048
Married           0.729   0.857
StressByMarried  -0.285   0.126
```

(i) Provide a brief interpretation for the intecept and the coefficients of Edu3 and StressByMarried.

(ii) Test whether the coefficient for the interaction term, StressByMarried, is significantly different from zero. Provide the degrees of freedom for the test statistic.

(iii) Explain how you would test whether the coefficient for Edu2 equals the coefficient for Edu3. (No calculations are required.)

(iv) A model without the education indicator variables gives the following output. Conduct a test to determine whether education is a significant predictor of CESD, when PSS and marital status are in the model.

```
      Source |       SS       df       MS              Number of obs =     267
-------------+------------------------------           F(  3,   263) =  126.94
       Model |  16828.6651     3  5609.55504           Prob > F      =  0.0000
    Residual |  11622.5334   263  44.1921421           R-squared     =  0.5915
-------------+------------------------------           Adj R-squared =  0.5868
       Total |  28451.1985   266  106.959393           Root MSE      =  6.6477
```

	Coef.	Std. Err.
Intercept	12.03187	.690126
Stress	1.334526	.1023292
Married	.7402888	.8545503
StressByMarried	-.2632816	.1261736

(c) Suppose that two observations are collected from a sample of n individuals, one observation before a treatment $(y_{11}, y_{12}, ..., y_{1n})$ and a second observation after the treatment $(y_{21}, y_{22}, ..., y_{2n})$. Each set of observations is normally distributed.

Commonly, a paired t-test would be conducted to determine whether there is a difference in means before and after treatment, that is, to test $H_0 : \mu_1 = \mu_2$. Explain how one could instead use a linear regression model to test this hypothesis, specifying the dependent variable and the covariate(s). Derive the test statistic for the paired t-test and the test statistic based on the linear regression model and show that they are the same. Do you expect to obtain exactly the same p-value from both approaches? Explain.