NAME: _____ **EXAM NUMBER:** _____

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 24th, 2018

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 5 question pages.

Question 1

A study is performed in which a random sample of 5000 women is screened for breast cancer by mammography. Of the 5000 women, 200 test positive for the disease. However, not all women who test positive for breast cancer by mammography actually have it. The presence of disease can only be definitively diagnosed by biopsy. Biopsies were performed on all of those who tested positive by mammography and the presence of breast cancer was confirmed in 170 of the 200 women.

(a) Find an approximate 95% confidence interval for the probability that a woman with a mammography test that is positive for breast cancer actually has the disease.

(b) In this part of the question, you may take as true the following assumption proposed by the study investigators: A woman who has a mammography test that is negative for the presence of breast cancer has a probability equal to zero of having the disease.

(i) Using the study investigators' assumption and all the information given above, estimate the probability a woman in the population has breast cancer.

(ii) Find an approximate 95% confidence interval for the probability that a woman in the population has breast cancer.

(c) A critic of the study disputes the investigators' claim from Part (b) and believes that the more appropriate assumption is the following: A woman who has a mammography test that is negative for the presence of disease has a probability equal to .06 of having the disease. Using the critic's assumption rather than the investigators' assumption, re-answer Part (b)(i).

(d) To address the critic's concerns, the study investigators decide to collect additional data. A sample of the 4800 women who tested negative for breast cancer by mammography will be asked to have a biopsy. What sample size would you recommend to estimate the probability that a women with a negative mammography test actually has the disease with a 95% error bound of $\pm .02$?

Question 2

Allergen-specific IgE has been proposed as a biomarker for the severity of asthma. Researchers are conducting a study to explore the relationship between allergen-specific IgE levels in serum and the frequency of asthma attacks in a sample of adults with asthma. For each individual, they determine the number of asthma attacks in the past month and serum concentration of allergen-specific IgE. A log10 transformation is used for allergen-specific IgE. Some descriptive information regarding the data is provided below:

	Mean	SD
Number of attacks (X)	1.108	1.125
log10(IgE concentration) (Y)	1.953	1.090

Correlation between X and Y: 0.366

Number of attacks	0	1	2	3
Mean log10(IgE)	1.54	1.83	2.59	2.42
n	15	9	7	6

(a) Suppose that a simple linear model is fit by regressing $Y = \log10(\text{IgE})$ on number of attacks, X. Calculate the estimated regression coefficients (i.e., the intercept and slope).

(b) Provide a quantitative interpretation of the slope coefficient that you obtained in Part (a).

(c) The analysis of variance table for the linear regression of log10(IgE) on number of attacks is provided below. Conduct a test to determine whether number of attacks has a significant linear relationship with log10(IgE), using a $\alpha = .05$.

Source	Df	Sum Sq
Model	1	5.73
Residuals	35	37.06

(d) The researchers would like to compare the linear regression model for the data, $Y = \beta_0 + \beta_1 X$, to a model that uses three dummy variables to code for the four levels of number of attacks, in order to determine which is a better fit for the data. The model with 3 dummy variables gives the following analysis of variance table:

Source	Df	Sum Sq
Model	3	6.76
Residuals	33	36.02

Conduct a test to compare the two models using $\alpha = .05$. What do you conclude?

(e) Suppose that the investigators decide to reverse the roles of X and Y and now want to use log10(IgE) to predict the nuber of asthma attacks a person will have in a month. What sort of model should they use and how would you interpret the coefficient of the log10(IgE) variable in this model? Explain briefly in each case.

Question 3

A study was conducted in 338 HIV-infected veterans to compare the current standard treatment to a proposed new medication regimen. The study also considered whether the race of the patient had an influence on outcomes, operationalized here as whether or not the patient developed AIDS symptoms during a three-year follow-up period. The following data emerged from the study:

Development of AIDS Symptoms by Race and Treatment

Race	Treatment	AIDS Symptoms During Follow-up	
		Yes	No
White	New Regimen	14	93
White	Standard Treatment	32	81
Black	New Regimen	11	52
Black	Standard Treatment	12	43

Define the relevant varables as follows:
$X_{1i} = 1$ if individual i received the new regimen and 0 if they received the standard treatment;
$X_{2i} = 1$ if individual i was white and 0 if they were black;
$Y_i = 1$ if individual i exhibited AIDS symptoms during follow-up and 0 if they did not.

(a) Suppose we first analyze only the data from the 220 white patients. Find the estimated coefficients in the logistic regression model $\text{logit}[\text{P}(Y_i = 1)] = \beta_0 + \beta_1 X_{1i}$.

(b) For the model you fit in Part (a), is there evidence of a significant difference between the two treatments in the emergence of AIDS symptoms during follow-up, at the $\alpha = .05$ level? Summarize your conclusions in a sentence, and explain your reasoning.

(c) Consider the following estimated parameter vector and variance-covariance matrix from fitting a multiple logistic regression model, $\text{logit}[\text{P}(Y_i = 1)] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$, to the full sample:

```
 Predictor  Estimate         Variance-Covariance Matrix
 Intercept   -1.0736          0.06914 -0.03305 -0.05651
      X1     -0.7195         -0.03305  0.07783  0.00333
      X2      0.0555         -0.05651  0.00333  0.08330

Log-likelihood:  -167.576
```

Produce a 95% confidence interval for the coefficient of X_1 and comment on whether it is statistically significantly different from 0. How does this compare to your answer in part (b)?

(d) Consider a model that adds the product term, $X_1 X_2$, as a predictor to investigate whether there is an interaction effect between treatment regimen and race. Suppose the estimated log-likelihood is -166.884 for the model that includes X_1, X_2 and $X_1 X_2$. Carry out a likelihood-ratio test to assess whether there is a significant interaction effect and explain the practical implications of your findings. How does this fit with your answers in (b) and (c)?

Question 4

Suppose that $X_1, \ldots, X_n$ are iid continuous random variables with cdf $F(x)$.

(a) What is the joint distribution of $F(X_1), \ldots, F(X_n)$?

(b) Find $P(F(X_{(1)}) \leq 0.5)$, where $X_{(1)} = \min_{i=1,\ldots,n} X_i$.

(c) Find the joint distribution of $F(X_{(1)})$ and $F(X_{(n)})$, where $X_{(1)} = \min_{i=1,\ldots,n} X_i$ and $X_{(n)} = \max_{i=1,\ldots,n} X_i$.

Question 5

A random quantity, X, is said to follow a Geometric distribution if its probability mass function takes the form: $f_X(x \mid \theta) = \theta(1-\theta)^x$; x=0, 1, 2, ...; $0 < \theta < 1$; with $\mathrm{E}(X) = \theta^{-1}(1-\theta)$ and $\mathrm{Var}(X) = \theta^{-2}(1-\theta)$.

(a) Let $X_1, X_2, \ldots, X_n$ be iid Geometric(θ). Characterize the appropriately normalized limiting distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) Consider the following estimator

$$\tilde{\theta}_n = \frac{n + (1 + \bar{X}_n)}{(n+1)(1+\bar{X}_n)}.$$

Establish whether $\tilde{\theta}_n$ is a consistent estimator of θ and characterize its limiting distribution (appropriately normalized).

(c) Derive the MLE of θ and characterize its limiting distribution.

(d) In the Bayesian setting, consider the following prior $\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$, so that $\theta \sim \mathrm{Beta}(a, b)$; where $a, b > 0$ are contants and $B(\cdot, \cdot)$ is a function of a, b.

(i) Derive the posterior distribution $f(\theta \mid x_1, \ldots, x_n)$.

(ii) Find the posterior expected value $\theta_n^* = E(\theta \mid x_1, \ldots x_n)$.