NAME: ______________________________ **EXAM NUMBER:** __________

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 29th, 2014

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 5 pages.

Question 1

Consider the following data set, aiming to relate a person's genetic predisposition towards skin cancer (melanoma), as measured by hair color, and the probability of actually getting the disease.

	Hair Color			
Melanoma	Red	Blond	Dark	Total
Yes	24	20	10	54
No	6	10	20	36
Total	30	30	30	90

(a) Show how to test whether the occurrence of melanoma is independent of a person's hair color. Specifically, state the null and alternative hypotheses; give the form of the test statistic and provide its null distribution; and define the p-value in terms of a probability involving the test statistic. (There is no need to calculate the actual test statistic or p-value for these data.)

(b) Give the name of the sampling distribution of the number of blonds with melanoma, assuming we observe a total of $n = 90$ individuals, and provide an estimate for its parameters.

(c) A famous oncologist claims that the joint probability distribution of melanoma and hair color is:

	Hair Color			
Melanoma	Red	Blond	Dark	Total
Yes	0.2	0.15	0.1	0.45
No	0.15	0.2	0.2	0.55
Total	0.35	0.35	0.3	1.0

Perform a test to determine whether or not your sample is consistent with the oncologist's opinion. Calculate the value of the test statistic, give the rejection region for the test and state your real-world conclusions. (Note: Unlike in part (a) you do need to do the actual data calculations!)

(d) This is clearly an observational study. If we are interested in the relationship between hair color and skin cancer risk, should we account for other variables? If yes, explain why, using a specific illustrative example. If no, briefly justify your answer. (Your response should not be more than a couple of sentences.)

(e) Does the data structure suggest that the subjects were collected as a simple random sample? If yes, what supports your intuition? If no, what would be the reason for the chosen design? (Your response should not be more than a couple of sentences.)

(f) Consider collapsing the blond hair and red hair categories into a light hair category. Compute the odds ratio for melanoma comparing the light and dark hair-color groups and provide an approximate 95% C.I. for this quantity.

Question 2

(a) In the context of multiple linear regression

(i) Define multicollinearity in words.

(ii) State one way to quantify the severity of multicollinearity.

(iii) Describe two important ways in which severe multicollinearity could affect the inferences from the model.

(iv) Describe two remedies that can be applied in a data analysis to lessen the effects of severe multicollinearity.

(b) When a simple linear regression model, $E(y) = \beta_0 + \beta_1 x$, is fit to data from 50 subjects, the estimated slope coefficient, $\hat{\beta}_1$, has a standard error of 8. Suppose the investigators enlarge the sample to 100 subjects and rerun the model. What is your best estimate for the new standard error of $\hat{\beta}_1$? Provide your calculation.

(c) Researchers collected measurements of C-reactive protein (CRP) and body mass index (BMI) from a sample of colorectal cancer patients. The sample includes men and women. Some of the patients were undergoing chemotherapy and some were not.

(i) When CRP (dependent variable) is regressed on BMI (independent variable), the regression coefficient for BMI is positive and statistically significant. However, when CRP is regressed on BMI and a binary indicator for sex (male=1, female=0), the regression coefficient for BMI is negative and statistically significant. Provide an explanation of how this could happen, including a graphical sketch (e.g., a scatterplot). Assume that the findings are not attributable to erroneous data or outliers.

(ii) The researchers hypothesize that patients with higher BMI will have higher levels of CRP, but in patients undergoing chemotherapy, CRP will increase more rapidly with BMI. Using CRP as the dependent variable, write out a model that you could fit to the data and the corresponding test or tests you would use to evaluate this researchers' theory.

Question 3

In a disease surveillance program suppose $N = nk$ randomly chosen people are available for testing. The test is 100% accurate, but expensive, so pooling is used: the N people are divided into n groups each consisting of k people. The blood from each group of k people is pooled into one sample and that is tested. If all k people are disease free, the pooled test will be negative; if any of the k people have the disease, the pooled test will be positive. Let p be the probability that a randomly chosen person has the disease.

(a) What is the probability that a particular pooled sample from a set of k people tests positive?

(b) Let X be the number of groups (pooled samples) in the surveillance program that test positive. What is the distribution of X? Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

For the remaining parts of the problem suppose we treat p as a random variable and assume $p \sim U(0, 1)$.

(c) Suppose that $X = 1$ (i.e. that exactly one of the n pooled samples has tested positive.) Find the conditional distribution of $p|X$ and the corresponding expected value, $\mathrm{E}(p|X)$. Explain what this tells you in real-world terms about the disease.

(d) Write down general expressions for the density, $p|X$ and expected value, $\mathrm{E}(p|X)$, simplifying them as much as you can. (You may find it helpful to note that $(1 - a^m)^l = \sum_{i=0}^{l}(-1)^i a^{mi}$.)

Question 4

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \mu^2)$, $\mu > 0$. The parameter μ is unknown.

(a) What is the Cramer-Rao lower bound for unbiased estimation of the parameter μ?

(b) Find $\hat{\mu}$, the maximum likelihood estimator (MLE) of the parameter μ. (Note: You should keep in mind that μ is positive in this exercise. Your estimate should also have this property.)

(c) What is the approximate distribution of $\sqrt{n}(\hat{\mu} - \mu)$? Briefly justify your answer.

(d) Suggest a variance stabilizing transformation that could be used to set approximate confidence intervals for μ and show what the form of the resulting intervals would be.

Note: You do not need to complete part (b) in order to answer parts (c) and (d).

Question 5

Let Z and V denote independent random variables, where $Z \sim N(0,1)$ and $V \sim \chi^2_m$.

(a) Write down the joint density of Z and V.

(b) Obtain the density of $T = Z/\sqrt{V/m}$.

(c) Let $X_1, \dots, X_n$ be a random sample drawn from a $N(\mu, \sigma^2)$ distribution. Derive the asymptotic distribution of

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}}$$

when $n \to \infty$.

Note: You may find it useful to remember that the densities for the χ^2_m and t_ν distributions are, respectively,

$$f(x) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} x^{(\frac{m}{2}-1)} e^{-\frac{x}{2}}, x > 0, m > 0$$

and

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} (1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}, \nu > 0.$$