NAME: ______________________________ **EXAM NUMBER:** __________

PRELIMINARY COMPREHENSIVE EXAMINATION

Department of Biostatistics

8:30 AM - 11:30 AM

September 23rd, 2013

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 pages.

Question 1

Suppose that in keeping records on the costs associated with a sequence of health-care procedures, a health-policy analyst maintains a database with one transaction per line, rounding each entry in the database to the nearest dollar. To be more specific, when the decimal portion of the cost (i.e., the number of cents) is from 0.00 to 0.49, then the entry would be rounded down to the nearest dollar, and when the decimal portion is from 0.50 to 0.99, the entry would be rounded up.

(a) To simplify matters, assume that rounding errors have a continuous uniform distribution on $(-0.5, 0.5)$ and are independent of one another. What are the mean and variance of this distribution?

(b) Continuing with the distributional assumption from part (a), what is the probability that at least 9 out of the next 10 transactions will have a rounding error of a quarter of a dollar or more in magnitude (i.e. .25+ either up or down)?

(c) Still continuing with the distributional assumption from part (a), what is the probability that the accumulated rounding error across the next 100 transactions will exceed \$5.00 (i.e. overstate the total cost by that amount)?

(d) A well-known comment attributed to George Box, one of the leading statisticians of the 20th century, is that "All models are wrong; some models are useful." In this context, can you think of anything "wrong" with using a continuous uniform distribution on $(-0.5, 0.5)$ as a way of representing the analyst's rounding rule? If so, explain what you see as 'wrong' and comment on whether the continuous uniform distribution might still be regarded as useful in this context. If not, explain why you think the continuous uniform distribution provides a representation of the rounding rule that cannot be improved in a meaningful way.

Question 2

It's a nice, warm, summer day, and you, a biostatistician, are minding your own business when suddenly a non-statistical investigator bursts into your office. 'Help', she says, 'I'm writing a grant and have a few 'stats' questions.' Answer each of her questions in a couple of sentences. (You may use an example if it helps explain a concept.)

(a) With regard to variable selection, what is the difference between confounding and effect modification (i.e. an interaction effect)?

(b) What is the difference between a fixed and a random effect?

(c) What is the difference between standard deviation and standard error?

(d) With regard to hypothesis testing, what is the difference between a Type I error and a Type II error?

(e) With regard to hypothesis testing, what is the difference between the p-value and the Type I error?

Question 3

In a randomized controlled trial, breast cancer patients were assigned to an intervention group (n = 39) or a control group (n = 34). CESD score, a measure of depressive symptoms, was obtained at baseline (prior to randomization) and at 6 months (at the end of treatment). The CESD score takes on values from 0 to 60, with a higher score indicating worse symptoms. The table below provides statistics on the CESD scores from the trial.

Table 1. Summary statistics for CESD scores

Variable	Control	Intervention	Overall
	Mean (SD)	Mean (SD)	Mean (SD)
	n = 34	n = 39	n = 73
Baseline CESD	11.1 (9.5)	11.3 (9.2)	11.2 (9.3)
6-month follow-up CESD	13.1 (11.5)	9.5 (6.5)	11.1 (9.3)
Change in CESD (follow-up minus baseline)	2.0 (9.8)	-1.8 (8.3)	-0.05 (9.17)

(a) Test for a mean difference between the intervention and control groups on change in CESD score from baseline to follow-up. Show your work.

For Parts (b)-(d): A linear regression model is fit with 6-month follow-up CESD as the outcome variable and baseline CESD, an indicator for treatment group (1 for intervention and 0 for control), and the interaction between baseline CESD and the indicator as covariates. The following output is obtained.

Table 2. Linear regression model estimates

Variable	Parameter estimate	Standard error
Intercept	5.33	2.06
Baseline CESD	0.70	0.14
Group indicator	0.28	2.86
Baseline CESD x group indicator interaction	-0.36	0.20
R^2=0.332		
Root mean square error (RMSE) = 7.750		

Table 3. Covariance matrix of parameter estimates

	Intercept	Baseline CESD	Group indicator	Interaction
Intercept	4.25	-0.22	-4.25	0.22
Baseline CESD	-0.22	0.02	0.22	-0.02
Group indicator	-4.25	0.22	8.18	-0.44
Interaction	0.22	-0.02	-0.44	0.04

(b) Provide a real-world interpretation of the parameter for the interaction term.

(c) Based on the output in Table 2, what is the expected difference in mean CESD at follow-up for individuals **in the intervention group** who differ in baseline CESD by one unit? Give a point estimate and construct a 95% confidence interval for this quantity.

(d) Suppose that the parameter estimates in the previous model are the true parameter values. Assuming that the error terms are normally distributed, find the probability that an individual in the control group who had a CESD score of 10 at baseline would have a CESD score of 11 or higher at 6-month follow-up.

Question 4

Suppose that each of a random number, N, of insects lays X_i eggs, where the X_i's are independent, identically distributed random variables with values greater than 0. The total number of eggs laid is $H = X_1 + \ldots + X_N$. Assume the hierarchical model

$$\begin{aligned} &N \sim \text{Poisson}(\lambda) \\ &H|N = X_1 + \ldots + X_N \\ &P(X_i = x) = \frac{-(1-p)^x}{xlog(p)}, \quad x = 1, 2, \ldots, \quad i = 1, \ldots, N \quad 0 < p < 1. \end{aligned}$$

Show all your steps for each of the following:

(a) Derive the moment generating function of X_i.

(b) Find $E(H)$.

(c) Find $Var(H)$.
(Hint: If $Y \sim$ Geometric(p) with $P(Y = y) = p(1-p)^{y-1}$, $y = 1, 2, \ldots$, $0 \leq p \leq 1$, then $E(Y) = 1/p$.)

(d) Use the moment generating function of X_i to derive the moment generating function of H.

Question 5

Two possible study designs are being considered to check whether a certain genetic abnormality increases the risk of Alzheimer's disease. The study investigators plan to test the null hypothesis that there is no association between the genetic abnormality and Alzheimer's disease and to calculate the odds ratio comparing the risk of Alzheimer's disease with and without the genetic abnormality. As the consulting statistician, you are asked to provide advice about the study designs. Note that Alzheimer's is a disease generally associated with old age, with risk beginning around age 65.

(a) The first proposed study is a case-control design. The sample will consist of 100 randomly selected 75-year olds with Alzheimer's disease and 130 75-year old controls who do not have the disease. All persons will be given a test to determine whether or not they carry the genetic abnormality. It is known that approximately 20% of controls have the genetic abnormality. Find the power of this case-control study to detect an odds ratio of 3.0 using a 2-sided test with a Type 1 error rate of $\alpha = .05$ assuming all the information the investigators have provided is correct.

(b) The second proposed design is a prospective cohort study. A random sample of N 65-year olds without Alzheimer's disease will be recruited and followed for 10 years to determine whether they develop the illness. (You may assume that all subjects are followed until the end of the 10 years-i.e. there are no drop outs or deaths.) Each of the N subjects will also be given the genetic test to determine whether they carry the abnormality.

The investigators project that about 4% of the N persons will develop Alzheimer's over the 10 year follow-up. They would like to design the study to detect an odds ratio of 3 with 90% power using a two-sided test with $\alpha = .05$. Assuming all the information that the investigators have provided is correct, what is your recommended sample size, N, for this study? You should feel free to use any information given in both parts (a) and (b) to answer this question.

(c) In your role as consulting statistician, write a short paragraph (a few sentences) for the investigators that briefly discusses some of the strengths and weaknesses of the two study designs. The goal is to provide a summary that would help to the investigators decide which design to use.