

University of Illinois at Chicago
School of Public Health
Epidemiology-Biostatistics Division

PRELIMINARY EXAMINATION
Ph.D. in Biostatistics
Part II

Friday, September 10, 2010, 1 pm

This is a take-home exam. You are not to communicate about any aspect of this exam with anyone except Drs. Bhaumik (dbhaumik@psych.uic.edu), Chen (hychen@uic.edu), Demirtas (demirtas@uic.edu), Freels (sallyf@uic.edu), Hedeker (hedeker@uic.edu) and Xie (huixie@uic.edu). There are seven questions in this exam. You need to answer only three of them. Only the chosen three will be counted towards your score. Please start each question on a separate page. Before you turn in your answers, number your pages consecutively in the upper right-hand corner, and put your code number (not your name) next to the page number. Turn in your completed exam to Dr. Bhaumik (Room 457) on September 17, between 10 am and 1 pm.

Code Number:

1. The data for this question are from the Schizophrenia data that is discussed in the Longitudinal Data Analysis class. This study examined the effectiveness of drug, relative to placebo, in reducing the severity of schizophrenia symptoms. In class, using a variety of methods, we noted that there was a significant drug by time interaction. Here, examine the degree to which this posited relationship is moderated by a subject's gender. The dataset is located at: <http://tigger.uic.edu/~hedeker/SCHIZREP.DAT.txt> and contains the following variables:

field 1: Patient ID

field 2: IMPS79 (7-point severity scale)

field 3: Week - from 0 to 6

field 4: Drug (0=placebo, 1=drug)

field 5: sex (0 = female 1 = male)

Perform a reasonable longitudinal data analysis that examines whether there is a significant drug by time interaction. Does gender moderate this interaction? If so, describe how the effect of drug might vary for males and females.

2. This exercise is about generating multivariate binomial data with specified marginal characteristics and association structure. This type of data frequently arise in toxicology studies. Let X have a binomial distribution with parameters n and p .
 - (a) Show that Pearson's coefficients of skewness and excess kurtosis are $\frac{1-2p}{\sqrt{np(1-p)}}$ and $\frac{1-6p+6p^2}{np(1-p)}$, respectively.
 - (b) Using the Fleishman's power polynomials approach, generate univariate binomial data of size 100 and 10000 with specified n and p . This RNG technique is designed for continuous data, so round the resulting numbers to the nearest integer. Make sure no range violations occur, i.e., if a realization turns out to be negative, make it 0. If it exceeds n , force it to be equal to n . Calculate the first four moments (mean, variance, skewness and kurtosis). Repeat this process 1000 times, and report the average empirical estimates along with the theoretical results. You need to use the R functions we have used in Computational Statistics course. Note that at least three (n, p) combinations should be employed. You are free to use identical marginals.
 - (c) Extend what you did in part b to the bivariate case with a specified Pearson correlation. Pick three levels of correlation and report the average empirical correlation across 1000 simulation replicates for $3*3 = 9$ scenarios (There are three correlation quantities and three (n, p) combinations).

3. If S^2 is the sample variance based on a sample of size n from a normal population, we know that $(n - 1)S^2/\sigma^2$ has a χ_{n-1}^2 distribution. The conjugate prior for σ^2 is the inverted gamma pdf, $IG(\alpha, \beta)$, given by $f(\sigma^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \sigma^{-2(\alpha+1)} e^{-\frac{1}{\beta\sigma^2}}$, $0 < \sigma^2 < \infty$, where α and β are positive constants. Find the posterior distribution of σ^2 . Find the mean of this distribution (the Bayes estimator of σ^2).
4. Let Y denote the height of a person in centimeter at age 20. We know that the height of a person is affected by both genetic and environment factors. Let X denote the environmental factor measured. Denote the genotype of a person by G , which takes three values: 0, 1, and 2. The distribution of the genotype in the general population follows the Hardy-Weinberg law of equilibrium, that is,

$$P(G = 0) = p^2, P(G = 1) = 2p(1 - p), p(G = 2) = (1 - p)^2.$$

Since genotyping a person can be expensive, only a fraction of the subjects under study are genotyped. More specifically, let (Y_i, X_i) , $i = 1, \dots, n$, be a random sample from the general population. Subjects who are very tall or very short are more likely to be selected from the random sample for genotyping. Assume the selection probability is

$$P(S = 1|Y, X, G) = \frac{\exp(-3.0 + |Y - 175|/5)}{1 + \exp(-3.0 + |Y - 175|/5)}.$$

For those selected, G is measured. For others, G is missing. Assume that m of the n subjects were selected for genotyping. Assume further that genotype G and environment factor X are independent in the general population. That is, $p(G, X) = p(G)p(X)$. Under the model that

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_{12} G_i X_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and that different ϵ s are independent, answer the following questions based on data obtained from this study design.

- Find out the likelihood for the observed data.
- Derive formulas to be used in the E-step and the M-step of the EM algorithm.
- Give details of your plan for estimating the variance of β_{12} estimator.
- If we are not sure about the independence assumption on gene and environment factors, is there any way you can make the estimator of β_{12} robust against this assumption?

5. Batchelor and Hackett (1970) have reported the results of a study of 16 acutely burned patient received from one to four grafts. For each graft, the time in days to rejection of the graft (T) was recorded as well as an indicator variable Z which had a value of 1 if the graft was a good match of HLA skin type and 0 if it was a poor match. The survival times of some grafts were censored by the death of the patient (T +). The data is recorded below.

Patient	(T,Z)
1	(29, 0), (37, 1)
2	(3, 0), (19, 1)
3	(15, 0), (57+, 1), (57+, 1)
4	(26, 0), (93, 1)
5	(11, 0), (16, 1)
6	(15, 0), (21, 1)
7	(20, 1), (26, 0)
8	(18, 1), (19, 0)
9	(29, 0), (43, 0), (63, 1), (77, 1)
10	(15, 0), (18, 0), (29, 1)
11	(38, 0), (60+, 1)
12	(19, 0)
13	(24, 1)
14	(18, 0), (18, 0)
15	(19, 0), (19, 0)
16	(28+, 0), (28+, 0)

The survival time of an allograft is thought to depend on the degree of HLA matching between the patient and the donor and on the strength of the patient's immune response.

- Examine whether or not there appears to be a random patient effect due to differing immune response.
 - Use the Cox proportional hazards model to test for the effect of HLA matching. First ignore the clustering due to patients and analyze the data as 34 independent observations; second, account for patients in the model. Compare your results.
 - Explore the use of parametric models for this data. Choose the best parametric model and compare results to b).
6. Let X_1, X_2, \dots, X_{2n} be iid $N(0, 1)$ rv's. Define

$$U_n = \frac{X_1}{X_2} + \frac{X_3}{X_4} + \dots + \frac{X_{2n-1}}{X_{2n}}, V_n = X_1^2 + X_2^2 + \dots + X_n^2, \text{ and } Z_n = \frac{U_n}{V_n}$$

Find the limiting distribution of Z_n .

7. Let $X_i, i = 1, 2, \dots, n$ be *i.i.d* from a one-parameter exponential family with distribution

$$p(x; \theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\}.$$

In this setting, $\sum_{i=1}^n T(X_i)$ is sufficient for θ . Explain what this means.

Assuming Regularity conditions which allow differentiation under the integral sign, show that $n^{-1} \sum_{i=1}^n T(X_i)$ is the maximum likelihood estimator of $ET(X_i) = -d'(\theta)/c'(\theta)$, where \prime denotes differentiation with respect to θ . Do this mean that a solution $\hat{\theta}$ of

$$n^{-1} \sum_{i=1}^n T(X_i) + d'(\theta)/c'(\theta) = 0$$

is a maximum likelihood estimator of θ ?

Now consider the estimation of θ by least squares. Show that $n^{-1} \sum_{i=1}^n T(X_i)$ is also the least squares estimator of $-d'(\theta)/c'(\theta)$.