

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

PRELIMINARY EXAMINATION
Ph.D. in Biostatistics
Part I

Friday, September 10, 2010
9 am to 1 pm

Please read the following instructions carefully before answering the questions.

- This is a closed book exam. No books or notes are permitted. You may use a calculator.
- There is one question from each PhD course offered. You are required to answer questions from Advanced Statistical Inferences (question 1) and Linear Models (question 2). Please choose to answer three from the remaining six questions. Please note that only the chosen three in addition to the two required will be counted towards your score even if you choose to answer more than three questions.
- Write your answers on the paper provided. Start each question on a separate page. Number your pages consecutively in the upper right-hand corner, and put your code number (not your name) next to the page number.

Code Number:

Advanced Statistical Inferences

1. Let X_1, X_2, \dots, X_n be iid satisfying all the eight regularity conditions discussed in the class. Show that any consistent sequence $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N\left(0, \frac{1}{I(\theta_0)}\right)$$

where θ_0 is the true value of θ and $0 < I(\theta_0) < \infty$.

Linear Models

2. Consider the following linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

where

\mathbf{y} : is the observable vector of dimension $n \times 1$

\mathbf{X} : is the known design matrix of dimension $n \times p$

β : is the vector of parameters of dimension $p \times 1$

\mathbf{e} : is the error vector of dimension $n \times 1$

Assume that $\mathbf{e} \sim N_n(\mathbf{0}, \Sigma)$, $r(X) = p$, and $\Sigma = \sigma^2 I$, where σ^2 is unknown. Derive the maximum likelihood estimates (mle) of β and σ^2 .

Generalized Linear Models

3. Suppose that one has n pairs of measurements $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The $2n$ values being normally distributed and independently with variance σ^2 . Every pair (x_i, y_i) has an unknown mean $E(x_i) = E(y_i) = \mu_i$ for $1 \leq i \leq n$. Find the maximum likelihood estimator of σ^2 , and show that the estimator found is not consistent as n becomes large. Can you find a consistent estimator?

Computational Statistics

4. The Type I, II, III, and IV generalized logistic density functions are

$$f^I(x; a) = \frac{ae^{-x}}{(1+e^{-x})^{a+1}}, \quad -\infty < x < \infty, \quad a > 0.$$

$$f^{II}(x; a) = \frac{ae^{-ax}}{(1+e^{-x})^{a+1}}, \quad -\infty < x < \infty, \quad a > 0.$$

$$f^{III}(x; a) = \frac{\Gamma(2a)}{(\Gamma(a))^2} \frac{e^{-ax}}{(1+e^{-x})^{2a}}, \quad -\infty < x < \infty, \quad a > 0.$$

$$f^{III}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{e^{-bx}}{(1+e^{-x})^{a+b}}, \quad -\infty < x < \infty, \quad \min(a, b) > 0.$$

i) Describe a random generation method that can efficiently simulate random variates from these densities.

ii) How do we generate multivariate data whose univariate components marginally follow one of the above densities? Assume that Pearson correlation matrix is either available or specified.

Bayesian Inferences

5. Let X_1, \dots, X_n be a random sample from a $N(\theta, \sigma^2)$ population, and suppose that the prior distribution on θ is $N(\mu, \tau^2)$. Here we assume that σ^2 , μ , and τ^2 are all known.

(a) Find the joint pdf of \bar{X} and θ .

(b) Derive $f(\bar{x}|\sigma^2, \mu, \tau^2)$, the marginal distribution of \bar{X} .

(c) Derive $f(\theta|\bar{x}, \sigma^2, \mu, \tau^2)$, the posterior distribution of θ .

Missing Data Analysis

6. Let Y_1 and Y_2 be two binary variables. A random sample is drawn and data on (Y_1, Y_2) are collected. Variable Y_1 is fully observed and Y_2 is subject to missing values. The completely observed cases are summarized in the following table

	$Y_1 = 0$	$Y_1 = 1$
$Y_2 = 0$	97(n_{00})	28(n_{01})
$Y_2 = 1$	53(n_{10})	52(n_{11})

and the data on Y_1 with Y_2 missing are summarized in the following table

$Y_1 = 0$	$Y_1 = 1$
30(N_{+0})	20(N_{+1})

Let $p_{jk} = P(Y_1 = j, Y_2 = k)$, $j = 0, 1$ and $k = 0, 1$. Assume the missing data are missing at random.

- Write out the observed data likelihood.
- Find the maximum likelihood estimator of $p_{00}, p_{01}, p_{10}, p_{11}$.

Survival Analysis

7. Consider the problem of calculating a nonparametric estimate of the bivariate distribution of X =age at first marriage and Y =age at first birth. The purpose would be to examine the nature of the association between these two variables without imposing any parametric model. A random sample of the population is drawn; each individual is asked whether or not they have ever been married, or ever had a child, and if so they are asked how old they were when the event occurred. Each person's current age is also noted.
- Set up notation for your data. Clearly you need to allow for either or both events to be right-censored at the person's current age. Is this an example of homogeneous or heterogeneous censoring? Explain.
 - Describe the standard nonparametric estimates for the marginal distributions of age at first marriage and age at first birth.
 - Suppose you have a large sample, $N=5,000$. Set up notation for your data and describe in detail how you will estimate the bivariate distribution.
 - Now suppose you have a small sample, $N=50$. What problems might arise for the procedure you describe in 3)? How will you revise your procedure to handle this?
 - What assumptions are being made? Specifically address the fact that you are using a cross-sectional sample.

Longitudinal Data Analysis

8. Assume the following linear mixed model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + v_i + e_{ij}$$

where y_{ij} is the dependent variable for subject i ($i = 1, \dots, N$) at time j ($j = 1, \dots, n_i$), and t_{ij} is a time variable taking on values $0, 1, \dots, n_i - 1$. The random subject effects v_i are normally distributed in the population with mean 0 and variance σ_v^2 , and the errors e_{ij} are also normally distributed with mean 0 and variance σ_e^2 (and the random effects and errors are uncorrelated).

- (a) Based on this model, what is the expectation and variance of y_{ij} ?
- (b) Write down the likelihood function and then outline the maximum likelihood solution for parameters β_0 and β_1 .
- (c) Now, suppose that the model is

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + v_{0i} + v_{1i} t_{ij} + e_{ij}$$

where, the random subject effects are distributed normally as:

$$\begin{bmatrix} v_{0i} \\ v_{1i} \end{bmatrix} \sim \mathcal{N} \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{v_0}^2 & \sigma_{v_0 v_1} \\ \sigma_{v_0 v_1} & \sigma_{v_1}^2 \end{bmatrix} \right\}$$

How do your answers to (a) and (b) change?