Division of Biostatistics
College of Public Health
Qualifying Exam II
Part I

1-5 pm, August 24, 2012
Closed Book

1. Write the question number in the upper left-hand corner and your exam ID code in the right-hand corner of each page you turn in.
2. Do **NOT** put your name on any of your answer sheets.
3. Start each problem on a separate sheet of paper.
4. There are 4 questions, each worth 25 points for a total of 100 points. Answer each question as completely as you can being sure to show your work and justify your answers.

1. The Television, School and Family Smoking Prevention and Cessation Project (TVSFP) was a study designed to determine the efficacy of a school-based smoking prevention curriculum in conjunction with a television-based prevention program, in terms of preventing smoking onset and increasing smoking cessation.

   The study used a 2 X 2 factorial design, with four intervention conditions determined by the cross-classification of a school-based social-resistance curriculum (CC: coded 1 = yes, 0 = no) with a television-based prevention program (TV: coded 1 = yes. 0 = no). Randomization to one of the four intervention conditions was at the school level, while much of the intervention was delivered at the classroom level.

   The subset of the complete study data consists of 1600 seventh-grade students from 135 classes in 28 schools in Los Angeles. The response variable, a tobacco and health knowledge scale (THKS), was administered before and after randomization of schools to one of the four intervention conditions. The scale assessed a student's knowledge of tobacco and health.

   (a) (3 points) Identify possible sources of correlated data from this study description and provide a brief explanation for each possible source of correlation.

   (b) (6 points) Unfortunately, all pretest data have been lost, so investigators will consider models with posttest data only. They will consider models for the ith student in the jth classroom of the kth school of the form:

   $$Y_{ijk} = X_{ijk}\beta + \mathrm{b}_k^{(3)} + \mathrm{b}_{jk}^{(2)} + \epsilon_{ijk}$$

   where $\epsilon_{ijk} \sim N(0, \sigma_1^2)$, $\mathrm{b}_{jk}^{(2)} \sim N(0, \sigma_2^2)$, and $\mathrm{b}_k^{(3)} \sim N(0, \sigma_3^2)$.
   Explain in the context of this study what it would mean if:
   i. σ_2^2 was large relative to σ_1^2 and σ_3^2
   ii. σ_3^2 was large relative to σ_2^2
   iii. σ_1^2 was large relative to σ_2^2 and σ_3^2

   (c) (2 points) Instead of considering a model like the one outlined above, the investigators consider the vector of all observations from the same school (Y_k) for school k. School k has M_k classrooms. Define:

$$v_k = \begin{bmatrix} \mathrm{b}_k^{(3)} + \mathrm{b}_{1k}^{(2)} \\ \mathrm{b}_k^{(3)} + \mathrm{b}_{2k}^{(2)} \\ \vdots \\ \mathrm{b}_k^{(3)} + \mathrm{b}_{M_k k}^{(2)} \end{bmatrix}$$

The linear model can thus be considered as:

$$Y_k = X_k\beta + Z_k v_k + \epsilon_k$$

For this model to make sense, give the dimensions of Y_k and Z_k assuming that there are n_k total observations from school k.

(d) (3 points) Write the actual values for the first four rows of the Z_k matrix corresponding to two observations from classroom 1 and two observations from classroom 2 and assuming that school k has exactly 5 classroms ($M_k = 5$ for this question only).

(e) (3 points) Still assuming as before that $\mathrm{b}_{jk}^{(2)} \sim N(0, \sigma_2^2)$ and $\mathrm{b}_k^{(3)} \sim N(0, \sigma_3^2)$, describe $\Sigma = Cov(v_k)$.

(f) (4 points) Instead of fitting a mixed model, the investigators decide they would like to use a generalized estimating equation (GEE) approach instead. Explain to the investigators how you would specify the model using a GEE approach. Be sure to (1) state which clusters would be identified and (2) discuss how the assumptions about clustering in GEE are different than what was true in the mixed effects model above.

(g) (2 points) Assuming that all assumptions from both the mixed model and the GEE approach are reasonable, would you expect a noticeable difference in the estimated treatment effect between the two approaches (LMM vs. GEE)? Why or why not?

(h) (2 points) Suppose the investigators wanted to consider the binary outcome of scoring below a certain threshold, instead of treating score as a continuous outcome. They consider again a mixed model and a GEE approach, this time using a logit link so that they can estimate the treatment effect as an odds ratio. Again assuming that all assumptions are reasonable, would you expect a noticeable difference in the estimated treatment effect between the two approaches (LMM vs. GEE)? Why or why not?

2. Bruce et al. (1973, *American Heart Journal*, **65**: 546-562) conducted a study which assessed physical conditioning in normal individuals. As part of the study, subjects ran on a treadmill for different durations and different physiological measurements were obtained immediately afterward including heart rate (beats/minute). Let y_i denote the heart rate for subject i and x_i be her duration on the treadmill (in minutes). Consider the following linear model relating the two variables:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.$$

   a. (4 points) Write down the likelihood for this model, assuming n subjects.

   b. (8 points) Assume β_0 and σ^2 are fixed and known, but β_1 is unknown. Derive the maximum likelihood estimate (MLE) of β_1. Don't just provide the solution; show your work.

   c. (9 points) Now assume that the unknown parameter β_1 comes from a $N(m_1, v_1)$ distribution. Derive the distribution of β_1 given the data $(y_1, \ldots, y_n; x_1, \ldots, x_n)$ and all other parameters $(\beta_0, \sigma^2, m_1, v_1)$. In a Bayesian analysis, this is known as the posterior distribution of β_1, though no knowledge of Bayesian inference is required to solve this part.

   d. (4 points) Show that the mean of the posterior distribution of β_1 is approximately equal to the MLE when v_1 is large.

3. Consider a *method comparison study* where two methods or instruments are compared for *agreement*. For example, one may want to know whether an oral thermometer and a tympanic thermometer (put in the ear) produce sufficiently close temperature readings. If they do, one can use them interchangeably, or use the one that is cheaper. If there is close agreement between the two methods, we expect to see a sample scatter-plot of X_2 versus X_1 "close" to the 45^o line.

   One simple statistical model applicable here is the bivariate normal (BVN) distribution. Note that (X_1, X_2) is said to be BVN$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, if its pdf is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},$$
$$-\infty < x_1 < \infty, -\infty < x_2 < \infty.$$

   (a) (3 points) What does close agreement mean in terms of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ? Explain.

   (b) (3 points) Consider a transformation given by $Y_1 = (X_1 + X_2)/2$ and $Y_2 = X_1 - X_2$. Determine the joint distribution of (Y_1, Y_2) and identify all the parameters in terms of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ.

   For the following parts, assume that $(X_{1i}, X_{2i}), i = 1, \ldots, n$, is a random sample from the above pdf, where X_{1i} and X_{2i} are the readings from the ith subject using Method 1 and 2, respectively.

   (c) (12 points) Assume $\rho = 0$ and suppose we want to test $H_0 : \sigma_1 = \sigma_2$ against the alternative $H_1 : \sigma_1 \neq \sigma_2$. Derive the likelihood ratio test (LRT) that has level α and describe its critical region in as explicit form as possible.

   (d) (2 points) If there is close agreement between X_1 and X_2, what kind of scatter plot do we expect for $(Y_{1i}, Y_{2i}), i = 1, \ldots, n$? Discuss. Here $Y_{1i} = (X_{1i} + X_{2i})/2$ and $Y_{2i} = X_{1i} - X_{2i}$.(This scatter plot is called the Bland-Altman plot.)

   (e) (2 points) Consider again $H_0 : \sigma_1 = \sigma_2$ against the alternative $H_1 : \sigma_1 \neq \sigma_2$, but now ρ is unknown. Show that, regardless of ρ, $\sigma_1 = \sigma_2$ if and only if Y_1 and Y_2 are uncorrelated.

   (f) (3 points) Using (e), suggest a level α test procedure for H_0 in terms the Y's,

identifying the test statistic and the associated critical region. There is no need to derive it.

4. Suppose that the collected right-censored data are $\{X_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i), i = 1, 2, \cdots, n\}$. Assume that T_i and C_i are independent. Consider an Exponential model for T_i given covariate Z_i as $f(t|z) = \lambda z \exp(-\lambda z t)$, where $f(\cdot)$ is the density function.

   (a) (10 points) Derive the score equation for λ and the asymptotic variance of the MLE of λ.

   (b) (8 points) Let covariate Z_i be constant and $Z_i = 1, i = 1, 2, \cdots, n$. Denote the MLE of λ by $\hat{\lambda}$. Derive the MLE of the median survival time of T, $\hat{t}_{0.5}$, and its corresponding asymptotic standard error estimator.

   (c) (7 points) Dr. Crank has decided to conduct a simulation study to assess the performance of the estimated median survival time $\hat{t}_{0.5}$ derived in (b). He would like to set true $\lambda = 1, 2$ and the sample size of simulated data set is $n = 400$.
   Please suggest a concrete simulation plan for him, be specific about how to generate the survival data and what are the quantities that you will use to evaluate the properties of $\hat{t}_{0.5}$.

Division of Biostatistics
College of Public Health
Qualifying Exam II
Part II

8 am—5 pm, August 25, 2012

1. There are two data analysis projects for this part. Submit a separate report for each project with your exam ID code on the title page of the reports. Do **NOT** put your name on any page of your reports.

2. Time allocation

8 am—12 pm	Project #1
12 pm—1 am	Lunch Break
1 pm—5 pm	Project #2

3. The datasets are saved as read-only on the qualifier exam drive. You need to first download them to the desktop of your computer before you start working on it. At the end of each time period, print a copy of your final report and also save an electronic copy on the desktop with your exam ID as the file name.

4. Part II is open book and you are allowed to bring up to 10 books and unlimited class notes as references.

5. Each report should be self-contained. Follow the instruction of each question to prepare your answer.

6. Each project is worth 50 points.

7. No access to internet is allowed during the exam.

8. Login information

   For the exam in the lab, all students need to login with the following account:

   Username: bioexam

   Password: bioexam

   The dataset for the test can be found under My Computer in the T: Drive. They will be the only files available. Each student should copy the file to the desktop of their machine before starting the exam.

1. All patients who had undergone either brachytherapy, external-beam radiation, or radical prostatectomy (surgery to remove the prostate) as primary therapy for localized prostate cancer during a 4-year period (June 1, 1995, to May 31, 1999) at the University of Michigan, Departments of Radiation Oncology and Urology-Surgery, were offered participation in an institutional review board-approved cross-sectional survey. A total of 650 men participated (response rate 75%).

   Investigators are interested in health-related quality of life (QOL) post-treatment, as measured by the Expanded Prostate Cancer Index Composite (EPIC) survey. In particular, you will investigate **bowel-related quality of life**. This is measured with a composite score that ranges from 0 to 100, with higher values indicating higher quality of life.

   The primary goal of this study is to investigate differences in long-term bowel-related QOL between the three treatment types (brachytherapy, radiation, surgery).

   The variables available to you are summarized in the table below. In addition to the treatment variable and bowel QOL, there are several other patient characteristics that may be important.

Variable	**Description**	**Comments**
id	Unique subject identifier	
tx	Treatment	Values are {brachy, radiation, surgery} for the three treatment types
qtime	Time between treatment and survey completion	Years
bqol	Bowel-related Quality of Life (QOL) score	Ranges from 0 to 100. Higher is better QOL.
age	Patient age at survey completion	Years
Tumor/cancer characteristics		
gs	Gleason score	Values are from 2 to 10. Higher is worse.
tstage	Tumor stage (T-stage)	T-stages take the values {1,2,3}. Higher is worse.
psa	Prostate Specific Antigen (PSA)	Units are ng/mL. Higher is worse.

Using these data, answer the questions below. Clearly indicate which response is for which question.

(a) Create a "Table 1" that describes the sample by treatment type, and test for overall differences across treatment types (do not test for pairwise differences, just overall differences). Do not include the outcome (bowel QOL) in your table. Include an accompanying paragraph describing and justifying any statistical tests that you use.

(b) Using appropriate statistical methods, determine if there are differences in long-term bowel-related QOL between the three treatment types. Be sure to consider potential confounders. To summarize your findings, write a short (1-2 page) summary that includes:

(i) Description of statistical methods used, and appropriate justification

(ii) Summary of results from your final model

(iii) Any concerns you have about conclusions drawn from this observational study

(c) Cancer researchers at The Ohio State University would like to perform a similar study using patients from their Radiation/Oncology and Surgery Departments, but are interested in adding a control group for comparison. Briefly (1 paragraph) describe how you would recommend selecting control subjects for such a study.

2. Fontanella, Early and Phillips (2007) present results from a study of determinants of aftercare placement for psychiatrically hospitalized adolescents. A subset of the original variables, described below, provide the data set for this question.

Table 1.8 Code Sheet for Variables in the Adolescent Placement Study

Variable	Description	Codes/Values	Name
1	Identification Code	1 - 508	ID
2	Placement	0 = Outpatient or Day Treatment 1 = Intermediate Residential 2 = Residential	PLACE
3	Age at admission	years	AGE
4	Race	0 = White 1 = Non-white	RACE
5	Gender	0 = Female 1 = Male	GENDER
6	Neuropsychiatric disturbance	0 = None 1 = Any	NEURO
7	Emotional disturbance symptoms	0 = Not Severe 1 = Severe	EMOT
8	Danger to others	0 = Unlikely 1 = Not Unlikely	DANGER
9	Elopement Risk	0 = No History of risk 1 = History of risk	ELOPE
10	Length of hospitalization	days	LOS
11	Behavioral symptoms score*	0 - 9	BEHAV
12	State custody	0 = No 1 = Yes	CUSTD
13	History of violence	0 = No 1 = Yes	VIOL

*: Behavioral symptom score is based on the sum of three symptom subscales (oppositional behavior, impulsivity, and conduct disorder) from the CSPI.

Consider the three category outcome variable PLACE,

$$\text{PLACE} = \begin{cases} 0 & \text{Outpatient or Day Treatment} \\ 1 & \text{Intermediate Residential} \\ 2 & \text{Residential} \end{cases}$$

Use the outcome variable PLACE and fit the multinomial logistic regression model with $\mathrm{PLACE} = 0$ as the referent outcome and consider as possible model covariates all other variables in the table above.

The steps should include:

1) a complete univariable analysis,
2) an appropriate selection of variables for a multivariate model (this should include scale identification for continuous covariates and assessment of the need for interactions),
3) an assessment of fit of the multivariate model,
4) an evaluation of the diagnostics statistics for the multinomial logistic regression model
5) preparation and presentation of a table containing the results of the final model (this table should contain point and interval estimates for all relevant odds ratios).

Organize your analyses and present a 3-page report that summarizes your results. Interpret the major findings in lay terms. In your report, be sure to provide justification for each of your model building steps. Key output may be included in an appendix or as tables or figures that you refer to in your report.