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

1-5 pm, June 7, 2013
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. Investigators at Leland University are planning to run a clinical trial in which they will recruit participants from different clinics in Franklin County. Their plan is to randomize participants within each clinic to treatment group with M participants from each clinic assigned the test drug and a different M participants from each clinic assigned placebo. Assume the response data resulting from this study design are generated from the following linear mixed model:

$$Y_{ijk} = \mu_j + a_i + b_{ij} + e_{ijk},$$

where Y_{ijk} denotes the response value of the kth subject ($k = 1, \ldots, M$) assigned the jth treatment ($j = 1, 2$) within the ith clinic ($i = 1, \ldots, C$), μ_j is the fixed mean response to the jth treatment, a_i is the random effect of the ith clinic, b_{ij} is a random interaction effect of the ith clinic and jth treatment, and e_{ijk} is the random effect of the kth subject assigned the jth treatment within the ith clinic. The random variables a_i, b_{ij}, and e_{ijk} are mutually independent random variables for all i, j, and k each with an expected value of 0 and respective variances equal to σ_a^2, σ_b^2, and σ_e^2.

   a. Provide an explicit expression for the correlation between the response values of two different subjects from the *same clinic* assigned the *same treatment*; i.e., $\mathrm{Corr}(Y_{ijk}, Y_{ijk'})$, where $k \neq k'$.

   b. Provide an explicit expression for the correlation between the response values of two different subjects from the *same clinic* assigned *different treatments*; i.e., $\mathrm{Corr}(Y_{i1k}, Y_{i2k'})$.

   c. Provide explicit expressions for the expected value and variance of the average response in treatment group j; i.e., the expected value and variance of

$$\overline{Y}_{\cdot j \cdot} = \frac{1}{CM} \sum_{i=1}^{C} \sum_{k=1}^{M} Y_{ijk}$$

   d. Provide explicit expressions for the expected value and variance of the estimated treatment effect

$$\hat{\tau} = \overline{Y}_{\cdot 1 \cdot} - \overline{Y}_{\cdot 2 \cdot}$$

   e. The investigators would like to perform a two sided test to determine if the mean response differs across treatment groups; i.e., $H_0 : E(\hat{\tau}) = 0$ vs. $H_a : E(\hat{\tau}) \neq 0$. Assuming that σ_a^2, σ_b^2, and σ_e^2 are known, they will use the following test statistic:

$$W = \frac{\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}.$$

   Assuming a true treatment effect of $E(\hat{\tau}) = 1$, $\sigma_a^2 = 0.8$, $\sigma_b^2 = 0.2$, $\sigma_e^2 = 1$, $C = 30$ clinics, and $M = 20$ participants in each clinic assigned to each treatment, use large sample arguments to calculate the approximate power of the two sided test at a two-sided type I error rate of 0.05.

f. Dr. A.P. Keaton will be performing the statistical analyses for this study. Dr. Keaton, however, is a PhD in Economics and doesn't have a strong background in statistics. He thinks that he can use a simple t-test to compare treatment groups (i.e., he will ignore within clinic correlations in the response values). What impact would you expect this to have on the power and type-I error rate of the test of treatment effect? Use the expressions you derived earlier to justify your answer. Also, your answer should refer to general circumstances, not under the specific treatment effect, sample size, and variances used to compute the power in part e.

2. A paper is investigating the association between hospital characteristics and hospital quality scores (a continuous variable). Specifically, they are interested in the associations between hospital location (Urban/Rural) and teaching status (teaching hospital/non-teaching hospital) with hospital quality score. .

   The general model the researchers wish to fit is:

   $$Y_i = \beta_0 + \beta_1(U_i) + \beta_2(T_i) + \epsilon_i \tag{1}$$

   where Y_i is the quality score for hospital i, U_i is an indicator variable for urban ($U = 1$ if the hospital is urban, 0 if rural) and T_i is the indicator variable for teaching status ($T_i = 1$ if the hospital is a teaching hospital; 0 otherwise). The error terms (ϵ_i) are assumed to be independent with constant finite variance and mean 0.

   You may find the following matrix inversions helpful for this question:

   $$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}^{-1} = \frac{1}{4}\begin{pmatrix} 3 & -2 & -2 \\ -2 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$$

   $$\begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

   $$\begin{pmatrix} 7 & 4 & 3 \\ 4 & 4 & 2 \\ 3 & 2 & 3 \end{pmatrix}^{-1} = \frac{1}{10}\begin{pmatrix} 4 & -3 & -2 \\ -3 & 6 & -1 \\ -2 & 1 & 6 \end{pmatrix}$$

   (i) (3 pts) In the context of this question, give the standard interpretation for β_0, β_1, and β_2.

   (ii) (3 pts) Assume that the distribution of urban/rural and teaching/non-teaching is as shown in table 1. Assume further that teaching hospitals have higher quality scores than non-teaching hospitals. The researchers are concerned that teaching status is confounding the association between hospital location and quality scores. Based on this distribution of hospitals, state whether teaching status is a confounder of the association between location and quality score. Provide an explanation for your choice. If more information is needed, state clearly what additional information you would need.

   (iii) (4 pts) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the standard ordinary least squares estimates of β_1 and β_2. Assuming the distribution of hospitals shown in table 1, give the expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of $\bar{Y}^{00}$, $\bar{Y}^{10}$, $\bar{Y}^{01}$, and $\bar{Y}^{11}$, where $\bar{Y}^{ab}$ indicates the group mean when $U = a$ and $T = b$.

Table 1: Distribution of hospitals in Scenario 1.

	Urban	Rural
Teaching	20	20
Non-Teaching	20	20

(iv) (5 pts) It turns out that the researchers didn't realize that there were no rural hospitals that were also teaching hospitals. The distribution of hospitals in the study therefore looks like that of table 2. Given this distribution of hospitals, find expressions for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ in terms of $\bar{Y}^{00}$, $\bar{Y}^{10}$, and $\bar{Y}^{11}$.

Table 2: Distribution of hospitals in Scenario 2.

	Urban	Rural
Teaching	20	0
Non-Teaching	20	20

(v) (3 pts) Based on your answer above, provide a careful interpretation of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ in the context of this problem and assuming the distribution of hospitals is as in table 2.

(vi) (4 pts) Finally, assume that in a different study, the distribution of urban/rural and teaching/non-teaching is as shown in table 3. Give the expression for $\hat{\beta}_1$ in terms of $\bar{Y}^{00}$, $\bar{Y}^{10}$, $\bar{Y}^{01}$, and $\bar{Y}^{11}$ assuming the distribution of hospitals is as in table 3.

Table 3: Distribution of hospitals in Scenario 3.

	Urban	Rural
Teaching	20	10
Non-Teaching	20	20

(vii) (3 pts) Compare and contrast your expressions for $\hat{\beta}_1$ in all three scenarios.

3. This question contains two parts, which involve survival data subject to various types of truncation and/or censoring.

   (a) Suppose failure time T is observed subject to $W \leq T \leq C$, where (W, T, C) is a random vector. We observe a sample of independent and identically distributed $(w_1, x_1, \delta_1), \ldots, (w_n, x_n, \delta_n)$ where $x_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$.

      (i) (3 pts) Identify the truncation and/or censoring pattern.

      (ii) (6 pts) Derive the likelihood function or conditional likelihood function that could be used to make inference on T (i.e., estimate the distribution or survival function of T). Clearly define any notation and assumption you may need.

      (iii) (6 pts) Provide a nonparametric estimator for the survival function of T based on observed data. (You may define $t_1^* < t_2^* < \ldots < t_D^*$ as the distinct observed failure time.)

   (b) Suppose failure time T is observed if and only if $T \leq W$, where (T, W) is a random vector. The observed data are $\{(t_i, w_i), i = 1, \ldots, n)\}$. Assume T and W are independent, and the corresponding density functions are f and g respectively.

      (i) (4 pts) What is the **sampling** joint density of (T, W)?

      (ii) (6 pts) Derive the likelihood function or conditional likelihood function that could be used to make inference on T.

4. Let $U_{1:m} < U_{2:m} < \cdots < U_{m:m}$ be the order statistics of a random sample of size m from a standard Uniform distribution.

   (a) (4 pts.) Find the mean and variance of $U_{s:m}$.

   (b) (8 pts.) Show that for $1 \leq r < s \leq m$, $Y_1 = U_{r:m}/U_{s:m}$ and $Y_2 = U_{s:m}$ are independent random variables. (Note: This property has applications in the simulation of uniform and hence general order statistics.)

   (c) (2 pts.) Identify by name along with the relevant parameters the distributions of Y_1, Y_2.

   (d) (7 pts.) Suppose $s = [mp] + 1$ where p is in $(0, 1)$ and $[\cdot]$ represents the greatest integer function. Show that as $m \to \infty$, $U_{s:m}$ converges in probability to p.

   (e) (4 pts) Show that given $U_{m:m} = w$, $(U_{1:m}, U_{2:m}, \cdots, U_{m-1:m})$ behave like the vector of order statistics from a random sample of size $m-1$ from a Uniform distribution over $(0, w)$.

You may assume that for a random sample of size n from an absolutely continuous cdf $F(x)$ with pdf $f(x)$, the joint pdf of k selected order statistics $X_{r_1:n}, \ldots, X_{r_k:n}$ where $1 \leq r_1 < r_2 < \cdots < r_k \leq n$ has the following multinomial type representation:

$$\frac{n!}{(r_1-1)!(r_2-r_1-1)!\cdots(r_k-r_{k-1}-1)!(n-r_k)!}$$
$$\times[F(x_1)]^{r_1-1}[F(x_2)-F(x_1)]^{r_2-r_1-1}$$
$$\cdots[F(x_k)-F(x_{k-1})]^{r_k-r_{k-1}-1}[1-F(x_k)]^{n-r_k}$$
$$\times f(x_1)\cdots f(x_k),$$
$$x_1 < x_2 < \cdots < x_k.$$

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

8 am—5 pm, June 8, 2013

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: Buckeyes2013! (case-sensitive)
   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. A double-blind, randomized crossover trial was conducted at OSUWMC to characterize the impact of a fast-food-type meal (high saturated fat) compared to a healthier meal (moderate level of “healthy” fats) on postprandial triglyceride responses in breast cancer survivors and benign controls. The main hypothesis for the study was:

The fast-food-type meal will provoke larger and longer-lasting postprandial (post-meal) changes in triglycerides than the healthier meal in both groups, with relatively larger changes in response to the fast-food-type meal in survivors compared to controls.

The study design was as follows: subjects each came into the Clinical Research Center (CRC) twice. During one visit they consumed the fast-food meal and during the other they consumed the healthier meal. The order in which they received the meals was randomized (i.e., which visit had which meal was random). The two meals were designed so that they looked and tasted the same – thus both the subjects and the experimenters were blind to the type of meal. During each visit to the CRC the subject consumed the meal early in the morning and remained in the CRC for 8 hours. Blood draws were taken before the meal (time 0) and at the following seven additional times: 1.5, 2, 3, 4, 5, 6, 7 hours post-meal. Triglyceride levels were measured in the blood samples for each time point. A marker of inflammation (the cytokine IL-6) was also measured at a subset of these time points (hours 0, 2, 4, 7).

A total of 57 subjects participated in the study. However, not all subjects had useable data for both visits (for various reasons; you may assume the data are missing completely at random). There are 10 subjects who only have data for one visit and the remaining 47 have data for both visits.

The data from this study are available to you in two forms: WIDE and LONG. Data dictionaries for each data file are below:

WIDE Form	**Variable Name**	**Description**
(tg_wide.csv)	subject	Unique subject identifier
	Visit	Study visit (1 or 2)
	group	Subject type: 1 = Cancer survivor, 0 = Control
	satfat	Meal type: 1 = fast-food type, 0 = healthier meal
	Triglyceride1-Triglyceride8	Triglyceride values (measured 8 times)
	il6_1-il6_4	IL-6 values (measured 4 times)

LONG Form	**Variable Name**	**Description**
(tg_long.csv)	subject	Unique subject identifier
	Visit	Study visit (1 or 2)
	group	Subject type: 1 = Cancer survivor, 0 = Control
	satfat	Meal type: 1 = fast-food type, 0 = healthier meal
	time	measurement number (1 to 8)
	hour	timing of the measurement (hours since meal)
	triglyceride	triglyceride value
	il6	IL-6 value

Using these data, answer the questions on the next page.

You should include all answers in one report, but please clearly label the sections corresponding to each question.

(a) Ignoring the groups (cancer survivors, controls), use appropriate statistical techniques to build a model and test for a difference in the triglyceride response to the fast-food meal compared to the healthy meal. To summarize your findings, write a short (1-2 page) summary that includes:
   (i) Discussion of what (if any) within-subject correlations you are concerned about given the study design.
   (ii) Detailed description of your model, including how it handles the correlations you described in (i).
   (iii) Whether there is evidence of a difference in triglyceride response for the two meals. Remember that the hypothesis was that there would be "*larger and longer-lasting postprandial (post-meal) changes*" for the fast-food meal compared to the healthier meal, so make sure to address this in your summary.

(b) Add the group indicator into your model in a way that allows you to test whether there are group differences in the triglyceride response. Briefly describe how you modified your model. Then test the second part of the investigators' hypothesis, which was that there would be "*relatively larger changes in response to the fast-food-type meal in survivors compared to controls.*"

(c) Another hypothesis of the investigators is that high triglyceride responses to the meal are associated with high inflammatory responses to the meal. Inflammation is measured by the level of the cytokine IL-6 in the blood, which was only measured at pre-meal (time 0), 2 hrs, 4 hrs, and 7 hrs post-meal. In 1-2 paragraphs describe an idea for testing the investigators' hypothesis using the data collected. You do not have to run any analyses, just describe how you might go about analyzing the data to address the hypothesis. Be specific about the statistical methods, models, etc. you would use.

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

8 am—5 pm, June 8, 2013

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: Buckeyes2013! (case-sensitive)
   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.

2. Basal cell carcinoma (BCC) tumors are the most common skin cancer and are highly immunogenic. A study was conducted to assess how immune-cell related gene expression in an initial BCC tumor biopsy was related to the appearance of subsequent BCC tumors. A total of 138 BCC patients, obtained from dermatology outpatient clinics, were screened eligible and they were not on medications that would have promoted skin cancer. Patients were followed up every six months for a maximum of three years after the biopsy date of the BCC tumor. At each assessment point, patients reported the month and year of any new BCC removals that had occurred within previous six months.   Follow-up BCC pathology data were also verified both to validate patient self-report data and identify any additional BCC tumors not reported by patients. BCC tumor-free time period was defined as the time from the biopsy date of the BCC tumor to the appearance of the first subsequent new primary BCC tumor. The censoring date for participants free of new tumors was the last contact date.

The following variables are included in the dataset q2_bcc.csv:

```
monthsToRecur:    months until recurrence or censoring
recur:            whether or not there was a recurrence
CD25pcr:          levels of mRNA for CD25, the alpha chain of the interleukin
                  (IL)-2 receptor expressed on activated T-cells and B-cells,
                  measured using real-time PCR
CD68pcr:         levels of mRNA for CD68, a marker for monocytes/macrophages,
                  measured using real-time PCR
Female:           indicator for female sex
Race:             race (character variable with name of race as the value —
                  all but 1 subject were white)
education3:       education level, collapsed into 3 levels (1=high school or
                  less, 2=some college/college grad, 3=professional/graduate
                  degree)
Age1:             age in years at the tumor removal
BCCdx:            tumor type (1=Nodular, 2=Mixed, 3=Superficial)
Subject:          subject ID
```

You are asked to import the dataset into statistical software and perform the appropriate data analysis to answer the following questions. Summarize your results in a 3-page report in lay terms. In your report, be sure to provide justification for each of your model building steps and interpret the major findings. Key output may be included in an appendix or as tables or figures that you refer to in your report.

1. What is the median recurrence-free survival time of all BCC patients?

2. CD25 and CD68 are important predictors of interest. Dichotomize CD68 into two equal-sized groups and use an appropriate test to see whether there is any difference in hazard rates between high CD68 and low CD68 groups.

3. If the investigators are more interested in the early departures between the two groups, how could you modify the above test?

4. Using the dichotomized CD68 indicator as the primary predictor, complete a thorough regression analysis to determine how CD68 affects the survival of BCC patients,

controlling for other relevant covariates. Explain any model assumptions, assess their appropriateness for these data, and explore what to do if assumptions fail.

5. Predict the probability of surviving more than 10 months for a group of female BCC patients who are college graduates with tumor removal age at 60, mixed type of tumor.

6. Another group of researchers have run a logistic regression using recurrence or not as the binary outcome. Run a logistic regression using the same set of predictors you used in part 4 and compare the results. Comment on which model is more appropriate.