Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 1: Methods and Applications

4 hours

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.

1. Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Normal(\mu, 1)$. This question concerns inference for μ, which could be any real number.

   (a) (4 pts) Write down the likelihood function to make inference about μ and state the maximum likelihood estimator (MLE) for μ (no calculation needed).

   (b) (7 pts) Suppose we need to further restrict the parameter space to $\mu \in \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, i.e., μ can only take on integer values. Find the MLE $\hat{\mu}$ for μ.

   (c) (7 pts) Identify the distribution of $\hat{\mu}$.

   (d) (7 pts) Show that $\hat{\mu}$ is unbiased.

2. Suppose that $X_1, \ldots, X_n$ are iid samples from a continuous distribution with pdf given by

$$f(x) = \frac{1}{b} \exp\left\{-\frac{x-a}{b}\right\} I\{x \geq a\},$$

where $-\infty \leq a \leq \infty$ and $b > 0$ are unknown constants, and $I(\cdot)$ is the indicator function.

(a) (5 pts) Let $Y_1 \leq \ldots \leq Y_n$ be the order statistics of $X_1, \ldots, X_n$. Find the joint distribution of $Y_1, \ldots, Y_n$.

(b) (6 pts) Let $Y_0 = 0$. $W_i = Y_i - Y_{i-1}$ for $1 \leq i \leq n$ are the *spacings*. Prove that W_i's are distributed independently, and show that the marginal pdfs are given by:

$$\frac{1}{b/n} \exp\left\{-\frac{w_1 - a}{b/n}\right\} \qquad \text{for } W_1$$

$$\frac{1}{b/(n-i+1)} \exp\left\{-\frac{w_i}{b/(n-i+1)}\right\} \qquad \text{for } W_i, 2 \leq i \leq n$$

(c) (7 pts) Find the MLEs of a and b.

(d) (7 pts) Denote $\hat{a}$ and $\hat{b}$ as the MLEs of a and b. Find the marginal distributions of $\hat{a}$ and $\hat{b}$. Are $\hat{a}$ and $\hat{b}$ independent? Why? If they are dependent, also find their joint distribution.

3. A study of an experimental drug was conducted in a sample of n mice. Mice were treated with the drug or placebo at the start of the study and followed until J mice died. Of primary interest was the effect of the drug on the mortality hazard which was modeled using the following proportional hazards model:

$$\alpha(t|X_i) = \sum_{j=1}^{J} \alpha_{0j} e^{\beta X_i} I(t_{j-1} < t \leq t_j)$$

where $X_i = 1$ if mouse i was treated with the drug (0 otherwise), $\alpha_{01}, \ldots, \alpha_{0J}$ are positive constants, and $t_1 < t_2 < \cdots < t_J$ are the unique death times in the data set with $t_0 = 0$. The data for each mouse consist of $(X_i, \widetilde{T}_i, D_i)$ $(i = 1, \ldots, n)$, were $\widetilde{T}_i = \min(T_i, t_J)$, T_i = time of death (in years), and $D_i = I(\widetilde{T}_i = T_i)$. Assume no tied death times.

(a) (5 pts) Show that the log-likelihood equals

$$l(\boldsymbol{\alpha}, \beta) = \sum_{j=1}^{J} \left[\log(\alpha_{0j}) + \beta X_{i_j} - \alpha_{0j}(t_j - t_{j-1}) \sum_{l \in R_j} e^{\beta X_l} \right]$$

where X_{i_j} denotes the treatment indicator of the mouse that died at t_j and $R_j = \{l : \widetilde{T}_l > t_{j-1}\}$.

(b) (3 pts) Find $\widehat{\alpha}_{0j}$, the maximum likelihood estimate of α_{0j} given β $(j = 1, \ldots, J)$.

(c) (4 pts) Let $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_{01}, \widehat{\alpha}_{02}, \ldots, \widehat{\alpha}_{0J})'$. Provide the expression that would be used to find the value of β that maximizes the log-profile likelihood, $l(\widehat{\boldsymbol{\alpha}}, \beta)$. Simplify your expression as much as possible.

(d) (6 pts) Suppose the investigators of the study decided to not use a piece-wise constant baseline hazard function and instead chose to leave the baseline hazard unspecified and use Cox's partial likelihood function to estimate β. Derive the expression that would be solved to find the maximum partial likelihood estimate $\hat{\beta}$ and compare it to the expression from part c.

(e) (7 pts) The investigators decided to stick with the piece-wise constant hazard model, but added a mouse-specific frailty Z_i to account for heterogeneity in the hazard function not explained by treatment group. Their new model was as follows:

$$\alpha(t|X_i, Z_i) = \sum_{j=1}^{J} \alpha_{0j} Z_i e^{\beta X_i} I(t_{j-1} < t \leq t_j)$$

where Z_i is distributed gamma with pdf

$$g(z_i) = \frac{z_i^{1/\delta - 1}}{\delta^{1/\delta} \Gamma(1/\delta)} e^{-z_i/\delta}.$$

Derive a marginal likelihood, not conditional on $Z_1, \ldots, Z_n$, that could be used to perform inference on β.

4. Consider a sample of i.i.d. random variables $\mathbf{T} = (T_1, T_2, \ldots, T_n)$, such that $T_i \sim \text{Exp}(\theta)$, $\theta > 0$, having pdf

$$f_\theta(t) = \frac{1}{\theta} e^{-t/\theta} I(t > 0) \ ,$$

where $I(\cdot)$ denotes the indicator function. Rather than observing the random variables T_i directly, we observe $Y_i = \min\{T_i, C_i\}$, $i = 1, 2 \ldots, n$, where C_i is called a censoring time. In addition, we observe

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases} \qquad\qquad i = 1, 2 \ldots, n \ .$$

Thus, the data consist of the pairs $\{Y_i, \delta_i\}$, $i = 1, 2, \ldots, n$ and the goal is to carry out inference for the parameter θ.

(a) **Fixed Censoring**
Suppose that $C_i = c$ for all $i = 1, 2, \ldots, n$, where $c > 0$ is a known constant.

(i) (4 points) Find the maximum likelihood estimator (MLE), $\hat{\theta}_f$, of θ.

(ii) (4 points) Is the MLE $\hat{\theta}_f$ from part (i) a consistent estimator of θ? Justify your answer.

(iii) (4 points) Find the Fisher information for θ based on the sample. Describe how the information varies as a function of the censoring value c.

(b) **Random Censoring**
Now suppose that the censoring times C_i are also random, and that $C_i \sim \text{Exp}(\mu)$. Assume that $\{C_i \ , i = 1, 2 \ldots, n\}$ and $\{T_i \ , i = 1, 2, \ldots, n\}$ are independent.

(i) (2 points) Find $\mathbb{P}(\delta_i = 1)$.

(ii) (2 points) Find the distribution of Y_i.

(iii) (5 points) Find the MLE, $\hat{\theta}_r$, of θ under the random censoring model.

(iv) (4 points) Is $\hat{\theta}_r$ a consistent estimator of θ? Justify your answer.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 2: Data Analysis #1

4 hours

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the instructions of each question to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam**.
4. Computer Login Information:

   Username: bioexam
   Password: xxx
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report on the desktop of the computer with the file name: Day2_*examID*, where *examID* is your assigned ID code. Also save a copy on the Flash drive provided by the proctor with the same file name.

A National Toxicology Program (NTP) study examined the carcinogenic (cancer-causing) potential of the chemical urethane. Since urethane is produced naturally during fermentation processes, the NTP examined urethane both in the presence and absence of ethanol. Male mice were randomly assigned to a dose of urethane (0, 10, 30, or 90 ppm) and a dose of ethanol (0 or 5%), with 48 mice per combination. The mice were exposed to their assigned treatment via drinking water starting at five weeks of life and continuing for two years. At the end of the study, the mice were sacrificed and examined for a specific type of tumors, called hemangiosarcomas. Mice were housed four per cage during the study, with the treatment held constant within each cage.

The Environmental Protection Agency (EPA) would like to use the NTP data to determine safe levels of exposure to urethane. They have hired you to analyze the data and provided you a data set (Excel file "q2_2015_day2_urethane.xls") containing the following variables:

Variable	**Description**
Animal	unique ID number for each animal
cage	unique ID number for each cage
wt_5	weight at 5 weeks of life (i.e., prior to exposure)
udose	dose of urethane (in ppm)
ethanol	1 if exposed to ethanol, 0 if unexposed
hem	1 if animal developed one or more tumors, 0 if not

Your task is to build a model relating urethane dose to the incidence of hemangiosarcoma and to use this model to estimate the Extra Risk. Extra Risk at dose *d* of urethane, *ER*(*d*), is calculated as follows:

$$ER(d) = \frac{\pi(d) - \pi(0)}{1 - \pi(0)}$$

where $\pi(d)$ and $\pi(0)$ are the probabilities of a hemangiosarcoma given *d* ppm and 0 ppm of urethane, respectively. Things you should consider when building your model:

1. Presence of ethanol may affect carcinogenicity of urethane (i.e., the dose response relationship).
2. Results will be sensitive to assumed trend in dose and thus you must select the most appropriate trend.
3. Animals in the same cage drank from the same water source, which may impact results.
4. Whether or not you should adjust for body weight in your analysis.

Specific instructions for writing a report summarizing your findings are on the next page.

After building your model generate a short report for the EPA containing the following five components.

Please label the components in your report with the corresponding numbers (1-5).

1) A detailed description of your statistical methods (this will be reviewed by statisticians so it is okay for this to be somewhat technical). This must include a description of how you selected your model.
2) A table containing appropriate summary statistics for body weight at five weeks of age.
3) A table containing the following information for each treatment group:
    a. Proportion of animals who experienced a hemangiosarcoma
    b. Estimated probability of a hemangiosarcoma based on your final model and a 95% confidence interval
    c. Estimated extra risk (no confidence interval needed)
4) A graph containing both the dose-response curve (i.e., estimated probabilities at various doses) and observed proportion with hemangiosarcoma in each dose group.
5) A one-paragraph summary of the results of your analysis for a non-statistical audience. The summary must contain the following information:
    a. Results of a statistical test determining whether or not the dose-response relationship between urethane and hemangiosarcoma incidence is affected by ethanol exposure.
    b. Results from a statistical test determining if there is any effect of urethane on incidence of hemangiosarcoma.
    c. A goodness-of-fit assessment (may be done using an informal graphical assessment).
    d. The urethane dose corresponding to ER=0.1 (an approximation is sufficient); this will be used as a safety dose by the EPA.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 3: Data Analysis #2

4 hours

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the project instructions to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam except if needed to download software add-ons (e.g., R packages, Stata files)**.
4. Computer Login Information:

   Username: bioexam
   Password: xxx
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report (in **one file**) on the desktop of the computer with the file name: **Day2_*examID***, where *examID* is your assigned ID code.

A recent randomized controlled trial was conducted to compare two treatment regimens for behavior management for children with attention-deficit /hyperactivity disorder (ADHD). All of the children had severe aggression. The two treatments that were compared were:

Basic: parent training in behavior management + stimulant medication
Augmented: parent training in behavior management + stimulant medication + risperidone

The trial lasted 9 weeks, and the protocol was for children to come into the clinic for weekly visits. However, some children dropped out of the study and did not complete the full 9 weeks. The primary study goal was to determine whether adding the antipsychotic medication risperidone reduced aggressive behaviors. That analysis has already been completed. Instead, you will focus on evaluating the occurrence of adverse events in the two groups.

Risperidone is not without side effects. A major concern is the occurrence of adverse events (AEs) that are rated by a clinician as "severe" – some examples might be vomiting, headache, trouble falling asleep, etc. At each study visit, parents reported whether any AEs had occurred, and a clinician determined its severity (i.e., "severe" or "not severe"). You have been provided data on the severe AEs reported for each study participant during the course of the trial. To facilitate your data analyses, the data are provided both in "wide" form and in "long" form as described below:

"Wide" Data Set (q2_2016_day2_wide.csv)

Variable	**Description**
subnum	Unique child identifier
trt	Treatment group (0=Basic, 1=Augmented)
site	Clinical site (1,2,3,4)
age	Age of the child (at week 0)
gender	Sex of the child (1=male, 2=female)
vnumLastVisit	Last study visit: week number (e.g., 2=completed through week 2 of study)
daysLastVisit	Last study visit: days since treatment start *Note: a child's visits might not be exactly 7 days apart, hence a child who completed the entire study – 9 weeks – might not have his last visit exactly on day 9x7=63). Similarly a child who completed 4 visits might not have his last visit on exactly day 4x7=28.*
days1-days8	Days since treatment started when **1st through 8th** AE occurred *Note: No subject had more than 8 reported AEs*
ae1-ae8	Indicator for if an AE occurred – 1 if AE occurred, missing otherwise *Note: these variables take the value 1 if the corresponding "days" variable has a non-missing value, and are missing otherwise.*

"Long" Data Set (q2_2016_day2_AElong.csv)

Variable	**Description**
subnum	Unique child identifier
daysAE	Day on which an AE was reported (days since treatment start)

Note: If a child never had any AEs reported, they do not appear in this "long" data set.

For the purposes of this exam, you will focus on **three specific research questions**:

1. Is there a difference between the treatments regimens in the occurrence of at least one adverse event?
2. The study was conducted at four different sites (clinics). Since the AEs are rated by clinicians, there is concern that there may be a “site effect”. Is there evidence of a “site effect” in the occurrence of at least one adverse event?
3. In addition to caring about whether a child experienced any AEs, we also care about the cumulative experience of AEs – i.e., did children in one treatment group experience more AEs across the whole study period than children in the other treatment group?

Your Assignment

You are asked to perform the appropriate data analysis to answer each of the three research questions above. Summarize your results in a 1-2 page report in lay terms that explains (a) the analysis methods you used, (b) the results that answer the questions, and (c) alternative strategies you considered and the strengths/weaknesses of your chosen approach. Key output from statistical software may be included in an appendix or as tables or figures that you refer to in your report.

Please clearly label the sections in your report corresponding to each research question.