Biostatistics PhD Comprehensive Exam: Theory

June 1 - 4, 2021

Instructions: Please adhere to the following guidelines:

- The PhD Theory Comprehensive Exam will be administered on Tuesday, June 1 at 9:00am (central time); you have until Friday, June 4 at 12:00pm (central time) to complete the exams and place your responses into your respective Box folder. You may (should) place draft solutions in your Box folder throughout the examination period; the latest version submitted prior to the deadline will be considered the final version. In addition, please also email your final version to Drs. Andrew Spieker (**andrew.spieker@vumc.org**) and Robert Greevy (**robert.greevy@vumc.org**) prior to the deadline (dual submission helps ensure the exam is received).
- There are six equally weighted problems of varying length and difficulty. Note that not all subproblems are weighted equally. You are advised not to spend too much time on any one problem.
- Answer each question clearly and to the best of your ability. Partial credit will be awarded for partially correct answers.
- Be as specific as possible, show your work when necessary, and please write legibly for any handwritten responses.
- This is an open-book and open-notes examination, but it is an *individual effort*; do not discuss any part of this exam with anyone. Vanderbilt University's academic honor code applies.
- Please email any clarifying questions to:
  Dr. Andrew Spieker (andrew.spieker@vumc.org),
  Dr. Matt Shotwell (matt.shotwell@vumc.org), and
  Dr. Bob Johnson (robert.e.johnson@vumc.org).

---

1. $\boxed{\textbf{25 pts}}$ Let $(\Omega, \mathcal{F}, P)$ denote a probability space, and let $\{A_n \in \mathcal{F}\}_{n=1}^{\infty}$ denote a sequence of events, each having associated probability measure $\mathrm{P}(A_n) = \frac{1}{n^2}$. Let $X_n(\omega) = n^2 \mathbb{I}_{A_n}(\omega) - 1$ denote a sequence of random variables, where

$$\mathbb{I}_{A_n}(\omega) \quad = \quad \begin{cases} 1 & \text{if } \omega \in A_n \\ 0 & \text{otherwise} \end{cases}.$$

(a) For each n, determine the values of $\mathrm{E}[X_n]$ and $\mathrm{Var}[X_n]$.

(b) Determine the distribution function, $F_{X_n}(t)$, of X_n.

(c) On separate graphs, plot $F_{X_n}(t)$ for $-5 \leq t \leq 20$ when $n = 2$, $n = 3$, and $n = 4$ (it is acceptable to use R or to draw the figure legibly by hand). Briefly explain the behavior of $F_{X_n}(t)$ as n grows.

(d) Let $X \equiv -1$ denote a degenerate random variable with CDF $F_X(t) = \mathbb{I}(t \geq -1)$. Show that

$$\lim_{n \longrightarrow \infty} |F_{X_n}(t) - F_X(t)| = 0 \text{ for all } t \in \mathbb{R}.$$

Does X_n converge to X in distribution?

(e) Prove that $X_n \xrightarrow{a.s.} X$.

(f) Prove that there exists no random variable Y such that $X_n \xrightarrow{\mathcal{L}^1} Y$.

---

2. $\boxed{\textbf{25 pts}}$ Your client is a doctor seeking to model the time it takes patients to receive medical care in her solo practice. You may assume time to be measured in discrete, integer-valued (non-negative) units. Let δ_n denote the number of patients arriving to the clinic at time n, with probability mass function given by $\mathrm{P}(\delta_n = k) = \alpha^k(1-\alpha)^{1-k}$ for $\alpha > 0$ and $k = 0, 1$ (that is, no more than one patient can arrive at a single time). You may further assume that the δ_n's are mutually independent.

An arriving patient waits in a queue (if there is one), which is served by a single receptionist. When arriving to the front of the queue, the patient is directed to the examination room and receives one of a number of medical care services. The time to render that service is distributed as a discrete random variable S with probability mass function given by

$$\mathrm{P}(S = k) = \begin{cases} p_k & \text{if } k = 1, 2, \ldots K \\ 0 & \text{otherwise} \end{cases},$$

for some fixed and known value K (you may assume that the services received by the patients are mutually independent). Let W_n denote the total time a patient arriving at time n will spend until he or she receives care (that is, the time spent in queue prior to receiving the service).

(a) Determine an expression for the expected time between arrivals (in terms of α).

(b) Determine an expression for the expected service time (in terms of $p_1, \ldots, p_K$).

(c) Argue that $W_{n+1} = (W_n + S_n\delta_n - 1)^+ = \max(0, W_n + S_n\delta_n - 1)$, where S_n are i.i.d. random variables distributed like S. Provide a plain-language interpretation of this equation for your client.

(d) Argue that $\{W_n\}$ is a Markov chain, and describe the transition probabilities in terms of α and p_k.

For the remainder of this question, suppose $K = 2$, with $p_1 = 1 - \beta$ and $p_2 = \beta$, for some $\beta \in (0, 1)$.

(e) Describe the transition probabilities in terms of α and β.

(f) Determine the expected time between arrivals and the expected service time in terms of α and β.

(g) Determine conditions on α and β such that the stationary (steady-state) distribution exists. Provide a plain-language interpretation of these conditions for your client.

(h) Determine the stationary distribution, denoted by $\pi_1, \pi_2, \cdots$, and explicitly name the family of distributions to which it belongs.

(i) Determine the expected waiting time in steady-state.

(j) Suppose that $\alpha = 4/5$ (that is, 4 patients arrive every 5 units of time, on average). Determine the maximum value that β can take such that a stationary distribution exists.

(k) Suppose that $\alpha = 4/5$ and $\beta = 0.24$. Determine the expected waiting time, in steady-state. Provide a plain-language interpretation of this result for your client with respect to individual service times.

---

3. $\boxed{\textbf{25 pts}}$ Survival analysis methods often focus on modeling the hazard function, which uniquely determines the distribution of the (continuous) survival time, T. Let $\lambda_i(t)$ denote the subject-specific time-varying hazard function for independently sampled subjects $i = 1, \ldots, n$. One way to model the subject-specific hazard is to consider it equal to some "baseline" hazard, $\lambda_0(t)$, times a positive-valued random variable, G, that we refer to as the *frailty*:

$$\lambda_i(t) = \lambda(t|G = g) = \lambda_0(t)g$$

Assume without loss of generality that $\mathrm{E}[G] = 1$. When $G = 1$, the subject-specific hazard corresponds to the hazard of an "average" subject. Subjects having $G > 1$ have a higher hazard (lower mean survival), while those with $G < 1$ have a lower hazard (higher mean survival). Variation in G serves as a source of variation in time-to-event outcome apart from that which is explainable by the hazard function alone. Because a subject's frailty cannot be observed, a frailty distribution must be assumed. One choice for G is the inverse Gaussian distribution with probability density function depending upon $\mu > 0$ and $\tau > 0$:

$$f_G(g; \mu, \tau) = \sqrt{\frac{\tau}{2\pi g^3}} \exp\left(-\frac{\tau(g-\mu)^2}{2\mu^2 g}\right), \text{ for } g > 0.$$

Denote this distribution as $\mathrm{IG}(\mu, \tau)$.

(a) Express the expectation and variance of $G \overset{d}{=} \mathrm{IG}(\mu, \tau)$ in terms of μ and τ, and determine the values of μ and τ such that $\mathrm{E}[G] = 1$ and $\mathrm{Var}[G] = \sigma^2$.

(b) Assume a frailty distribution parameterized by $G \overset{d}{=} \mathrm{IG}(1, 1/\sigma^2)$. Under this parameterization, it is possible to show that the conditional hazard function is given by $\lambda(t|T \geq t) = \lambda_0(t)(1 + 2\sigma^2\Lambda_0(t))^{-1/2}$, where $\Lambda_0(t)$ is the baseline cumulative hazard function. In the specific case where $\sigma^2 = 1$ and $\lambda_0(t) = 1$,

   i. Determine the baseline cumulative hazard function, $\Lambda_0(t)$.

   ii. Use R to plot the density function of $f_G(g)$.

   iii. Use R to plot the conditional hazard function $\lambda(t|T \geq t)$.

   What does this suggest about the frailty of survivors as t increases?

(c) An interesting property of the inverse Gaussian distribution is that it is related to first passage times in a Brownian motion. Suppose (W_s) is a Wiener process (a standard Brownian motion) where $s \geq 0$. Define $S_a = \inf\{s > 0 : W_s \geq a\}$ where $a > 0$ is a real constant. S_a is the random time it takes the Wiener process to first equal or exceed a. Prove that $S_a \overset{d}{=} \lim_{\mu \longrightarrow \infty} \mathrm{IG}(\mu, a^2)$.

(d) Now $X_s = \nu s + \phi W_s$ where $\nu > 0$ and $\phi > 0$ (note that (X_s) is known as a Brownian motion with drift), and let $S_a = \inf\{s > 0 : X_s \geq a\}$.

   i. Use R to demonstrate empirically that $S_a \overset{d}{=} \mathrm{IG}\left(\frac{a}{\nu}, (\frac{a}{\phi})^2\right)$.

   ii. Describe the Brownian motion with drift that corresponds to the frailty distribution $\mathrm{IG}(1, 1/\sigma^2)$.

---

4. $\boxed{\textbf{25 pts}}$ Suppose we collect multiple independent data points on some outcome Y at each of K distinct values of some exposure X. Consider the "no-intercept" linear regression model

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1N_1} \\ y_{21} \\ \vdots \\ y_{2N_2} \\ \vdots \\ y_{K1} \\ \vdots \\ y_{KN_K} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_1 \\ x_2 \\ \vdots \\ x_2 \\ \vdots \\ x_K \\ \vdots \\ x_K \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \\ \vdots \\ \epsilon_{K1} \\ \vdots \\ \epsilon_{KN_K} \end{pmatrix},$$

for some real-valued, unknown parameter β. In this problem, you may assume the errors ϵ_{kj} to be pairwise independent, to be of mean zero, and to have constant variance σ^2.

(a) Determine the least squares estimator for β—namely, the estimator that minimizes the following quantity:

$$\sum_{k=1}^{K} \sum_{j=1}^{N_k} (y_{kj} - x_k \beta)^2.$$

(b) Show that the least squares estimator you derived in part (a) also minimizes the following quantity:

$$\sum_{k=1}^{K} N_k (\overline{y}_k - x_k \beta)^2,$$

where $\overline{y}_k = N_k^{-1} \sum_{j=1}^{N_k} y_{kj}$ denotes the sample mean value of the outcome Y among all observations with exposure value $X = x_k$.

(c) Show that the least squares estimator you derived in part (a) is exactly the same as the weighted least squares estimate obtained from the linear model

$$\begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_K \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \\ \vdots \\ \epsilon_K^* \end{pmatrix},$$

where the weights are given by N_k for $k = 1, \ldots, K$.

---

5. $\boxed{\textbf{25 pts}}$ Let $\ell(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$ denote the sum of squared errors for a linear regression model, where $\boldsymbol{y}$ is an n-vector and $\mathbf{X}$ is an $n \times p$ matrix of covariates. The vector of coefficients, $\boldsymbol{\beta}$, is said to be *estimable* if and only if $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}')$ implies that $\boldsymbol{\beta} = \boldsymbol{\beta}'$, for all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ that minimize (globally) $\ell(\boldsymbol{\beta})$. In plain language, $\boldsymbol{\beta}$ is estimable if and only if $\ell(\boldsymbol{\beta})$ possesses a unique global minimum. Note that a global minimum must satisfy the estimating equation $\ell'(\boldsymbol{\beta}) = \mathbf{0}$, where $\ell'(\boldsymbol{\beta})$ denotes the gradient evaluated at $\boldsymbol{\beta}$.

(a) Let $\widehat{\boldsymbol{\beta}}$ denote a global minimum of $\ell(\boldsymbol{\beta})$. Write an expression to approximate $\ell'(\boldsymbol{\beta})$ in a neighborhood about $\widehat{\boldsymbol{\beta}}$ using a first-order Taylor expansion. Argue that $\ell''(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \neq 0$ is a condition for estimability of $\boldsymbol{\beta}$, where

$$\ell''(\widehat{\boldsymbol{\beta}}) = \left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right]_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}}.$$

(b) Compute the value of $\ell''(\widehat{\boldsymbol{\beta}})$. What does the estimability condition imply about the matrix $\mathbf{X}$?

(c) Now consider the ridge-penalized residual sum of squares $\ell(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$. Show that the ridge regression estimate of $\boldsymbol{\beta}$ can be obtained by ordinary least squares regression using an augmented data set $\mathbf{X}'$ and $\boldsymbol{y}'$, where $\mathbf{X}'$ is the covariate matrix $\mathbf{X}$ augmented with p additional rows defined by $\sqrt{\lambda}\mathbf{I}$, and $\boldsymbol{y}'$ is $\boldsymbol{y}$ is augmented with p zeros. Using the augmented covariate matrix $\mathbf{X}'$, argue that $\boldsymbol{\beta}$ is always estimable.

(d) When $\ell(\boldsymbol{\beta})$ is a likelihood function, similar logic defines an estimability condition for a maximum likelihood estimate, where $-\ell''(\widehat{\boldsymbol{\beta}})$ is the observed Fisher information. What does the estimability condition imply about the observed Fisher information matrix?

(e) Consider the data augmentation method described in part (c). Would the augmented data $\mathbf{X}'$ and $\boldsymbol{y}'$ ensure estimability for a maximum likelihood estimate? Explain why, or why not.

---

6. $\boxed{\textbf{25 pts}}$ Let $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with unknown mean μ and known variance $\sigma^2 = 1$. Suppose you are asked to use the sample mean, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$, to decide between the following three decisions:

- State that $\mu < 0$
- Abstain from making a statement about the value of μ
- State that $\mu > 0$

For convenience, refer to these three decisions numerically as $d = -1$, $d = 1$, and $d = 0$, respectively. Further, let $L(\mu, d) = 1 - d \times \text{sign}(\mu)$ denote the *loss* function for this decision problem, and let $R(\mu, d) = \text{E}[L(\mu, d)]$ denote the *risk* (as a function of μ) associated with the decision rule d.

(a) Fill in the 3×3 table below with the corresponding values of $L(\mu, d)$:

Decision	Description of decision	$\mu < 0$	$\mu = 0$	$\mu > 0$
$d = -1$	State that $\mu < 0$			
$d = 0$	Abstain from statement			
$d = 1$	State that $\mu > 0$			

Very briefly explain this choice of a loss function.

(b) Consider the specific decision rule $\delta(\overline{x}) = \text{sign}(\overline{x}) \times \mathbb{I}(|\overline{x}| > 1)$. Plot $\delta(\overline{x})$ as a function of $\overline{x}$ (it is acceptable to use R or to draw the figure legibly by hand).

(c) Determine the risk function $R(\mu, \delta(\overline{x}))$ as a function of μ and n. You may of course use the notation $\Phi(\cdot)$ for the standard normal CDF in your response.

(d) Using your response to part (c), show that $R(0, \delta(\overline{x})) = 1$ for all n.

(e) Using your response to part (c), prove that $\lim_{n \longrightarrow \infty} R(\mu, \delta(\overline{x})) = \mathbb{I}(|\mu| < 1) + 0.5 \times \mathbb{I}(|\mu| = 1)$.

(f) On a single plot, graph $R(\mu, \delta(\overline{x}))$ as a function of μ for:

- $n = 1$
- $n = 10$
- $n = 50$
- $n = 100{,}000$

This plot should be consistent with the statements in parts (d) and (e).

(g) The decision rule δ_1 is said to be *asymptotically dominated* by the decision rule δ_2 if for all values of μ,

$$\lim_{n \longrightarrow \infty} R(\mu, \delta_2) \leq \lim_{n \longrightarrow \infty} R(\mu, \delta_1),$$

and there exists at least one value of μ (call it μ^*) for which

$$\lim_{n \longrightarrow \infty} R(\mu^*, \delta_2) < \lim_{n \longrightarrow \infty} R(\mu^*, \delta_1).$$

Propose a decision rule that you would expect to asymptotically dominate $\delta(\overline{x})$ based on an extremely simple modification to $\delta(\overline{x})$. Although you needn't redo the math, please heuristically argue your choice.

(h) A decision rule is said to be *asymptotically admissible* within a class of decision rules if it cannot be asymptotically dominated by another decision rule in that class. Consider the set of decision rules for this problem of the form $\delta_a(\overline{x}) = \text{sign}(\overline{x}) \times \mathbb{I}(|\overline{x}| > a)$ with $a > 0$. Argue heuristically that no asymptotically admissible decision rule exists within this special class.
