Vanderbilt University Biostatistics Comprehensive Examination

PhD Theory Exam Series 2

May 20–May 23, 2025

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Tuesday, May 20 at 9:00am, and will be due on Friday, May 23 at 5:00pm. This deadline is strict: late submissions will not be accepted.
- To turn in your exam, please use your assigned Box folder and e-mail your word-processed exam to Dr. Andrew Spieker by the deadline. This level of redundancy is designed to ensure that your exam is received by the deadline. If you would like to e-mail exam drafts along the way, that is perfectly acceptable—do not be concerned about spamming my inbox.
- There are three problems. Note that not all questions and their sub-questions are weighted equally. You are advised to pace yourself and to not 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 in your responses.
- You may consult reference material (e.g., course notes, textbooks), though the work you turn in must be your own (this means no generative AI). This is an *individual effort*. Do not communicate about the exam with anyone. Vanderbilt University's academic honor code applies.
- Please direct clarifying questions by e-mail to Dr. Andrew Spieker.

1. $\boxed{\text{25 pts}}$ Suppose $\{X_k\}_{k=1}^{\infty}$ is a sequence of independent random variables such that $X_k \sim \text{Poisson}(\lambda = 1/k)$.

(a) Show that the sequence $\{X_k\}_{k=1}^{\infty}$ converges in $\mathcal{L}^1$ to zero. That is, show that

$$\lim_{k\to\infty} \mathrm{E}[|X_k|] = 0.$$

(b) Define the following partial sum:

$$S_n = \sum_{k=1}^{n} X_k \quad \text{for } n = 1, 2, \ldots$$

Determine whether there a finite constant, c, such that S_n converges in $\mathcal{L}^1$ to c. Justify your answer.

(c) Use the Lyapunov central limit theorem to identify suitable sequences, $\{\mu_n\}_{n=1}^{\infty}$ and $\{\sigma_n\}_{n=1}^{\infty}$, such that

$$\frac{S_n - \mu_n}{\sigma_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

You should support your answer by verifying that the Lyapunov condition is satisfied.

(d) Illustrate the result of part (c) with a very simple simulation. It is completely adequate to choose a single value of n for which to illustrate this. **Please include your R code as a supplementary appendix.**

---

2. $\boxed{\text{40 pts}}$ Suppose that $X_1, \ldots, X_{2n+1}$ are independent random variables (note that the number of observations is always odd by this notation), each with common density function:

$$f_X(x;\theta) = \frac{1/\pi}{1+(x-\theta)^2}; \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Note that this distribution has a median of θ and an undefined mean. For this problem, you may use computational tools such as WolframAlpha to compute any pesky integrals (there will be a couple).

(a) Possibly after consulting reference material, state the asymptotic distribution of the sample median, $X_{(n+1)}$, and use this result to conclude that $\widetilde{\theta}_{2n+1} = X_{(n+1)}$ is a consistent estimator of θ.

(b) Consider maximum likelihood estimation of θ. Determine the score equation for θ based on the $2n+1$ observations (do not attempt to solve it by hand, as no closed-form solution exists).

(I) State the maximum number of real-valued solutions that this score equation can possess.

(II) State why the score equation must possess at least one real-valued solution.

(III) Let $\theta^{(0)}$ denote an initializer for a Newton-Raphson approach to numerically identify a root of the score equation. Determine the form of a Newton-Raphson step, $\theta^{(j+1)}$, starting from $\theta^{(j)}$.

(IV) Assume that the initializer has been chosen so that the Newton-Raphson algorithm converges to the global maximizer of the likelihood (namely, the MLE). Determine the asymptotic relative efficiency of the MLE relative to $\widetilde{\theta}_{2n+1}$.

(c) Indeed, $f_X(x;\theta)$ represents a Cauchy distribution having known scale $\gamma = 1$ and a location parameter of θ. Reeds (1985) investigates certain properties of this distribution. Use information provided in this paper to conclude the approximate probability, in large samples, that the log-likelihood possesses a single local maximum that is hence the global maximum (*Hint*: You need not read very much of the paper at all to gain the information you need to answer this. In fact... try reading the abstract...). If you're inclined, put this paper on your list of papers to read in greater detail later (post-exam...).

(d) The result of part (c) underscores the importance of selecting a "good" value for $\theta^{(0)}$. It should seem intuitive that selecting $\theta^{(0)}$ to be a consistent estimator of θ will mitigate the risk of false or failed convergence in large samples. Choosing $\theta^{(0)}$ to be a consistent estimator has the *additional* advantage that asymptotic efficiency can be achieved by merely applying a *single* Newton-Raphson step beyond $\theta^{(0)}$, even if $\theta^{(0)}$ is itself inefficient (aptly named the "one-step estimator"). Explore this by simulation, with $\theta = 0$ and $2n + 1 = 5{,}001$. Consider the following estimators for θ, described as follows:

(I) $\widehat{\theta}_a$: The sample median (i.e., $\widetilde{\theta}_{2n+1}$ from part (a)).

(II) $\widehat{\theta}_b$: Initialize $\theta^{(0)} = \widetilde{\theta}_{2n+1}$ and take a single Newton-Raphson step.

(III) $\widehat{\theta}_c$: Initialize $\theta^{(0)} = \widetilde{\theta}_{2n+1}$ and iterate Newton-Raphson until convergence.

(IV) $\widehat{\theta}_d$: Initialize $\theta^{(0)} = \overline{X}_{2n+1}$ and take a single Newton-Raphson step.

(V) $\widehat{\theta}_e$: Initialize $\theta^{(0)} = \overline{X}_{2n+1}$ and iterate Newton-Raphson until convergence.

(VI) $\widehat{\theta}_f$: Initialize $\theta^{(0)} = 0.5$ and take a single Newton-Raphson step.

(VII) $\widehat{\theta}_g$: Initialize $\theta^{(0)} = 0.5$ and iterate Newton-Raphson until convergence.

Note: "until convergence" means to *any* local optimum; further, please exclude any cases of convergence failure from consideration (they should be rare if they occur at all). Report the empirical mean and variance of $n^{1/2}\widehat{\theta}_a, \ldots, n^{1/2}\widehat{\theta}_g$ based on $M = 10{,}000$ simulation iterations. Thoughtfully comment on and carefully account for your findings. **Include your R code as an appendix.**

(e) Again, let $\theta = 0$ and consider the procedure associated with $\widehat{\theta}_e$ in particular. Provide a mathematical argument that this procedure fails to converge in probability to the proper value of $\theta = 0$ about 45% of the time. You may use the fact that $\overline{X}_{2n+1}$ has density $f_X(x;0)$ without further justification.

(f) Illustrate the result of part (e) by simulation under a sample size of $2n + 1 = 5{,}001$; in doing so, you will likely find it helpful to compare this approach to one of the methods of part (d) that is already understood to have good behavior—determine the proportion of times each procedure converges to the same quantity across many simulation replicates. **Again, include your R code as an appendix.**

---

3. $\boxed{\text{35 pts}}$ Suppose we are in the setting of a matched-pair case-control study. For the i^{th} observation in the k^{th} matched pair, let:

$$Y_{ik} = \begin{cases} 1 & \text{case} \\ 0 & \text{control} \end{cases} \quad \text{and} \quad X_{ik} = \begin{cases} 1 & \text{exposed} \\ 0 & \text{unexposed} \end{cases},$$

for $i = 1, 2$ and $k = 1, \ldots, K$ (*any* asymptotics in this problem are to be taken as $K \to \infty$). The model that includes unconstrained intercepts for each matched pair is given by:

$$\operatorname{logit}(\mathrm{P}(Y_k = 1 | X_k = x_k)) = \alpha_k + \beta x_k. \tag{1}$$

The odds ratio, e^{β}, is the parameter of interest. The goal of this problem is to rigorously argue that standard logistic regression does not produce a consistent estimate of β by breaking the problem into discrete steps.

(a) Prove that $K^{-1}\sum_{k=1}^{K}[X_{1k}(1 - X_{2k}) - e^{\beta} X_{2k}(1 - X_{1k})] \xrightarrow{p} 0$. *Hint*: Let n_{10} denote the number of matched pairs in which the case is exposed but the control is not, and let n_{01} denote the number of matched pairs in which the control is exposed but the case is not (together, these are referred to as the *discordant pairs*). Setting (without loss of generality) $Y_{1k} = 1$ to be the case and $Y_{2k} = 0$ to be the control, note that:

$$n_{10} = \sum_{k=1}^{K} X_{1k}(1 - X_{2k}) \quad \text{and} \quad n_{01} = \sum_{k=1}^{K} X_{2k}(1 - X_{1k}).$$

Considering the fact that e^{β} is also the odds ratio that compares the *exposure* between the a case and control within a matched pair, write the model as $\operatorname{logit}(\mathrm{P}(X_{ik} = 1 | Y_{ik} = y_{ik}) = \alpha_k^* + \beta y_{ik}$. Then, consider the estimating function $\mathbf{G}(\beta; X_{1K}, X_{2K}) = X_{1K}(1 - X_{2K}) - e^{\beta} X_{2K}(1 - X_{1K})$, show that it has expectation zero, and continue forward from there using standard asymptotics.

(b) Prove that $n_{10}/n_{01} \xrightarrow{p} e^{\beta}$. *Hint*: Continue forward from the result of part (a).

(c) You have previously learned about the duality between Poisson and logistic models for tabular data. To that end, let Z_{ijk} denote the ij^{th} entry in a 2×2 table within the k^{th} matched pair (row $i = 1$: exposed; column $j = 1$: case). Define Q_{ijk} and R_{ijk} as follows:

$$Q_{ijk} = \begin{cases} 1 & \text{exposed} \\ 0 & \text{unexposed} \end{cases} \quad \text{and} \quad R_{ijk} = \begin{cases} 1 & \text{case} \\ 0 & \text{control} \end{cases}.$$

The dual Poisson model based on this setup (no proof required) is given by $Z_{ijk} \sim \text{Poisson}(\lambda_{ijk})$, where

$$\log(\lambda_{ijk}) \quad = \quad \delta_{0k} + \gamma_k Q_{ijk} + \alpha_k R_{ijk} + \beta Q_{ijk} R_{ijk}. \tag{2}$$

This model's associated design matrix, $\mathbf{X}$, has $4 \times K$ rows and $3K + 1$ columns. Express $\mathbf{X}$ and $\boldsymbol{z}$ when there are four pairs: one in which both the case and control are unexposed, one in which they are both exposed, one in which only the control is exposed, and one in which only the case is exposed. *Note*: Should you wish to confirm any suspicions you may have for the remaining parts of this problem, it is advisable to use this very simple example as your anchor. However, present your answers to the remaining parts of the problem in the more general case.

(d) Recalling the form of the score equations for a generalized linear model, argue that $\mathbf{X}^{\intercal}\boldsymbol{z} = \mathbf{X}^{\intercal}\widehat{\boldsymbol{\lambda}}$. Conclude from this that the marginal totals for the 2×2 table associated with the k^{th} pair must match the marginal totals of its respective fitted values.

(e) The concordant pairs are those for which both the case and the control share an exposure status. Argue (possibly heuristically) that such pairs do not contribute to the maximum likelihood estimate of β.

(f) Argue that for the discordant pairs, $\widehat{\lambda}_{12k} = \widehat{\lambda}_{21k}$ and $\widehat{\lambda}_{11k} = \widehat{\lambda}_{22k}$, and that these values do not depend upon k.

(g) Argue that $\exp(\widehat{\beta}) = (\widehat{p}/(1 - \widehat{p}))^2$, where $\widehat{p} = \widehat{\mathrm{P}}(Y_k = 1 | X_k = 1)$.

(h) Argue that $\widehat{p} = n_{10}/(n_{10} + n_{01})$.

(i) Argue that $\exp(\widehat{\beta}) \xrightarrow{p} \exp(2\beta)$.
