Vanderbilt University Biostatistics Comprehensive Examination

PhD Theory Exam Series 2

May 31–June 3, 2022

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Tuesday, May 31 at 9:00am, and will be due on Friday, June 3 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 and Dr. Robert Greevy 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 our inboxes.
- There are six problems. Note that not all questions and 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.
- This exam is open-everything, but remains 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 and Dr. Bob Johnson.

---

1. [25 pts] Suppose $Y_1, \ldots, Y_n$ are independent random variables with $Y_k \sim \text{Poisson}(x_k\lambda)$, where $x_k > 0$ is a real constant and $\lambda > 0$. Define $\sigma_n^2 = \sum_{k=1}^n s_k^2$ where $\text{Var}[Y_k] = s_k^2$. Let $Z_k = Y_k - \text{E}[Y_k]$, and define $S_n = \sum_{k=1}^n Z_k$.

(a) Suppose $x_k = x_{k+1}$ for $k = 1, 2, \ldots, n$. Is it the case that $S_n/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$? Explain.

(b) Show that the Lindeberg condition holds true when $x_k = 1$ for $k = 1, 2, \ldots, n$.

(c) What conditions on the x_k's are required for the Lyapunov condition to hold true?

(d) Suppose $x_k = 1/k$ for $k = 1, 2, \ldots, n$. Is it the case that $S_n/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$? Explain.

(e) Suppose $x_k = 1/k^2$ for $k = 1, 2, \ldots, n$. Is it the case that $S_n/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$? Explain.

---

2. [25 pts] Let $p_n(u)$ be the discrete uniform distribution with support $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\right\}$. Let U_n be a random variable that is distributed according to $p_n(u)$.

(a) Find the value of $\mathrm{P}\{U_n \leq e^{-1}\}$ for each of $n = 3, 5, 7$.

(b) Express $\mathrm{P}\{U_n \leq u\}$, for $u \in (0, 1)$, as a simplified function of n for any integer $n \geq 1$.

(c) Suppose $U_1, U_2, \ldots$ is a sequence of independent random variables with U_n distributed according to $p_n(u)$. Let $U \sim \text{Uniform}(0, 1)$ (continuous) and suppose U and $U_1, U_2, \ldots$ are defined on the same probability space.

i. Prove that $U_n \xrightarrow{d} U$. You may produce a figure or figures to support your proof.

ii. Argue that $U_n \overset{a.s.}{\not\longrightarrow} U$.

iii. Can it be the case that $U_n \xrightarrow{p} U$? Explain.

---

3. [25 pts] Let $B(\cdot)$ be a standard Brownian motion defined over the interval $[0,1]$. Define the events

$$A_n = \left\{ \left| B\left(\frac{1}{n}\right) - B\left(\frac{1}{n+1}\right) \right| \geq \sqrt{2\ln n} \left| \frac{1}{n} - \frac{1}{n+1} \right|^{1/2} \right\}$$

for positive integers $n \geq 1$.

(a) Argue that the events $A_1, A_2, \ldots$ are independent.

(b) Derive an expression for $\mathrm{P}(A_n)$ as a function of $\Phi(\cdot)$, the standard normal CDF.

(c) Confirm the approximation bounds for $1 - \Phi(t)$, $t > 0$, given by:

$$\frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} \exp(-t^2/2) < 1 - \Phi(t) < \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2).$$

Also, for $n \geq 4$, confirm the inequality:

$$\frac{2\sqrt{2\ln(n)}}{2\ln(n)+1} \geq \frac{1}{\sqrt{\ln(n)}}.$$

(d) Prove that $\mathrm{P}(A_n,\ i.o.) = 1$.

(e) Argue that the the following claim holds:

$$\mathrm{P}\left\{ \sup_{s,t \in (0,1)} \frac{|B(s) - B(t)|}{|s-t|^{1/2}} = \infty \right\} = 1.$$

---

4. [25 pts] Consider the following measure, ν, of dispersion associated with a real-valued random variable, X having cumulative distribution function F_X:

$$\nu(F_X) \;=\; \iint |x_2 - x_1| dF_X(x_2) dF_X(x_1).$$

(a) Suppose you observe n i.i.d. random variables $X_1, \ldots, X_n$, each having cumulative distribution F_X. Consider the natural estimator of $\nu(F_X)$ given by:

$$\widehat{\nu}(X_1, \ldots, X_n) \;=\; \frac{1}{n^2} \sum_{1 \leq i,j \leq n} |X_i - X_j|.$$

Show that $\widehat{\nu}(X_1, \ldots, X_n)$ is biased for $\nu(F_X)$. Then, use a heuristic argument to explain why it makes sense for this estimator to be biased.

(b) Consider instead the following estimator:

$$\widetilde{\nu}(X_1, \ldots, X_n) \;=\; \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} |X_i - X_j|.$$

Show that $\widetilde{\nu}(X_1, \ldots, X_n)$ is unbiased for $\nu(F_X)$.

(c) Prove that $\sqrt{n}(\widetilde{\nu}(X_1, \ldots, X_n) - \widehat{\nu}(X_1, \ldots, X_n)) = o_p(1)$.

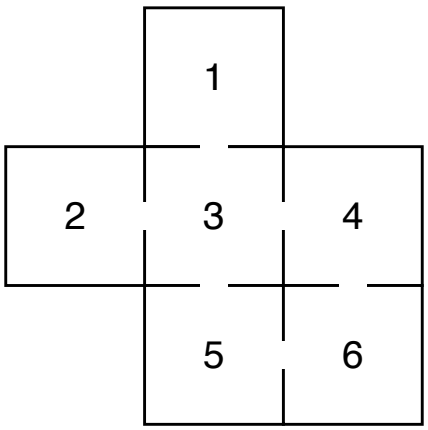
(d) Determine the numeric value of $\nu(F_X)$ if X follows a standard normal distribution. Then, for this specific setting, use Monte Carlo methods (e.g., in R) in order to:

- Confirm your calculation of $\nu(F_X)$.
- Illustrate the bias of $\widehat{\nu}(X_1, \ldots, X_n)$ and $\widetilde{\nu}(X_1, \ldots, X_n)$ as functions of n.
- Illustrate the relative efficiency of $\widehat{\nu}(X_1, \ldots, X_n)$ relative to $\widetilde{\nu}(X_1, \ldots, X_n)$ as a function of n.
- Illustrate the claim made in part (c).

For this problem, please include your R code as a clearly labeled supplementary appendix to your exam submission.

---

5. [25 pts] A group of investigators place a rat in a six-room setup (diagram shown below). Upon entering a room, the rat selects one of the available doors in leading out of the room at random (all equally likely), so that each of the six states is transient.



(a) Write the transition matrix associated with this Markov chain.

(b) Determine whether this Markov chain is regular (justify your answer).

(c) Determine whether this Markov chain is ergodic (justify your answer).

(d) Determine the expected number of room switches until the rat reaches room 5 for the first time, if starting from room 1.

(e) The investigators decide to make a career change and develop an escape game for humans inspired by the layout of the six-room setup above, where again room 1 is the entry point. However, there is a secret exit in room 6 that reveals itself according to the following rule:

- Each room has a point value equal to its room number (e.g., room 4 has a point value of 4).
- As the individual completes the k^{th} room switch, their total point value P_k is given by the sum of the point values of all rooms they have previously left (including prior re-entries to those rooms). As an example, an individual taking the path $1 \to 3 \to 2 \to 3 \to 4$ would accumulate a point value of $P_4 = 1 + 3 + 2 + 3 = 9$ upon entering room 4.
- The secret exit reveals itself if the individual has just entered room 6 *and* their total point value upon entry is a multiple of three.

The investigators, still acclimating to their new career choice, absentmindedly forget to inform an individual of the above escape room rule. Without a clear strategy, an individual behaves analogously to the rat, selecting one of the available doors at random (again, equally likely) each time they enter a room. Assume the individual spends exactly ten seconds in a room before making each room switch and that the transition occurs instantaneously. Use a Monte Carlo method (e.g., in R) to determine:

- The expected total time until escape.
- The median total time until escape.
- The 97^{th} percentile of total time to escape.

For this problem, please include your R code as a clearly labeled supplementary appendix to your exam submission.

(f) Continuing from the scenario of part (e), suppose the individual is in room 6 after K room switches and is starting to feel a little nervous when the investigators—realizing their omission of key information—use the overhead speaker to explain the escape rule and share their current point value, P_K. Describe a strategy that the individual could use to guarantee successful escape in exactly two additional room switches.

---

6. [25 pts] Consider the Nadaraya-Watson kernel smoother (NWKS) at $\boldsymbol{x}_0$, given by:

$$\widehat{f}(\boldsymbol{x}_0) = \frac{\sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i) y_i}{\sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)},$$

where the kernel function is given by:

$$K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i) = D\left(\frac{\sqrt{(\boldsymbol{x} - \boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0)}}{\lambda}\right).$$

(a) Identify the function $D(\cdot)$ that renders the Gaussian kernel.

(b) Show that $\widehat{f}(x_0)$ **is** differentiable when $K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)$ is the Gaussian kernel.

(c) Show that $\widehat{f}(x_0)$ **is not** differentiable when $K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)$ is the Epanechnikov kernel, given by:

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

(d) Show that $\widehat{f}(x_0)$ **is** differentiable when $K_\lambda(x_0, x_i)$ is the triweight kernel, given by:

$$D(t) = \begin{cases} (1 - t^2)^3 & |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

(e) Give two reasons why the triweight kernel may be preferable to either the Gaussian or Epanechnikov kernels.

(f) Partition the predictor variables into two groups as follows: $[\boldsymbol{x}, \boldsymbol{z}]$. An *interaction* occurs when there exists some $\boldsymbol{x} \neq \boldsymbol{x}'$ and $\boldsymbol{z} \neq \boldsymbol{z}'$ such that $\widehat{f}(\boldsymbol{x}, \boldsymbol{z}) - \widehat{f}(\boldsymbol{x}', \boldsymbol{z}) \neq \widehat{f}(\boldsymbol{x}, \boldsymbol{z}') - \widehat{f}(\boldsymbol{x}', \boldsymbol{z}')$. Argue that the NWKS is sufficiently flexible to model any such interaction.
