Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(Take-home portion)

May 26–27, 2022

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Thursday, May 26 at 9:00am, and will be due on Friday, May 27 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 two problems (Problems 7 and 8). 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. Simon Vandekar.

---

7. [30 pts] Respiratory syncytial virus (RSV) is a contagious virus that causes infections of the respiratory tract. The burden of RSV manifests itself primarily in children, among whom RSV-associated acute respiratory illness can sometimes be quite severe. Premature birth, younger age, and prior smoke exposure have been established as predictors of hospitalization for RSV. Even among hospitalized children, there are additional layers of outcome severity. One way to characterize severity is the need for major medical intervention (MMI, defined as at least one of the following while hospitalized: supplemental oxygen requirement, admission to the intensive care unit, and use of mechanical ventilation).

To better understand the epidemiology of RSV, a study was conducted among children under two years of age hospitalized for RSV-associated acute respiratory infection. The data set `rsv.csv` has been provided to you by e-mail, and the codebook can be found on the following page. You should include any software code you generate in your response to these problems in a clearly labeled .R file that you turn in with your exam.

Below, you are given three highly open-ended investigative problems. What that means is that there is no single correct answer. Here are some instructions on how to proceed, along with some advice:

(I) Only respond to the questions being asked.

(II) In your response to each question, you should summarize the approach you took in a way that your findings could be closely reproduced by an independent but statistically savvy investigator.

(III) The phrase "investigate the association," although open-ended, should be taken to mean that you are expected to produce both descriptive and inferential findings.

(IV) Similarly, the phrase "develop a prediction model" carries with it the implication that you should properly assess and report the predictive capacity of your model.

(V) If your analysis relies on specific assumptions, you should specifically state what the assumptions are, briefly defend them, and—where possible—assess possible violations to those assumptions.

(VI) Do not present the results of every single figure you generate and every single analysis you try. Be judicious in your presentation.

(VII) Remember the guiding principle that an analysis should be a simple as possible, but not simpler.

(a) Investigate the association between age and length of hospital stay.

(b) Establish a range of typical values for length of hospital stay for children with the following covariate profile:
- Eight months of age.
- Male.
- No wheezing.

(c) Develop a prediction model for major medical intervention based on the following variables:
- Age.
- Sex.
- Birth weight.
- Prior smoke exposure.
- Atopy.
- Family history of asthma.
- Respiratory rate.
- Wheezing.
- Cyanosis.

CODEBOOK FOR PROBLEM 7

`id`	unique study participant identifer
`age`	age at presentation (months)
`sex`	sex at birth (0=female; 1=male)
`bwt`	birth weight (kg)
`smoke`	prior smoke exposure (0=no; 1=yes)
`atopy`	child atopy (0=no; 1=yes)
`fhxasth`	family history of asthma (0=no; 1=yes)
`resp`	respiratory rate at presentation (breaths per minute)
`wheeze`	wheezing at presentation (0=none; 1=mild; 2=severe)
`cyan`	cyanosis at presentation (0=none; 1=mild; 2=severe)
`los`	length of hospital stay (days)
`mmi`	major medical intervention (composite outcome)

8. [20 pts] You are working with an investigative team that focuses on the use of magnetic resonance imaging (MRI) to understand variation in properties of the spinal cord. One goal in particular is to estimate specific quantiles of spinal cord volume (mm^3) in the population. As is often the case, there are many approaches that could be used to accomplish this goal. Consider the following two estimation approaches for estimating the p^{th} quantile of a distribution based on a total of n independently sampled subjects:

   - An estimate based on the presumption that spinal cord volume follows a normal distribution with mean μ and variance σ^2: $\widehat{Q}(p) = \widehat{\mu} + \Phi^{-1}(p) \times \widehat{\sigma}$, where $\widehat{\mu} = \overline{X}_n$ denotes the sample mean, $\widehat{\sigma}$ denotes the sample standard deviation, and $\Phi^{-1}(p)$ denotes the p^{th} quantile of the standard normal distribution.
   - An estimate that does not make specific assumptions about the underlying distribution of spinal cord volume, based on the inverse of the empirical cumulative distribution function. This can be obtained in R through the command `quantile(..., type = 1)`. **Specifying the 'type' argument is important, as the default approach is different from the one described here**.

   In this problem, your task is to compare these two estimation approaches using Monte Carlo (simulation) based techniques. Your code for all problems must be turned in as clearly labeled supplementary .R file.

   (a) Write simulation code in which you assess the finite-sample bias and variance of each of the two estimators described above when the assumption of normality is satisfied. Your simulation code should:
      - Generate spinal cord volume under a normal distribution: $X \sim \mathcal{N}(\mu = 80, \sigma^2 = 100)$.
      - Apply each of the two estimation techniques to estimate both the 50^{th} and 97.5^{th} percentiles of the spinal cord volume.
      - Vary the sample size across the following range: $n = 20, 40, 80, 160, 320, 640, 1280$.
      - Generate 10,000 replicates under each of the seven sample sizes.
      - Be fully reproducible and include comments that make clear the goal of key steps in your code.

   (b) For each of the two quantiles considered, present plots to compare the percentage bias and the variance of each estimator as functions of sample size (the figure will look nicer if you present the x-axis on a log-scale—e.g., with the option `log="x"`). Briefly summarize your findings.

   (c) Very briefly explaining your response, which of the two estimation approaches appears to be better in this scenario?

   (d) Repeat parts (a)-(c), instead re-defining spinal cord volume to follow a shifted Gamma distribution: $X = 60 + Z$, where $Z \sim \text{Gamma}(k = 4, \theta = 5)$. *Note*: $\text{E}[X] = 80$ and $\text{Var}[X] = 100$, just as in part (a).

   (e) Based on the overall results of this simulation study, briefly summarize your conclusions regarding which estimator is likely best if you don't know the underlying data generating mechanism *a priori*.
