Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(Take-home portion)

May 22–23, 2025

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Thursday, May 22 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 two problems (Problems 6 and 7). 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.
- 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.

---

6. [60 pts] An observational study was conducted to compare HIV viral load between patients receiving two different antiretroviral drugs (dolutegravir and efavirenz) six months after treatment initiation. The study included a group of patients prescribed dolutegravir and another (equally-sized) group of patients prescribed efavirenz. You were supplied with the file `viral-load.csv` for this exam; a codebook is presented on the following page. Please note that for this exam, you are expected to use statistical software (R, in particular). Please supply a knitted .html file together with a supplementary .Rmd or .Qmd file that is organized and annotated well (a .doc/.docx/.pdf file together with an organized and annotated .R file also suffices). To be clear, we need to be able to see your solutions without having to run your code, but we also need to be able to see the code that leads to your solutions.

In each of parts (a) through (g), you are asked to perform an analysis meeting a specific description. Where needed, you should state and defend any choices you make about the analysis (a small number of specifications are left open for you to interpret as the analyst). For each analysis, clearly state the quantity you're estimating (maximal credit will require interpreting parameters on a scientifically relevant scale); provide the information that would be expected from a scientific collaborator (i.e., point and interval estimates and—if applicable—a measure of strength of evidence).

(a) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying a two-sample t-test to the log-transformed six-month viral load.

(b) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying a two-sample t-test to the change in log-transformed viral load from baseline to six months.

(c) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying simple linear regression to the log-transformed six-month viral load.

(d) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying simple linear regression to the change in log-transformed viral load from baseline to six months.

(e) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying linear regression to the log-transformed six-month viral load, adjusting for log-transformed baseline viral load.

(f) Perform an analysis in which you compare patients receiving dolutegravir and efavirenz by applying linear regression to the change in log-transformed viral load from baseline to six months, adjusting for log-transformed baseline viral load.

(g) Repeat part (f) under the Bayesian framework in which you place the `brm()` function's default prior on $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$, representing the model's coefficients and root mean squared error of the linear regression model.

(h) Comment on and account for similarities and differences between the analyses you performed in parts (a) through (g). You need not include mathematical justification; heuristic responses are acceptable for this problem.

(i) Briefly discuss the relative advantages and limitations of the approaches employed in each of parts (a) through (g). It is sufficient to name one advantage and one disadvantage of each.

(j) Briefly explain the importance of pre-specifying analysis plans and how this problem helps illustrate this importance.

CODEBOOK FOR PROBLEM 6

`regimen`	Antiretroviral medicaiton (dolutegravir or efavirenz)
`viral_load_baseline`	Baseline viral load
`viral_load_6months`	Six-month viral load

7. [40 pts] Consider the setting of a randomized controlled trial to compare the mean of a continuous outcome, Y, between two arms. Each participant is randomized to receive either $X = 0$ (control) or $X = 1$ (experimental treatment). Enrollment is capped at a maximum of $n = 500$ participants, but you wish to allow for the possibility that the data may provide evidence of a difference between groups prior to completing enrollment. The purpose of this problem is to investigate (by simulation) what happens when you check the data in the interim and treat any result as if it were based on your *only* analysis of the data. Here are a few stipulations about the simulation setup that you should use:

- Each participant is randomized on the basis of a fair coin flip: $X \sim \text{Bernoulli}(p = 0.5)$.
- The outcomes are standard normal in each group: $Y_0 \sim \mathcal{N}(\mu_0 = 0, \sigma_0^2 = 1)$ and $Y_1 \sim \mathcal{N}(\mu_1 = 0, \sigma_1^2 = 1)$.
- Groups are compared using a t-test that assumes equal variances.
- The null hypothesis of no mean difference is rejected upon observing a (two-sided) p-value of $p < 0.05$.

Below are four scenarios describing different ways to monitor the data:

(I) No interim analysis are conducted and all $n = 500$ participants are measured, at which point a single p-value is recorded from a t-test.

(II) One interim analysis is conducted after measurement of $n = 250$ participants; enrollment is halted and the p-value of a t-test is recorded if $p < 0.05$, and otherwise continued until completion (i.e., $n = 500$) at which point the p-value of a t-test is determined from the entirety of the data.

(III) Four interim analyses are conducted after measurement of $n = 100$, $n = 200$, $n = 300$, and $n = 400$ participants; enrollment is halted and the p-value of a t-test is recorded when and if $p < 0.05$ at any interim analysis, and otherwise continued until completion (i.e., $n = 500$) at which point the p-value of a t-test is determined from the entirety of the data.

(IV) Nine interim analyses are conducted after measurement of $n = 50$, $n = 100$, ..., and $n = 450$ participants; enrollment is halted and the p-value of a t-test is recorded when and if $p < 0.05$ at any interim analysis, and otherwise continued until completion (i.e., $n = 500$) at which point the p-value of a t-test is determined from the entirety of the data.

Keep in mind that the purpose of this problem is to take these p-values at face value; each study replicate has only one p-value that is either determined in the interim (if significant) or at the end (if no interim analyses achieved significance). You are not "correcting" p-values for the having conducted interim analyses (there are ways to do this, but you have not learned them and are not expected to know or implement them).

(a) Simulate data under Scenario (I) with $M = 100{,}000$ replicates. Create/report the following:
- A histogram of the t-statistics across replicates (use bin sizes of 0.05).
- A histogram of the p-values across replicates (use bin sizes of 0.005).
- A table indicating the relative frequency of each sample size across replicates.
- The proportion of times across replicates in which the threshold $p < 0.05$ is achieved.

You must include R code as an appendix (.docx/.R or .Rmd/.html type formats are acceptable).

(b) Repeat part (a) for Scenario (II).

(c) Repeat part (a) for Scenario (III).

(d) Repeat part (a) for Scenario (IV).

(e) Briefly comment on your findings and account for the differences you see (your arguments can be heuristic rather than mathematical).

(f) Imagine now that the p-values of Scenarios (II) through (IV) were properly corrected for interim analysis. Comment on the ethical advantages of interim monitoring over no interim monitoring.

(g) Despite your answer to part (f), what might be a possible *disadvantage* of interim monitoring, even when p-values are properly corrected?
