Biostatistics MS Comprehensive Exam: Applied

June 3 - 4, 2021

Instructions: Please adhere to the following guidelines:

- The MS Applied Comprehensive Exam will be administered on Thursday, June 3 at 9:00am (central time); you have until Friday, June 4 at 5:00pm (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 problems of varying length and difficulty. Note that not all problems 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.
- Problem 5 involves the data set `er.csv` that was provided to you at the time of exam distribution. It is expected that you will use statistical software in producing a response to this question. Note also that you are welcome to use statistical software to answer other questions where applicable and appropriate. Wherever you do use statistical software, please attach your code in a clearly labeled appendix.
- Please email any clarifying questions to:
  Dr. Simon Vandekar (simon.vandekar@vumc.org), and
  Dr. Andrew Spieker (andrew.spieker@vumc.org).

---

1. $\boxed{\textbf{30 pts}}$ The following is a set of six TRUE or FALSE questions. For each sub-question, indicate which option you choose (TRUE or FALSE). Provide *very* brief (one-three sentence) justifications for your responses.

(a) **TRUE** or **FALSE** Suppose you seek to estimate an unknown population mean, μ. Then, the proportion of two-sided 95% confidence intervals based on normal theory derived from independent, equally-sized samples from the population that contain μ must always be 0.95.

(b) **TRUE** or **FALSE** A 95% Bayesian credible interval for an unknown parameter θ will necessarily have the property of 95% coverage when the likelihood used to derive it is correctly specified.

(c) **TRUE** or **FALSE** Even in sufficiently large samples, an α-level two-sample t-test of equality in means requires the outcomes to be normally distributed in order to achieve a type-I error rate of approximately α.

(d) **TRUE** or **FALSE** Consider estimation of the “slope” parameter β_1 in the context of simple linear regression where the mean model is correctly specified. In this setting, the ordinary least squares estimator $\widehat{\beta}_1$ must have minimal variance among all possible estimators of β_1 if and only if the errors are homoscedastic.

(e) **TRUE** or **FALSE** In a regression model, you must apply logarithmic transformations to any variables that show signs of skewness in order for the model to provide meaningful results.

(f) **TRUE** or **FALSE** Consider estimation of a population-average (marginal) association between age and mean systolic blood pressure. A study that includes $J = 5$ repeated measures on each of $N = 100$ independently sampled subjects will always contain the same degree of information about this association as a study of $N = 500$ subjects each measured once.

---

2. $\boxed{\textbf{20 pts}}$ Data were collected on $n = 25$ patients between the ages of 20 and 90 undergoing some surgical procedure related to their spine. Each patient had the posterior tibial nerve of each leg electrically stimulated at a standard location near the medial malleolus (ankle), and the cortical somatosensory evoked potential (SEP) was measured from electrodes placed at standard locations on the head. The table below provides descriptive statistics on `n35R` and `n35L` — the time (in milliseconds) to detection of the first negative SEP following stimulation of the right and left posterior tibial nerve, respectively. You may assume for the purposes of this problem that both `n35R` and `n35L` are approximately normally distributed.

Variable	Mean (SD)	Median (1st Qtl., 3rd Qtl.)	(Min., Max.)
`n35R`	35.1 (3.70)	34.8 (33.0, 37.8)	(23.4, 46.8)
`n35L`	35.4 (3.68)	34.8 (33.0, 37.2)	(25.2, 47.4)

(a) Construct a (two-sided) 95% confidence interval for the mean value of `n35R` in this population. You may of course use any statistical software of your choosing to determine, e.g., the quantile of a particular reference distribution.

(b) A collaborator interprets the confidence interval computed in part (a) as a range of values that encompasses approximately 95% of the `n35R` values in these data. Briefly discuss the degree to which you agree or disagree with your collaborator's interpretation.

(c) Construct a (two-sided) 95% confidence interval for the mean value of `n35L` in this population. You may of course use any statistical software of your choosing to determine, e.g., the quantile of a particular reference distribution.

(d) A collaborator states that since the confidence intervals computed in parts (a) and (c) overlap, the study provides evidence that the means of `n35R` and `n35L` are not different. Identify at least two major flaws in your collaborator's interpretation.

(e) Is it possible to use these data to determine a point estimate and/or construct a 95% confidence interval for the mean difference between `n35L` and `n35R` in this population? Do what you can, but if you cannot do something without further information, say so and briefly justify your response.

---

3. 20 pts You are working with a team that seeks to determine if C-reactive protein (CRP) levels exceeding 3 mg/dL are indicative of severe narrowing of the coronary arteries. The table below summarizes results from a case-control study of fifty subjects with severe narrowing and fifty subjects without severe narrowing.

	Severe narrowing	No severe narrowing	TOTAL
CRP > 3 mg/dL	33	8	41
CRP ≤ 3 mg/dL	17	42	59
TOTAL	50	50	100

In each of parts (a)-(d), you are provided with a statement of a scientific hypothesis. In each problem, your first task is to determine whether the study data can be used to evaluate the hypothesis. Then:

- If you determine that the answer is "no," provide a brief justification (one to two sentences).
- If you determine that the answer is "yes," briefly state the method you will use to test the hypothesis. Perform the analysis, and briefly summarize the results.

You will find it helpful to represent the data in the above table in a statistical software program of your choosing in order to perform the analyses you want to perform (you are not expected to complete this problem "by hand").

(a) **Hypothesis A**: The proportion with a positive CRP test (> 3 mg/dL) is greater than 0.6 among those with severe narrowing of the coronary arteries.

(b) **Hypothesis B**: The proportion with severe narrowing of the coronary arteries among those with a positive CRP test (> 3 mg/dL) is greater than 0.6.

(c) **Hypothesis C**: The difference in proportions of individuals with severe narrowing of the coronary arteries between those with a positive CRP test (> mg/dL) and those with a negative CRP test (≤ 3 mg/dL) is different from 0.2.

(d) **Hypothesis D**: The odds ratio that compares the odds of severe narrowing of the coronary arteries between those with a positive CRP test (> 3 mg/dL) and those with a negative CRP test (≤ 3 mg/dL) is different from one.

4. 15 pts Imagine that you are in a meeting to devise a statistical analysis plan for an upcoming cohort study of $N = 3{,}000$ subjects (to be sampled independently) in which the primary aim is to compare some continuous outcome, Y, between two groups defined by an exposure status. The exposure occurs in about one-third of individuals in the population. Is is suspected *a priori* that the outcome variance in the exposed group will be substantively larger in the exposed group as compared to the unexposed group (due to, e.g., heterogeneous response to the exposure). Despite this, your collaborator suggests Student's t-test (i.e., the t-test that assumes equal variances between groups) in order test the hypothesis of a nonzero mean difference, as this has been standard in the literature. As the group's biostatistician, you perform a simulation to illustrate the possible pitfalls of this approach. The R function to conduct the simulation is shown below (mind the line numbers on the left-hand side, as questions (a)-(e) below refer to them):

```
1     simulation <- function(N, p.exp, sigma0, k, nsim, seed)
2     {
3       set.seed(seed)
4       results <- matrix(0, ncol = 1, nrow = nsim)
5       for (j in 1:nsim)
6       {
7         x <- rbinom(n = N, size = 1, prob = p.exp)
8         N1 <- sum(x)
9         N0 <- N - sum(x)
10        y0 <- rnorm(n = N0, mean = 0, sd = sigma0)
11        y1 <- rnorm(n = N1, mean = 0, sd = sqrt(k * sigma0^2))
12        p <- t.test(x = y0, y = y1, var.eq = TRUE)$p.value
13        results[j, 1] <- as.numeric(p < 0.05)
14      }
15      return(colMeans(results))
16    }
```

You then run the simulation under certain parameters (shown below, along with the result).

```
> simulation(N=3000, p.exp=1/3,  sigma0=10, k=2, nsim=50000, seed=2021)
[1] 0.07926
```

(a) Briefly explain the major purpose of the `seed` argument (referred to on Line 3).

(b) Briefly explain what Lines 7-9 of the simulation code are doing.

(c) Briefly explain the role of the `k` argument (referred to on Line 11).

(d) In the language of hypothesis testing, what does the number "0.05" refer to on Line 13?

(e) In the language of hypothesis testing, what is the value returned by the function (Line 15)?

(f) Briefly summarize the main result of the simulation study and highlight the importance of its implications.

(g) What alternative test would you suggest to answer your collaborator's scientific question?

(h) Illustrate the advantage of the approach you identified in part (g) by re-running the simulation study under the same setup, but making a single key modification to the function. Specifically state the change you've made, and briefly summarize your findings.

---

5. 50 pts Suppose you are working with a team that is interested in understanding the link between recreational drug use and subsequent visits to the emergency room (ER). They collect data on $N = 400$ independently sampled individuals and survey them on their history of drug use. If a subject had an ER visit within five years of study enrollment, it was noted (data set: `er.csv`). The variables in the data set are as follows:

`agegrp`	age category, in years (0 = 12-17; 1 = 18-25; 2 = 26-34; 3 = 35-49; 4 = 50 or older)
`tob`	tobacco use (0 = no; 1 = yes)
`alc`	alcohol use (0 = no; 1 = yes)
`mrj`	marijuana use (0 = no; 1 = yes)
`illeg`	other illegal drug use (0 = no; 1 = yes)
`er`	ER visit within five years of enrollment (0 = no; 1 = yes)

(a) Briefly summarize the distribution of variables in the data set. You may include tables and figures if you choose, but your response should not exceed one double-spaced page (tables and figures included).

(b) Perform a principled analysis in which you investigate the association between recreational drug use and ER visits. When writing your response:

- Remember that this is an open-ended question in which there are multiple defensible ways to respond. Please select and implement only one approach (that may or may not involve results from more than one model).
- Describe the method you are implementing such that it could essentially be replicated without having to look at your code.
- Clearly state and briefly defend any choices you make in your analysis.
- Present and interpret your results using suitable measures of association that could be understandable to a scientific collaborator.
- You may include tables and/or figures if you believe they will aid the presentation of your response.
- Concede at least one limitation of the approach you choose.
- Your response should be no longer than two double-spaced pages (*including* tables and figures, but *excluding* code.
- Please present code only as an appendix.

(c) Perform a principled analysis in which you evaluate the overall out-of-sample predictive ability of recreational drug use as predictors for ER visits. When writing your response:

- Remember that this is an open-ended question in which there are multiple defensible ways to respond. Please select and implement only one approach (that may or may not involve results from more than one model).
- Describe the method you are implementing such that it could essentially be replicated without having to look at your code.
- Clearly state and briefly defend any choices you make in your analysis.
- Present and interpret your results using suitable measures of predictive ability that could be understandable to a scientific collaborator.
- You may include tables and/or figures if you believe they will aid the presentation of your response.
- Concede at least one limitation of the approach that you choose.
- Your response should be no longer than two double-spaced pages (*including* tables and figures, but *excluding* code.
- Present code only as an appendix.

---

6. $\boxed{\textbf{15 pts}}$ This problem is about understanding study design and model misspecification as sources of variability in simple linear regression. Let $i = 1, \ldots, 4n$ index study subjects, and consider four overall data generating mechanisms based on two study design scenarios and two outcome generation scenarios.

<u>Study design (under each study design, note that $\mathrm{E}[X] = 0$ and $\mathrm{Var}[X] = 1$):</u>

i. X is generated randomly, as per an observational study: $X_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

ii. X is fixed, as per an experimental design: $X_i = \begin{cases} -\sqrt{2} & \text{for } i = 1, \ldots, n \\ 0 & \text{for } i = n+1, \ldots, 3n \\ \sqrt{2} & \text{for } i = 3n+1, \ldots, 4n \end{cases}$

<u>Outcome generation mechanisms:</u>

i. $Y_i = X_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 5)$.

ii. $Y_i = X_i^2 + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 5)$.

Use computational (simulation) techniques in order to numerically determine the variance of $\widehat{\beta}_1$, as estimated using ordinary least squares, per the regression model $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$. Indeed, this regression model is correctly specified under outcome generation mechanism (i) above, but *not* under outcome generation mechanism (ii). Within each of the four simulation setups, simulate ten-thousand data sets, each under a total sample size of $4n = 120$, and compute the variance of the slope estimates $(\widehat{\beta}_1)$ across the simulation replicates. Please turn in your code as an appendix, and do not attempt to mathematically derive anything for this problem.

(a) Present a 2×2 table cross-tabulating the values of $\mathrm{Var}[\widehat{\beta}_1]$ according to the study design (random vs. experimental) and outcome generation mechanism (linear vs. quadratic).

(b) Within each study design, compare the variances between the two outcome generation mechanisms and use heuristic arguments to account for any differences or similarities you observe.

(c) Within each outcome generation mechanism, compare the variances between the two study designs and use heuristic arguments to account for any differences or similarities you observe.
