Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(In-class portion)

May 25, 2022

Instructions: Please adhere to the following guidelines:

- This exam begins on Wednesday, May 25 at 9:00am. You will have until 1:00pm to complete it.
- There are six problems of varying length and difficulty. 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, show your work when necessary, and please write legibly.
- This exam is closed-everything and is an *individual effort*. You are, however, permitted the use of a scientific calculator. Vanderbilt University's academic honor code applies.
- Please direct clarifying questions to the exam proctor.

---

1. [20 pts] A randomized controlled trial was conducted to assess whether a single nighttime dose of melatonin improves sleep among adults who previously reported not sleeping well. A total of ninety-six adults were randomized in a blinded fashion to one of two treatment groups (`grp=0`: placebo, and `grp=1`: melatonin). All participants adhered to their assigned treatment for one night and reported their sleep (hours) the following morning. The following page presents histograms of sleep levels (stratified by treatment group) along with Stata output from two versions of the two-sample t-test. Note that certain quantities have been intentionally excluded from the Stata output and replaced with the symbol "`%%%%`."

(a) In plain but statistically precise language, what hypothesis is being tested in this study?

(b) The standard error for the mean difference is excluded from the Stata output in both versions of the t-test. Determine their values.

(c) The t-statistic is excluded from the Stata output in both versions of the t-test. Determine their values.

(d) The p-value for the one-sided "greater" alternative is excluded from the Stata output in both versions of the t-test. Determine their values.

(e) The degrees of freedom associated with the equal-variance t-test is excluded from the Stata output. Determine its value.

(f) Determine the 97.5^{th} percentile of the t-distribution having the degrees of freedom specified in part (e).

(g) Irrespective of any of your responses to parts (a)-(f), which version of the t-test would you have chosen to conduct *a priori*? Very briefly justify your choice.

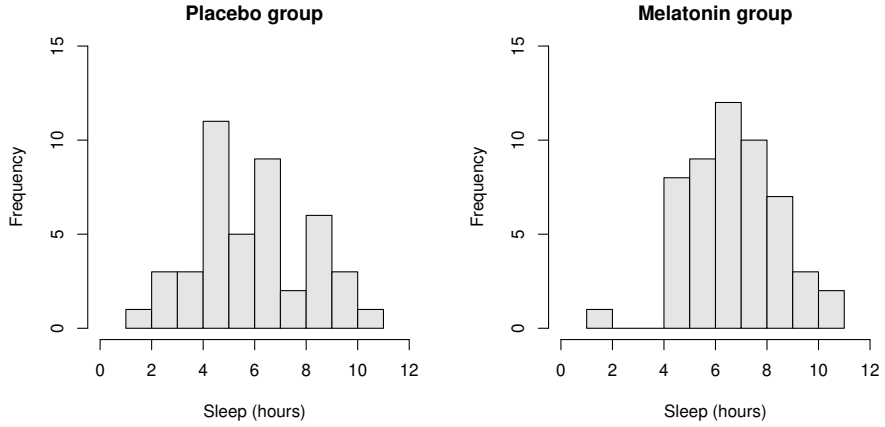
(h) Based on your choice in part (g), provide a concise write-up of the study results. In your response, be certain to provide proper measures of association, precision, and statistical strength of evidence.

(i) An alternative study design was proposed whereby all ninety-six participants reported sleep levels from the night prior to receiving their assigned treatments (placebo or melatonin). An analysis was then performed to determine if the mean *change* in sleep from baseline to treatment differed between treatment groups. Briefly describe the most likely advantage associated with this alternative design.

(j) Propose an alternative analytic strategy (separate from that proposed in part (i)) that would likely produce a similar advantage to that of part (i).

Supplementary Material for Problem 1

Histograms of sleep by treatment group:



Stata output for two different versions of the two-sample t-test:

```
. ttest sleep, by(grp)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      44    5.938007    .3109719    2.062754    5.310872    6.565141
       1 |      52    6.708712    .2405714    1.734785    6.225745     7.19168
---------+--------------------------------------------------------------------
    diff |            -.770706     %%%%                -1.540149   -.0012626
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =     %%%%
Ho: diff = 0                                     degrees of freedom =     %%%%

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0248         Pr(|T| > |t|) = 0.0496          Pr(T > t) =   %%%%


. ttest sleep, by(grp) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      44    5.938007    .3109719    2.062754    5.310872    6.565141
       1 |      52    6.708712    .2405714    1.734785    6.225745     7.19168
---------+--------------------------------------------------------------------
    diff |            -.770706     %%%%                -1.552504    .0110919
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =     %%%%
Ho: diff = 0                     Satterthwaite's degrees of freedom =  84.3865

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0266         Pr(|T| > |t|) = 0.0533          Pr(T > t) =   %%%%
```

2. [20 pts] A vaccine can be evaluated in part by the degree to which it produces an antibody response. The hemagglutination inhibition (HAI) titer measures antibody levels by leveraging the natural tendency of blood to arrange irregularly (agglutinate) when exposed to a virus sample. Specifically, serum antibody levels can be quantified by sequentially diluting the serum sample and exposing it to a fixed amount of virus; a participant's HAI titer is defined as the relative concentration of the final dilution in which agglutination does not occur (higher titers are therefore indicative of a greater concentration of antibodies).

A concern is that influenza vaccination may not elicit sufficiently protective antibody responses in older adults due to age-associated wanes in the immune system. A large randomized trial of adults over 65 years old was conducted to compare a series of two high-dose influenza vaccines (spaced four weeks apart) to a series of two standard-dose influenza vaccines (also spaced four weeks apart). HAI titers were measured at the first follow-up (visit 2: four weeks following the first dose, immediately prior to the second dose), and the second follow up (visit 3: four weeks following the second dose). The initial dilution of the serum sample for HAI titer evaluation occurred at a ratio of 1:5 (at which no serum samples displayed agglutination), and each sample was sequentially diluted by a factor of two until agglutination could be visibly detected. When organized in "long format," the data set contains the following variables:

`id`	unique study participant ID
`t`	study visit (2, 3)
`grp`	treatment assignment (0 = standard dose; 1 = high dose)
`logtiter`	log-transformed post-vaccine HAI titer (evaluated at visits 2 and 3)

The investigators propose the following mean model based on log-transformed titer values.

$$\mathrm{E}[\texttt{logtiter}_\texttt{t}|\texttt{grp}] \;=\; \beta_0 + \beta_1 1(\texttt{grp=1}) + \beta_2 1(\texttt{t=3}) + \beta_3 1(\texttt{grp=1}) \times 1(\texttt{t=3}).$$

In parts (b)-(h), you are asked to express certain quantities or hypotheses in terms of the model's coefficients. Specifically for parts (b)-(e), please show the work that leads to your answer.

(a) Briefly describe the primary justification for log-transforming the titer outcomes in this problem.

(b) Express the geometric mean titer (GMT) among high-dose recipients four weeks following the first dose.

(c) Express the GMT ratio that compares two high doses to two standard doses.

(d) Express the GMT ratio that compares two standard doses to a single standard dose.

(e) Express the GMT ratio that compares one high dose to two standard doses.

(f) Suppose you seek to evaluate whether the relative effect of high dose to standard dose changes from the first to the second dose. Express the null hypothesis, H_0.

(g) Suppose you seek to evaluate whether the relative effect of two doses to a single dose differs between the high dose and standard dose groups. Express the null hypothesis, H_0.

(h) Suppose you seek to evaluate whether the high dose series has any overall benefit relative to the standard dose series doses. Express the null hypothesis, H_0.

(i) Name and describe a valid approach that could be used to estimate the parameters of the model and perform all analyses associated with parts (b)-(h). Note that there are many correct responses.

---

3. [15 pts] A study was conducted to evaluate the association between kidney stone history (X) and coronary artery calcification (Y). A total of $N = 270$ study participants were sampled and surveyed regarding their kidney stone history, and their degree of coronary artery calcification was measured by the Agatston score and classified as a three-level variable. The variables collected in this study were as follows:

`ID`	unique study participant ID
`X`	kidney stone history (0=no history; 1=one prior episode; 2=at least two prior episodes)
`Y`	Coronary artery calcification (0=none; 1=mild; 2=severe)

The data from this study were then tabulated; the results are shown in the table below.

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	30	15	45
$X = 1$	18	12	60
$X = 2$	10	8	72

Consider the following three regression models that could be used to evaluate the association between kidney stone history and coronary artery calcification:

$$\text{logit}(\text{P}(Y \geq 1|X = x)) = \alpha_0 + \alpha_1 1(x \geq 1) \qquad \textbf{(MODEL 1)}$$

$$\log\left(\frac{\text{P}(Y = k|X = x)}{\text{P}(Y = 0|X = x)}\right) = \beta_{0k} + \beta_{1k} 1(x = 1) + \beta_{2k} 1(x = 2), \quad k = 1, 2 \qquad \textbf{(MODEL 2)}$$

$$\text{logit}(\text{P}(Y \leq k|X = x)) = \gamma_{0k} - \gamma_1 1(x = 1) - \gamma_2 1(x = 2), \quad k = 0, 1 \qquad \textbf{(MODEL 3)}$$

(a) Briefly explaining (although not proving) your response, which of the three models is/are saturated?

(b) Determine the values of the point estimates that would be produced (e.g., by the appropriate regression-based commands in Stata or R) for each of the following coefficients:

- α_0
- β_{21}
- γ_2

(c) Supposing you didn't actually have the data in hand, briefly discuss the relative advantages and disadvantages of each of the three models. You are permitted to provide your response in bullet-form.

(d) Irrespective of your response to part (c), suppose the investigators had settled on fitting a model in the spirit of **MODEL 1**, but also adjusting for age. Describe two settings that would lead you to support the decision to adjust for age. As part of your response, briefly state your reasons for support in each of the two settings you've identified.

(e) Irrespective of your response to part (c), suppose the investigators had settled on fitting a model in the spirit of **MODEL 3**, but also adjusting for coronary heart disease. Describe a setting that would lead you to be nervous about the decision to adjust for coronary heart disease. As part of your response, briefly state your reason for being nervous in that setting.

(f) Translate the hypothesis $H_0 : \beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 0$ into plain language. What other well-known procedure could be used to test the same null hypothesis? You do not need to do the calculations.

---

4. [20 pts] You are seeking to estimate an unknown population proportion, θ, based on $Y \sim \text{Binomial}(N, \theta)$. You conduct an analysis as a Bayesian by placing a Beta(α, β) prior distribution on θ. The table below describes the data from six separate and independent analyses.

	Prior		Data	
Analysis	α	β	Y	N
(1)	4	6	4	10
(2)	2	2	0	1
(3)	6	4	4	10
(4)	10	10	5	10
(5)	2	2	1	1
(6)	10	10	10	20

You may use the following information freely and without proof in this problem:

- This is a conjugate prior, such that the posterior distribution also follows a Beta distribution with hyperparameters α^* and β^* that depend upon the prior parameters and the data.
- The Beta(α, β) distribution has mean $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.

You are not expected to derive the posterior density for this problem. Your arguments in parts (a)-(d) should be concise and heuristic.

(a) Which of the statements below is true? Briefly explain using a heuristic argument (no math).

(i) The posterior mean for analysis (1) is higher than that of analysis (3).
(ii) The posterior mean for analysis (3) is higher than that of analysis (1).
(iii) The posterior means for analyses (1) and (3) are equal.

(b) Which of the statements below is true? Briefly explain using a heuristic argument (no math).

(i) The posterior mean for analysis (6) is higher than that of analysis (4).
(ii) The posterior mean for analysis (4) is higher than that of analysis (6).
(iii) The posterior means for analyses (4) and (6) are equal.

(c) Which of the statements below is true? Briefly explain using a heuristic argument (no math).

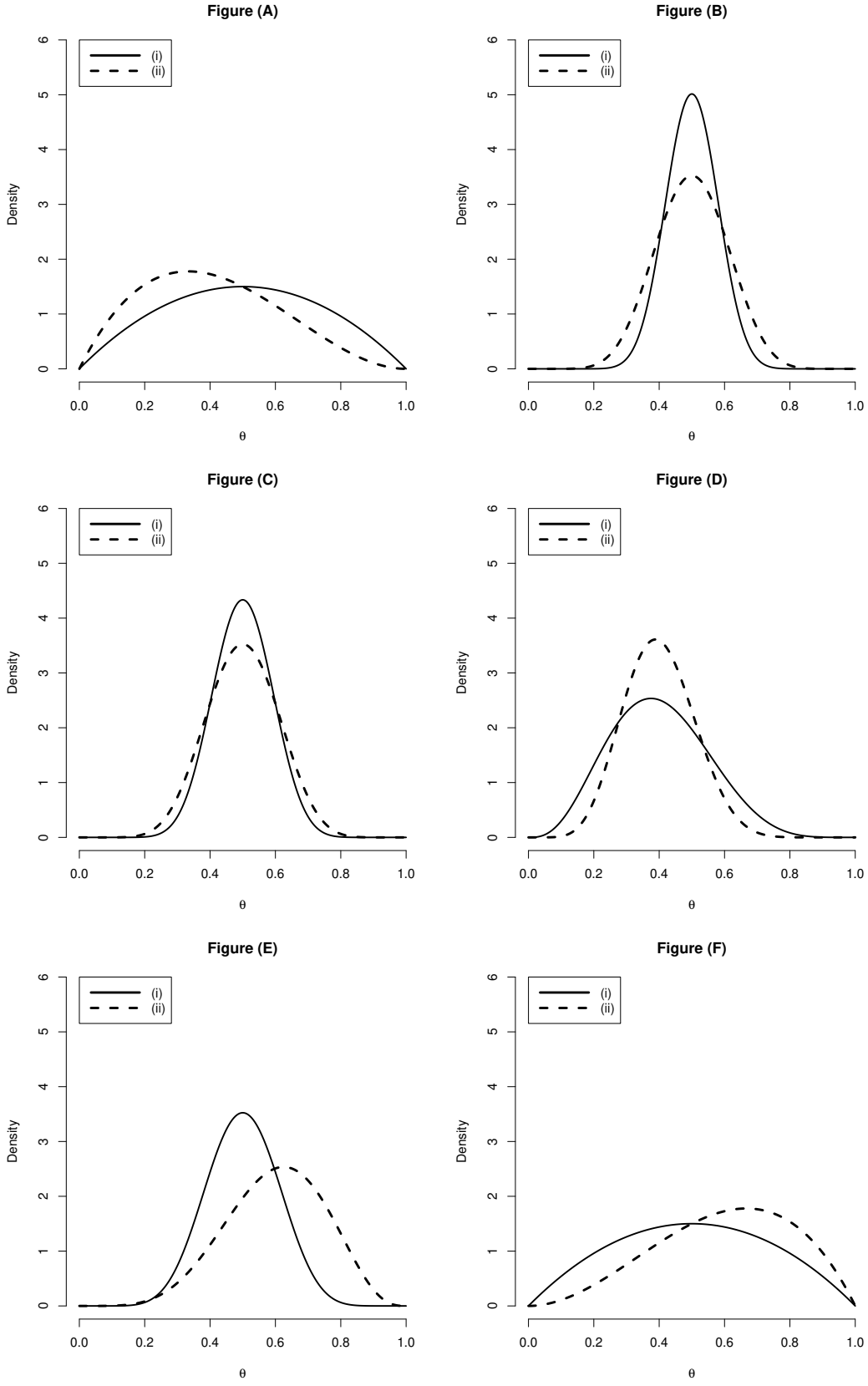
(i) The 95% quantile-based credible interval for analysis (2) is narrower than that of analysis (5).
(ii) The 95% quantile-based credible interval for analysis (2) is wider than that of analysis (5).
(iii) The 95% quantile-based credible intervals for analyses (2) and (5) have the same width.

(d) Which of the statements below is true? Briefly explain using a heuristic argument (no math).

(i) The 95% quantile-based credible interval for analysis (4) is narrower than that of analysis (6).
(ii) The 95% quantile-based credible interval for analysis (4) is wider than that of analysis (6).
(iii) The 95% quantile-based credible intervals for analyses (4) and (6) have the same width.

(e) The following page displays six figures—figures (A) through (F)—each displaying both the prior and the posterior densities for one of the six analyses. Match analyses (1) through (6) to figures (A) through (F). In your response, further match the curves in that figure—i.e., solid (i), and dashed (ii)—to the prior and posterior densities (they may match differently from figure to figure).

Supplementary Material for Problem 4



5. [15 pts] Code is shown below for a simulation study that compares the performance of two methods to form 95% confidence intervals for an unknown mean in a particular setting. You may find it helpful to know that a Gamma(s, r) distribution has skewness given by $2/\sqrt{s}$.

```
1    simulation <- function(N, s, r, nboot, nsim, seed) {
2      set.seed(seed)
3      lowerCI.miss <- upperCI.miss <- matrix(0, ncol = 2, nrow = nsim)
4      for (j in 1:nsim) {
5         x <- rgamma(N, shape = s, rate = r)
6         CI.x.1 <- c(mean(x) - qnorm(0.975)*sd(x)/sqrt(N), mean(x) + qnorm(0.975)*sd(x)/sqrt(N))
7         if (CI.x.1[1] > s/r) {lowerCI.miss[j,1] <- 1}
8         if (CI.x.1[2] < s/r) {upperCI.miss[j,1] <- 1}
9         samp <- matrix(sample(1:N, size = N*nboot, replace = TRUE), nrow = N, ncol = nboot)
10        bootstrap.x <- colMeans(matrix(x[samp], nrow = N, ncol = nboot))
11        CI.x.2 <- quantile(bootstrap.x, probs = c(0.025, 0.975))
12        if (CI.x.2[1] > s/r) {lowerCI.miss[j,2] <- 1}
13        if (CI.x.2[2] < s/r) {upperCI.miss[j,2] <- 1}
14     }
15     result <- data.frame(cbind(colMeans(lowerCI.miss), colMeans(upperCI.miss)))
16     row.names(result) <- c("METHOD.1", "METHOD.2")
17     names(result) <- c("LOW.CI.TOO.HIGH", "HIGH.CI.TOO.LOW")
18     return(result)
19   }
```

The results of this simulation are shown below under the specified parameters:

```
> simulation(N = 200, s = 1, r = 1, nboot = 10000, nsim = 5000, seed = 2022)

            LOW.CI.TOO.HIGH    HIGH.CI.TOO.LOW
METHOD.1             0.0188             0.0424
METHOD.2             0.0220             0.0374
```

(a) Briefly explain the major purpose of the `seed` argument (referred to on Line 2).

(b) Briefly describe the practical meaning of the quantity `s/r` (referred to on Lines 7, 8, 12, and 13).

(c) Briefly describe the theoretical justification associated with each of the two methods being compared in this simulation study (you do not need to prove anything).

(d) Briefly describe the importance of specifying a value for `nboot` that is sufficiently high.

(e) Briefly describe the importance of specifying a value for `nsim` that is sufficiently high.

(f) Which of the two methods has better performance under the specified parameters of the simulation study? As part of your response, state explicitly how you are evaluating performance.

(g) Briefly explaining your response, describe how you would expect the performance of each method to change if you were to substantially increase the parameter `N`.

(h) Briefly explaining your response, describe how you would expect the performance of each method to change if you were to substantially increase the shape parameter `s`.

(i) Briefly explaining your response, describe how you would expect the performance of each method to change if you were to substantially increase the rate parameter `r`.

---

6. [10 pts] The Michaelis-Menten relationship is a well known model of enzyme kinetics that relates the reaction velocity, V, to the substrate concentration, S, as follows:

$$V = \frac{v_{\max} S}{k + S}.$$

In this formula, $v_{\max}$ and k are real-valued constants representing the terminal velocity and how rapidly the reaction rises to its maximum rate (respectively). For a particular enzyme, the values of $v_{\max}$ and k are not known exactly but can be estimated in an experimental context. To that end, suppose you conduct an experiment in which you observe a collection of independent reaction velocities for several fixed substrate concentrations. Interestingly, despite the nonlinear relationship between S and V, it is possible to estimate $v_{\max}$ and k using simple linear regression as an intermediate step based on transformations of S and V. We will focus in this problem on one such clever approach.

(a) Show that the Michaelis-Menten model implies the following relationship:

$$\frac{1}{V} = \frac{1}{v_{\max}} + \frac{k}{v_{\max}} \times \frac{1}{S}.$$

(b) Consider a simple linear regression model with $Y = 1/V$ as the outcome and $X = 1/S$ as the predictor:

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

From an ordinary least squares fit to this model, you obtain point estimates $\widehat{\beta}_0 = 1$ and $\widehat{\beta}_1 = 5$. Keeping the result of part (a) in mind, obtain point estimates $\widehat{v}_{\max}$ and $\widehat{k}$.

(c) Suppose further that you obtain the following variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ from the ordinary least squares fit:

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.0025 & -0.03375 \\ -0.03375 & 0.6 \end{bmatrix}$$

Create Wald-based symmetric 95% confidence intervals for $v_{\max}$ and k based on the multivariate delta method (you are permitted to assume that all assumptions required by the delta method are satisfied without stating or justifying them). You may freely use the fact that $z_{0.975} \approx 1.96$.

(d) Based on the information supplied, is it possible to use another approach to easily obtain a (possibly asymmetric) 95% confidence interval for either $v_{\max}$ or k? Where it is possible to do so easily, do so—and briefly explain why you can.
