Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(In-class portion)

May 21, 2025

Instructions: Please adhere to the following guidelines:

- This exam begins on Wednesday, May 21 at 9:00am. You will have until 1:00pm to complete it.
- There are five problems of varying length and difficulty. Note that not all problems and sub-problems 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] Your colleague is the principal investigator of a pilot randomized trial to compare two exercise regimens for adults with back pain. A total of $N = 16$ participants were randomized to one of two daily regimens (A or B). Participants self-reported their pain on a subjective scale of 1 to 10 (lowest to highest) 30 days following randomization. The data are shown in the table below; group-specific sample means and variances are provided. For your reference, the approximate 95$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles of the t-distribution having between six and sixteen degrees of freedom are reported (you may not need all of this information).

Regimen A	2	2	3	3	4	5	7	8	$\overline{x}_A = 4.2500$	$\overline{s}^2_A = 5.0714$
Regimen B	3	4	5	7	8	9	9	10	$\overline{x}_B = 6.8750$	$\overline{s}^2_B = 6.6964$

df	6	7	8	9	10	11	12	13	14	15	16
$t_{0.95,\mathbf{df}}$	1.94	1.89	1.86	1.83	1.81	1.80	1.78	1.77	1.76	1.75	1.75
$t_{0.975,\mathbf{df}}$	2.45	2.36	2.31	2.26	2.23	2.20	2.18	2.16	2.14	2.13	2.12

(a) Form two-sided 95% confidence intervals for each group-specific mean, μ_A and μ_B.

(b) The following page presents the results of a two-sample t-test allowing unequal variances between groups (`APPROACH.1`). Based on this output, write a brief (i.e., approximately three-sentence) summary of the study's findings in a way that would be suitable for a scientific collaborator.

(c) Your colleague claims that statistical significance (or lack thereof) of a two-sample t-test can be inferred from merely checking whether confidence intervals for the group-specific means overlap. Briefly discuss the extent to which you agree or disagree with your colleague's claim.

The questions in parts (d)-(h) ask you to consider a variety hypothetical scenarios regarding the study. In supplying your responses, please consider each hypothetical scenario one at a time, completely separately from and independently of the others. Note that many of the questions below are deliberately open-ended, with no single correct answer. Your response to each question should be brief (two-three sentences).

(d) Suppose each participant self-reported their pain on the same scale of 1-10 at baseline. How might you incorporate this information into your analysis of the data? In your response, clearly state what the expected benefit would be.

(e) Suppose your colleague informed you that four of the patients did not perform any of their prescribed exercises over the 30-day intervention period (two from each of Regimens A and B), and therefore wants to exclude them from analysis. Briefly discuss the extent to which that is a good (or not-so-good) idea.

(f) Suppose your colleague received approval to conduct a larger Phase II study under a similar design. You're planning to pre-specify a Bayesian analysis of the mean difference. Describe how you might form a prior for the mean difference in a way that acknowledges the results of this pilot study while honoring the boundaries of the measurement's scale (i.e., 0-10). In your response, you need not worry about the mathematical challenges of obtaining the posterior distribution.

(g) Suppose the pilot study's original protocol had called for a Mann-Whitney test (also known as a Wilcoxon test) to be conducted as a secondary analysis (the results of this test are presented on the following page; `APPROACH.2`). Your colleague questions an apparent contradiction between the analysis with the following statement: "It doesn't make sense to me why the first analysis says there is a difference, but the second analysis says there's none." Offer a response to your colleague to help them better understand an appropriately nuanced interpretation (in your response, consider commenting on the difference between the two tests' target parameters, and be mindful of the appropriate interpretation of a p-value).

(h) Suppose you find out that one participant had enrolled in the study twice, assigned first to Regimen A, and then (following completion of their 30-day pain assessment) re-enrolled and assigned to Regimen B. You bring this to the attention of your colleague, who expressed no concern about any assumption violations associated with your analysis because the patient received "a whole new independent treatment." Supply a brief rebuttal to your colleague's lack of concern.

Supplementary Material for Problem 1

```
## Outcome data
yA <- c(2,2,3,3,4,5,7,8)
yB <- c(3,4,5,7,8,9,9,10)

## Two-sample t-test with unequal variances
APPROACH.1 <- t.test(yA, yB, var.equal = FALSE)

# > print(APPROACH.1)
# Welch Two Sample t-test
#
# data:  yA and yB
# t = -2.1643, df = 13.738, p-value = 0.04856
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -5.23094541 -0.01905459
# sample estimates:
#   mean of x mean of y
# 4.250     6.875

## Wilcoxon test (also known as Mann-Whitney test)
APPROACH.2 <- wilcox.test(yA, yB)

# > print(APPROACH.2)
# Wilcoxon rank sum test with continuity correction
#
# data:  yA and yB
# W = 13, p-value = 0.05031
# alternative hypothesis: true location shift is not equal to 0
```

2. [15 pts] Antibiotics are sometimes prescribed for treatment of acute otitis media (middle ear infection) in children. A cause for concern in the over-prescription of antibiotics is that it can result in resistant bacteria. A pilot study of $N = 16$ children was conducted to understand differences in time to symptoms resolving between individuals receiving antibiotics and not receiving antibiotics. Eight children were randomized to each group (day zero). In the group receiving antibiotics, three children reported symptoms resolving on day four, three reported symptoms resolving on day six, one reported symptoms resolving on day eight, and one reported symptoms resolving on day ten. In the group receiving no antibiotics, one child reported having symptoms resolve on day six, five reported symptoms resolving on day eight, and two children still experienced symptoms on day ten at which point they were censored. Kaplan-Meier curves are depicted for each treatment group on the following page, along with abridged results from a Cox proportional hazards regression fit to the data.

(a) Within each group, estimate the proportion of children still experiencing symptoms at one week.

(b) Within each group, estimate the proportion of children whose symptoms resolve within five days.

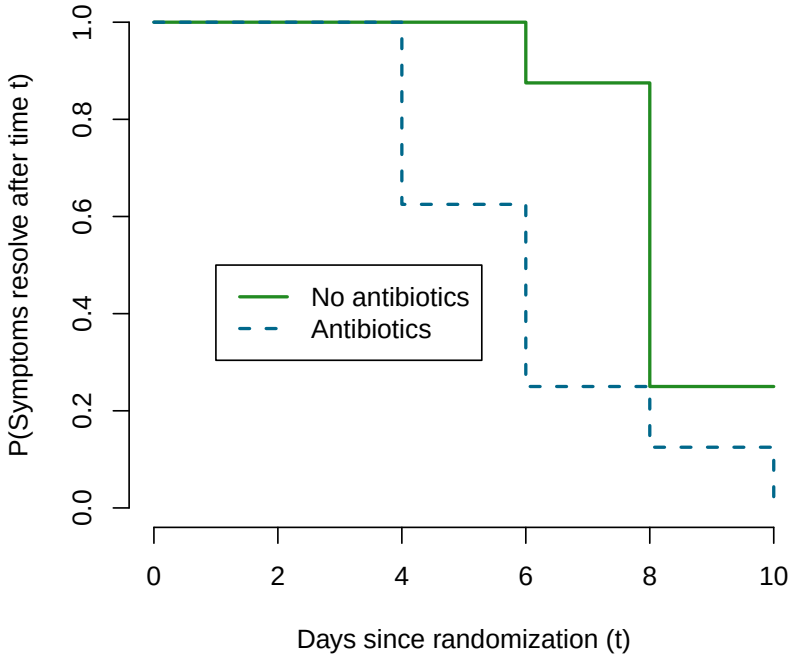
(c) Is it possible to obtain an approximate 95% confidence interval for the proportion of children receiving antibiotics whose symptoms resolve within five days without further information? If so, do so (and state the reasons why your confidence interval may only be approximate); if not, briefly explain why not.

(d) Within each group, estimate the median time to symptoms resolving.

(e) Based on the output from the Cox model, write a brief (i.e., approximately three-sentence) summary of the study's findings in a way that would be suitable for a scientific collaborator. Further, state and describe the most important assumptions that would need to hold in order for you to trust the validity of your conclusions.

(f) Briefly discuss the extent to which the Kaplan-Meier plot provides clinical evidence *against* prescribing antibiotics for acute otitis media in children right away.

Supplementary Material for Problem 2




```
## Survival library
library("survival")

## Data
Xi <- c(rep(0,8), rep(1,8))
Ti <- c(6,8,8,8,8,8,10,10,4,4,4,6,6,6,8,10)
Ci <- 1 - c(0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0)

## Cox Model
zz <- coxph(Surv(Ti,Ci) ~ Xi)

# > summary(zz)
# Call:
# coxph(formula = Surv(Ti, Ci) ~ Xi)
#
#   n= 16, number of events= 14
#
#      coef exp(coef) se(coef)     z Pr(>|z|)
# Xi 1.0539    2.8689   0.5497 1.917   0.0552 .
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#    exp(coef) exp(-coef) lower .95 upper .95
# Xi     2.869     0.3486    0.9768     8.427
```

3. [25 pts] Tuberculosis (TB) poses a significant global health challenge, with over ten-million new cases and more than one-million deaths reported in 2023. Effective treatment exists, but adverse drug reactions (ADRs) can interfere with real-world effectiveness. It has been established that older patients are at higher risk of ADRs; on the other hand, there is some evidence (albeit less compelling) that higher hemoglobin A1c (HbA1c) may be associated with a lower risk of TB-related ADRs. A study of $n = 900$ adults was conducted to understand the relationship between baseline HbA1c and ADRs. The table below shows the proportion of patients with ADR by diabetes status (defined as HbA1c levels over 6.5%).

	No ADR	ADR
No diabetes	356	72
Diabetes	409	63

(a) To investigate this association, consider (at least initially) using a chi-square test. Consider the following descriptions of the null hypothesis:

- H_0^A: There is no significant association between diabetes and occurrence of ADRs.
- H_0^B: Occurrence of ADRs in adults undergoing tuberculosis treatment is independent of diabetes.

Which of these two descriptions appears to be a more accurate reflection of a suitable null hypothesis? Briefly explain your answer.

(b) Calculate the expected table under the null hypothesis (cross-tabulated in the fashion shown above). Recall and utilize the following formula for the chi-square test statistic:

$$\chi^2 = \sum_i \sum_j \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}}.$$

(c) The following page presents key quantiles of chi-square distributions with various degrees of freedom. Although you will not be able to compute an exact p-value for the chi-square test, you should be able to supply a range. Briefly summarize your conclusion.

(d) Consider the following logistic regression model:

$$\text{logit}(\text{P}(\text{ADR}|\text{Diabetes})) = \beta_0 + \beta_1 \text{Diabetes}$$

Provide plain-language interpretations of $\exp(\beta_0)$ and $\exp(\beta_1)$; calculate point estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as they would be obtained from, say, the `glm()` command in R for logistic regression.

A colleague suggested a model in which we (1) include continuous HbA1c instead of the dichotomous diabetes variable, and (2) adjust for potential confounders. The proposed model is given by:

$$\text{logit}(\text{P}(\text{ADR}|\text{HbA1c}, \text{Age})) = \beta_0 + \beta_1 \text{HbA1c} + \beta_2 \text{Age}.$$

(e) Briefly discuss why age qualifies as a potential confounder.

(f) Suppose, hypothetically, that age had no association with HbA1c in this population. State whether you would prefer to adjust for age anyway; briefly justify your answer.

(g) Discuss the advantages and/or disadvantages of adjusting for continuous HbA1c instead of binary diabetes status.

(h) R output for this model is supplied on the following page. Write a brief summary of the findings in a way that would be suitable for a scientific collaborator.

Supplementary Material for Problem 3

$(1-q)$-quantiles, χ_{df} distirbution				
	df = 1	df = 2	df = 3	df = 4
$q = 0.01$	6.63	9.20	11.3	13.3
$q = 0.05$	3.84	5.99	7.81	9.49
$q = 0.10$	2.71	4.61	6.25	7.78
$q = 0.20$	1.64	3.22	4.64	5.99

Note: You will not need all of the information in the above table. As an example, the 0.9-quantile of a chi-square distribution with three degrees of freedom (i.e., the value c for which $\mathrm{P}(\chi^2_3 > c) = 0.1$) is 6.25.

```
## Call:
## glm(formula = adr ~ hba1c + age, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.97817 0.88667  -2.231  0.02568 *
## hba1c       -0.19177 0.06472  -2.963  0.00305 **
## age          0.03418 0.01853   1.845  0.06510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 760.88  on 899  degrees of freedom
## Residual deviance: 749.40  on 897  degrees of freedom
## AIC: 755.4
##
## Number of Fisher Scoring iterations: 4
```

4. [25 pts] Data from a sample of fifty-three American cities were collected to estimate the association between nitrogen oxide pollution (NO_x) and age-adjusted mortality. Two linear regression models were fit to these data, described as follows:

   (I) Including log-transformed NO_x (`log.nox`) as the predictor and mortality (`mortality`) as the outcome.

   (II) Allowing an interaction between yearly rainfall (`yr.rain`) and log-transformed NO_x.

   The following page depicts abridged R output from these models, as well as a scatter plot of log-transformed NO_x and yearly rainfall.

(a) Using the output of Model I, provide a point estimate of the mean mortality in a city with a NO_x level of 7.4 units. Briefly describe the assumptions that must hold in order for you to trust the validity of this estimate, as well as the extent to which these assumptions can be evaluated given only the output provided on the following page.

(b) Using the output of Model I, form an approximate 95% prediction interval (i.e., reference range) for age-adjusted mortality for a city with a NO_x level of 7.4 units. *Note*: You will not be able to establish this range exactly, but you should be able to approximate it. Please include as part of your response a description of the simplifications you make and the extent to which these simplifications may or may not be justified based on the output provided on the following page.

(c) Using the output of Model II, briefly comment on the extent of statistical evidence that the association between age-adjusted mortality and $\log(NO_x)$ is modified by yearly rainfall. Briefly describe the assumptions that must hold in order for you to trust this conclusion, as well as the extent to which these assumptions can be evaluated given only the output provided on the following page.

(d) Suppose that Nashville gets about 40 inches of rain per year and has a NO_x level of approximately 8 units. Determine the predicted age-adjusted mortality rate for Nashville according to Model II. Briefly describe how the scatter plot provided on the following page is useful for assessing the reliability of your answer to this question.

(e) Briefly discuss the extent to which you agree or disagree with the following statement:

"Since the coefficient for log(NO_x) is negative in Model II, and since its respective confidence interval excludes zero, Model II suggests that higher levels of log(NO_x) are associated with smaller age-adjusted mortality rates, adjusted for yearly rainfall."

(f) Briefly discuss the extent to which you agree or disagree with the following statement:

"The p-value for yearly rainfall in Model II is not significant; therefore, yearly rainfall is not associated with age-adjusted mortality."

(g) Briefly discuss a reason why you might be concerned about potential violations to the assumption of independent errors in this problem.

Supplementary Material for Problem 4

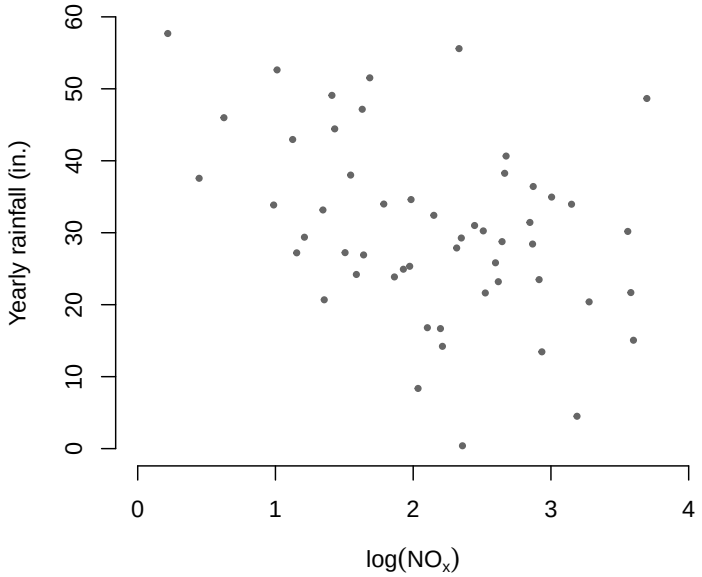
```
## RMS library
library(rms)

## Model I (and output)
MODEL.I <- ols(mortality ~ log.nox, x = TRUE, y = TRUE)

# > robcov(MODEL.I)
# Linear Regression Model
#
# sigma 51.2109
#
#               Coef    S.E.     t  Pr(>|t|)
# Intercept 892.0898 18.9308 47.12   <0.0001
# log.nox     2.3229  9.2266  0.25    0.8022


## Model II (and output)
MODEL.II <- ols(mortality ~ log.nox*yr.rain, x = TRUE, y = TRUE)

# > robcov(MODEL.II)
# Linear Regression Model
#
# sigma 35.5893
#
#                      Coef    S.E.     t  Pr(>|t|)
# Intercept         900.1794 30.9828 29.05   <0.0001
# log.nox           -38.7548 13.3323 -2.91    0.0055
# yr.rain            -0.6973  0.5981 -1.17    0.2493
# log.nox * yr.rain   1.6494  0.2880  5.73   <0.0001
```



5. [15 pts] Below is a sample of R code used to conduct a simulation. Investigate the code carefully before responding to the questions that follow.

```
1     simulation <- function(M=100000, n=100, beta0=0, sigmaY=3, seed=2025, beta1=1, rho, beta2) {
2       set.seed(seed)
3       est.res <- matrix(0, nrow = M, ncol = 2)
4       Sigma <- matrix(c(1, rho, rho, 1), nrow = 2, ncol = 2)
5       for (j in 1:M) {
6         predictors <- rmvnorm(n, mean = c(0,0), sigma = Sigma) # Bivariate normal
7         x <- predictors[,1]
8         z <- predictors[,2]
9         y <- beta0 + beta1 * x + beta2 * z + rnorm(n, 0, sigmaY)
10        zz1 <- lm(y ~ x)
11        zz2 <- lm(y ~ x + z)
12        est.res[j,1] <- coef(zz1)[2]
13        est.res[j,2] <- coef(zz2)[2]
14      }
15      res <- as.numeric(c(colMeans(est.res), apply(est.res, 2, sd)))
16      names(res) <- c("Unadj. Est.", "Adj. Est.", "Unadj. SE.", "Adj. SE.")
17      return(res)
18    }
```

The simulation was run six times under a combination of parameters (in particular, for two choices of `rho` and three choices of `beta2`), as shown below:

```
sim1 <- simulation(rho = 0, beta2 = 0)
sim2 <- simulation(rho = 0, beta2 = -1)
sim3 <- simulation(rho = 0, beta2 = 1)
sim4 <- simulation(rho = 0.5, beta2 = 0)
sim5 <- simulation(rho = 0.5, beta2 = -1)
sim6 <- simulation(rho = 0.5, beta2 = 1)
```

Unfortunately, the results of the simulation became scrambled when transcribed into the table below (who could have done such a terrible thing?):

	Unadj. Est.	Adj. Est.	Unadj.SE.	Adj. SE.
outputA	1.00	1.00	0.305	0.354
outputB	1.00	1.00	0.321	0.306
outputC	1.00	1.00	0.305	0.306
outputD	1.50	1.00	0.318	0.354
outputE	0.50	1.00	0.318	0.354
outputF	1.00	1.00	0.321	0.306

(a) For each simulation run, `sim1` through `sim6`, identify the corresponding row in the output (i.e., `outputA` through `outputF`. Some simulation runs provide the exact same output; as such, you may only be able to narrow it down to two rows.

(b) Now, having unscrambled the results, describe the key points that this study illustrates in a way that would be understandable to a scientific collaborator who has some training in applied biostatistics.

(c) You have not been supplied any output that considers the effect of varying the simulation parameters `n`, `beta0`, and `sigmaY`. What would be the most likely consequences of varying these parameters? You may consider each simulation parameter independently.
