

Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1
(In-class portion)

May 24, 2023

Instructions: Please adhere to the following guidelines:

- This exam begins on Wednesday, May 24 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 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. [25 pts] You are the biostatistician working with an investigative team conducting a phase 1 trial to evaluate an experimental drug for patients with hypertension. Specifically, systolic blood pressure (SBP, measured in mm Hg) was compared between the two groups following thirty days of medication use. The data from this study are presented in the table below:

Placebo control	125	130	135	140	135	145
Experimental drug	125	120	130	130	135	140

Let μ_0 and μ_1 denote the population mean 30-day SBP associated with the placebo control and experimental drug, respectively. Similarly, let σ_0^2 and σ_1^2 denote the respective population variances. Your primary goals are to estimate $\delta = \mu_1 - \mu_0$ and to test the hypothesis $H_0 : \delta = 0$. Below are approximate 95th and 97.5th percentiles of the t -distribution having between four and twelve degrees of freedom (you may not need all of this information).

df	4	5	6	7	8	9	10	11	12
$t_{0.95,df}$	2.13	2.02	1.94	1.89	1.86	1.83	1.81	1.80	1.78
$t_{0.975,df}$	2.78	2.57	2.45	2.36	2.31	2.26	2.23	2.20	2.18

-
- (a) Compute unbiased estimates of μ_0 , μ_1 , and δ (call them $\widehat{\mu}_0$, $\widehat{\mu}_1$, and $\widehat{\delta}$, respectively).
- (b) Compute unbiased estimates of σ_0^2 and σ_1^2 (call them $\widehat{\sigma}_0^2$ and $\widehat{\sigma}_1^2$, respectively).
- (c) Compute an unbiased estimate of $\text{Var}(\widehat{\delta})$ and explain what this quantity is in the language of sampling distributions.
- (d) Regarding a test of the hypothesis $H_0 : \delta = 0$, recall that we typically prefer to use the version of the t -test that allows unequal variances between groups. However, Student's t -test—which assumes a shared variance $\sigma^2 = \sigma_0^2 = \sigma_1^2$ —may be considered reasonable for this problem. Briefly explain why.
- (e) Determine the value of the t -statistic associated with Student's t -test.
- (f) State the approximate sampling distribution of the t -statistic of part (e) if the samples are independent and the group-specific variances are indeed equal. Under what further assumption(s) would this distribution be exact?
- (g) Although you cannot compute the p-value without further information, use your answer from parts (e) and (f) to determine whether the t -test reaches statistical significance at the nominal $\alpha = 0.05$ level (two-sided).
- (h) Construct an appropriate 95% confidence interval for δ . In what key way does this harmonize with your answer to part (g)?
- (i) The study's lead investigator sheepishly returns to you and informs you that the data presented in the table above are actually pre-treatment/post-treatment measurements from a crossover design. That is, each column actually corresponds to a single individual (the upper row denoting SBP following a month on placebo and the lower row denoting SBP following a month on the experimental drug). In light of this, complete the following four tasks taking the “paired” nature of the data into account (though do **not** go back and alter your responses to parts (a)-(h), in which you assumed the data were independent):
- Compute an unbiased estimate for δ .
 - Compute an unbiased estimate of $\text{Var}(\widehat{\delta})$.
 - Construct an appropriate 95% confidence interval for δ .
 - Determine whether the study provides sufficient evidence of a difference in means using an approach of your choice (though you cannot compute the exact p-value, you should be able to arrive at the correct conclusion based on a nominal level of $\alpha = 0.05$, two-sided).
- (j) Briefly describe one advantage and one disadvantage of the paired design as compared to the independent-group design.
-

2. [20 pts] Data from a cohort study of older adults was used to perform an analysis to understand the association between diabetes and coronary heart disease (CHD). Let X denote diabetes status (0 = no; 1 = yes) and let Y denote a three-level characterization of CHD (0 = no CHD; 1 = angina; 2 = myocardial infarction). The following page depicts Stata output for four commands, described as follows:

- A 2×3 cross-tabulation of diabetes status and CHD.
- Results from a multinomial logistic regression model with CHD as the outcome and diabetes as the predictor. Although you can infer the parameterization by examining the output, it is noted as follows:

$$\log \left(\frac{P(\text{chd} = k | \text{diabetes})}{P(\text{chd} = 2 | \text{diabetes})} \right) = \beta_{0k} + \beta_{1k} \text{diabetes} \quad \text{for } k = 0, 1.$$

- Results from a cumulative logit model with CHD as the outcome and diabetes as the predictor. Although you may recall the parameterization of the cumulative logit model, it is noted as follows:

$$\text{logit}(P(\text{chd} \leq k | \text{diabetes})) = \beta_{0k} - \beta_1 \text{diabetes} \quad \text{for } k = 0, 1.$$

This is sometimes referred to as a “proportional odds” or an “ordinal regression” model.

- Results from a logistic regression model with dichotomized CHD (`chd2`) as the outcome and diabetes as the predictor. Note that `chd2` is coded as (0 = no CHD; 1 = angina *or* myocardial infarction). Although you can infer the parameterization from this description, it is noted as follows:

$$\text{logit}(P(\text{chd} = 1 \text{ or } 2 | \text{diabetes})) = \beta_0 + \beta_1 \text{diabetes}.$$

Some of the software output has been abridged for ease of access to the salient information in this problem.

-
- Suppose you seek to compute a point estimate for the prevalence of angina among older adults with diabetes. For each of the four methods described above, state whether it is possible to use that method to do so. For those that allow you to compute the point estimate, use the software output on the following page do so; for those that do not, briefly describe why it is not possible.
 - Among the methods you identified as allowing you to compute a point estimate of the prevalence described in part (a), identify and briefly account for agreements/disagreements across your computations.
 - Suppose you seek to compute a point estimate for the odds ratio that compares the odds of CHD (i.e., either angina or myocardial infarction) between older adults with and without diabetes. For each of the four methods described above, state whether it is possible to use that method to do so. For those that allow you to compute the point estimate, use the software output on the following page do so; for those that do not, briefly describe why it is not possible.
 - Among the methods you identified as allowing you to compute a point estimate of the odds ratio described in part (c), identify and briefly account for agreements/disagreements across your computations.
 - Among the methods you identified as allowing you to compute a point estimate of the odds ratio described in part (c), choose *one* such method to further obtain a 95% confidence interval for that odds ratio and briefly justify your choice (note that there may be more than one justifiable choice).
-

Supplementary Material for Problem 2

* OUTPUT FOR CROSS-TABULATION OF DIABETES AND CHD

. tabulate diabetes chd

diabetes	chd			Total
	0	1	2	
0	524	57	75	656
1	56	7	16	79
Total	580	64	91	735

* ABRIDGED OUTPUT FOR MULTINOMIAL LOGISTIC REGRESSION MODEL

. mlogit chd i.diabetes, robust nolog baseoutcome(2)

	chd	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]	
0							
1.diabetes		-.6912406	.3094012	-2.23	0.025	-1.297656	-.0848253
_cons		1.944004	.1235415	15.74	0.000	1.701867	2.18614
1							
1.diabetes		-.5522417	.4863704	-1.14	0.256	-1.50551	.4010268
_cons		-.2744368	.1758387	-1.56	0.119	-.6190744	.0702007
2		(base outcome)					

* ABRIDGED OUTPUT FOR CUMULATIVE LOGIT PROPORTIONAL ODDS MODEL

. ologit chd i.diabetes, robust nolog or

	chd	Odds ratio	Robust std. err.	z	P> z	[95% conf. interval]	
1.diabetes		1.693234	.4567043	1.95	0.051	.9979946	2.872802
/cut1		1.382278	.0972503			1.191671	1.572886
/cut2		2.022455	.1166458			1.793833	2.251077

Note: Estimates are transformed only in the first equation to odds ratios.

* ABRIDGED OUTPUT FOR LOGISTIC REGRESSION MODEL ON DICHOTOMIZED OUTCOME

. logit chd2 i.diabetes, robust nolog

	chd2	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]	
1.diabetes		.4888324	.2663007	1.84	0.066	-.0331073	1.010772
_cons		-1.37869	.0974529	-14.15	0.000	-1.569694	-1.187686

3. [30 pts] A laboratory study of $n = 80$ independent cell cultures was conducted to evaluate two experimental chemotherapy agents—both individually and in combination—as compared to a control condition. The cell cultures were randomly assigned in the fashion described by the table below:

	Agent 2: No	Agent 2: Yes
Agent 1: No	$n = 20$	$n = 20$
Agent 1: Yes	$n = 20$	$n = 20$

Following incubation, the colony count was measured for each culture (more potent agents inhibit growth so that lower counts suggest greater efficacy). The two research assistants running this experiment were each randomly assigned ten cultures from each of the four groups, but were unfortunately given conflicting instructions from the principal investigator by mistake: Steve was told to use an incubation period of 24 hours and Jane was told to use an incubation period of 48 hours (you may assume *all other* aspects of the procedure did not differ between Steve and Jane). The variables measured in this study were as follows:

agent1	receipt of agent 1 (0=no; 1=yes)
agent2	receipt of agent 2 (0=no; 1=yes)
rsch	research assistant (1=Steve; 2=Jane)
count	colony count after incubation period

Stata output is shown on the following page for a Poisson regression model and post-estimation commands.

-
- (a) Briefly describe how the model addresses the principal investigator's mistake as described above.
- (b) Briefly describe the role of the robust standard error as it pertains to Poisson regression models.
- (c) Is it possible to use the Stata output on the following page to characterize the strength of evidence that Agent 1 has an overall effect on expected colonies per day? If so, do so; if not, explain why you can't.
- (d) Let θ_1 denote the ratio of expected colonies per day comparing cultures receiving Agent 1 alone to cultures receiving neither agent. Determine whether you can use the Stata output on the following page to determine each of the following. For the ones you can, do so; otherwise, briefly explain why you can't.
- A point estimate for θ_1 .
 - A 95% confidence interval for θ_1 .
 - A p-value for the hypothesis test $H_0 : \theta_1 = 1$.
- (e) Let θ_2 denote the ratio of expected colonies per day comparing cultures receiving both agents to cultures receiving Agent 2 alone. Determine whether you can use the Stata output on the following page to determine each of the following. For the ones you can, do so; otherwise, briefly explain why you can't.
- A point estimate for θ_2 .
 - A 95% confidence interval for θ_2 .
 - A p-value for the hypothesis test $H_0 : \theta_2 = 1$.
- (f) Let θ_3 denote the expected colonies per *hour* among cultures receiving Agent 2 alone. Determine whether you can use the Stata output on the following page to determine each of the following. For the ones you can, do so; otherwise, briefly explain why you can't.
- A point estimate for θ_3 .
 - A 95% confidence interval for θ_3 .
 - An approximate “educated guess” of a p-value for the hypothesis test $H_0 : \theta_3 = 0.42879$.
- (g) Characterize the strength of evidence that the two agents interact in their effects on expected colonies per day. In this sense, do they appear to bolster or hinder each other's potency? Justify your answer.
- (h) Despite your answer to part (g), why is it a mischaracterization to suggest that Agents 1 and 2 do not perform well when given in combination?
-

```
* REGRESSION MODEL OUTPUT
```

Poisson regression

Wald chi2(3) = 45.13

```
Prob > chi2    = 0.0000
```

Log pseudolikelihood = -218.22346

Pseudo R2 = 0.0879

* OUTPUT FOR SEVERAL POST-ESTIMATION COMMANDS

(1) [count]1.agent1 + [count]1.agent1#1.agent2 = 0

```
. test (1.agent1 = 0) (1.agent1#1.agent2 = 0)
```

```
( 1) [count]1.agent1 = 0
```

(2) `[count]1.agent1#1.agent2 = 0`

```
. lincom _cons + 1.agent2
```

```
( 1) [count]1.agent2 + [count]_cons = 0
```

count	Coefficient	Std. err.	z	P> z	[95% conf. interval]
(1)	2.428762	.0497435	48.83	0.000	2.331266 2.526257

4. [15 pts] Below is a sample of R code used to compare the finite-sample behavior of two competing methods via simulation. Investigate the code carefully before responding to the questions that follow.

```
1      simulation <- function(N, p, reps=10000, level=0.05, seed=2023) {
2          set.seed(seed)
3          reject <- matrix(0, nrow = reps, ncol = 2)
4          for (j in 1:reps)
5              {
6                  x0 <- rbinom(n = N/2, size = 1, prob = p)
7                  x1 <- rbinom(n = N/2, size = 1, prob = p)
8                  p0.hat <- sum(x0)/(N/2)
9                  p1.hat <- sum(x1)/(N/2)
10                 p.hat.null <- (sum(x0) + sum(x1))/N
11                 diff.hat <- p1.hat - p0.hat
12                 v.hat.A <- p0.hat * (1 - p0.hat)/(N/2) + p1.hat * (1 - p1.hat)/(N/2)
13                 v.hat.B <- p.hat.null * (1 - p.hat.null) * (1/(N/2) + 1/(N/2))
14                 z.A <- diff.hat/sqrt(v.hat.A)
15                 z.B <- diff.hat/sqrt(v.hat.B)
16                 p.val.A <- 2 * (1 - pnorm(abs(z.A)))
17                 p.val.B <- 2 * (1 - pnorm(abs(z.B)))
18                 reject[j,1] <- as.numeric(p.val.A < level)
19                 reject[j,2] <- as.numeric(p.val.B < level)
20             }
21         output <- colMeans(reject)
22         return(output)
23     }
```

The results of this simulation are shown below for a few combinations of values for N and p:

	output[1]	output[2]
N=40; p=0.6:	0.0715	0.0450
N=40; p=0.7:	0.0669	0.0531
N=40; p=0.8:	0.0662	0.0504
N=50; p=0.6:	0.0627	0.0605
N=50; p=0.7:	0.0648	0.0545
N=50; p=0.8:	0.0607	0.0550
N=60; p=0.6:	0.0527	0.0489
N=60; p=0.7:	0.0576	0.0476
N=60; p=0.8:	0.0595	0.0481

-
- (a) Briefly explain the major purpose of the `seed` argument (referred to on Line 2).
- (b) Identify the two methods being compared in this simulation study.
- (c) Identify the property that is being compared between the two methods you identified in part (b).
- (d) Briefly explain the most important reason to specify a value for the `reps` argument (referred to on Lines 3 and 4) that is sufficiently high.
- (e) 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.
- (f) When setting `N=6` and `p=0.999`, the simulation function returns the values `NA` for both `output[1]` and `output[2]`; this seems to occur even when varying the `seed` argument across a range of values. Identify the source of this problem.
- (g) Given the arguments `N=6`, `p=0.9`, and `reps=5`, determine the probability of running into the problem described in part (f) for both `output[1]` and `output[2]` given a random seed. Show your work and report your answer (though you needn't simplify; e.g., an answer of the form $\sqrt{3+0.3^2}$ is acceptable).
-

5. 10 pts Suppose you seek to model a process involving a positive-valued predictor X and an outcome Y such that $E[Y|X = x] = f(x)$ is a piecewise quadratic function with a single knot (that is, f is quadratic for $x \in (0, c]$ and quadratic for $x > c$). Further, you impose the constraint that f is differentiable over its domain. Treat the knot, c , as a fixed and known value.
-

- (a) Express f in terms of a basis expansion having the following form:

$$f(x) = \sum_{p=1}^P \beta_p h_p(x; c)$$

for P functions $h_1(x; c), \dots, h_P(x; c)$ that you are to determine.

- (b) You should have found in part (a) that $P = 4$, meaning the model uses four degrees of freedom. Briefly explain how you could have used knowledge about f and its constraints to arrive to that conclusion even without specifically deriving the basis expansion.
- (c) Show that under the *additional* constraint of being twice-differentiable over its domain, f must follow the form $f(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2$.
-