PHS2000B Exam 1

Epidemiologic Methods

March 11, 2021

Welcome to our 1st exam! As a reminder, you have 24 hours to complete and upload your exam to the Assignments tab on Canvas. You may refer to your notes and any books you would like, but please do not discuss the exam with any other students from class or your tutor. By submitting your exam on Canvas, you affirm that the work you are submitting is yours alone and that you have not discussed the questions or responses with anyone other than the instructional team.

We wish to emphasize that you are more than welcome to handwrite your responses on paper, paste them into a word document, or submit your write up as a scanned PDF file (please submit as a single file if possible). If you choose to handwrite, please circle or otherwise indicate your final answer. Alternatively, you are welcome to type your responses in your program of choice (e.g. Word, TeX, markdown, etc.) Please note that any math or drawings you want to include can be typeset into the document or can be pasted in as an image. You will NOT be graded based on the format you choose to submit your exam.

If you have clarifying questions for the instructional team, please email the whole team and we will get back to you as soon as possible. During our regular 11:30 am – 1:00 pm lecture period, you may also log on to the Zoom meeting and get an immediate response from the instructional team.

There is a Google doc where we will be posting any clarifications that arise from questions we receive during the exam. If we post a clarification, we will also post an announcement referring you to the Google doc.

Good luck!

Propensity Scores and Marginal Structural Models (13 points)

Baseline Injection-drug Use and HIV Infection (6 points)

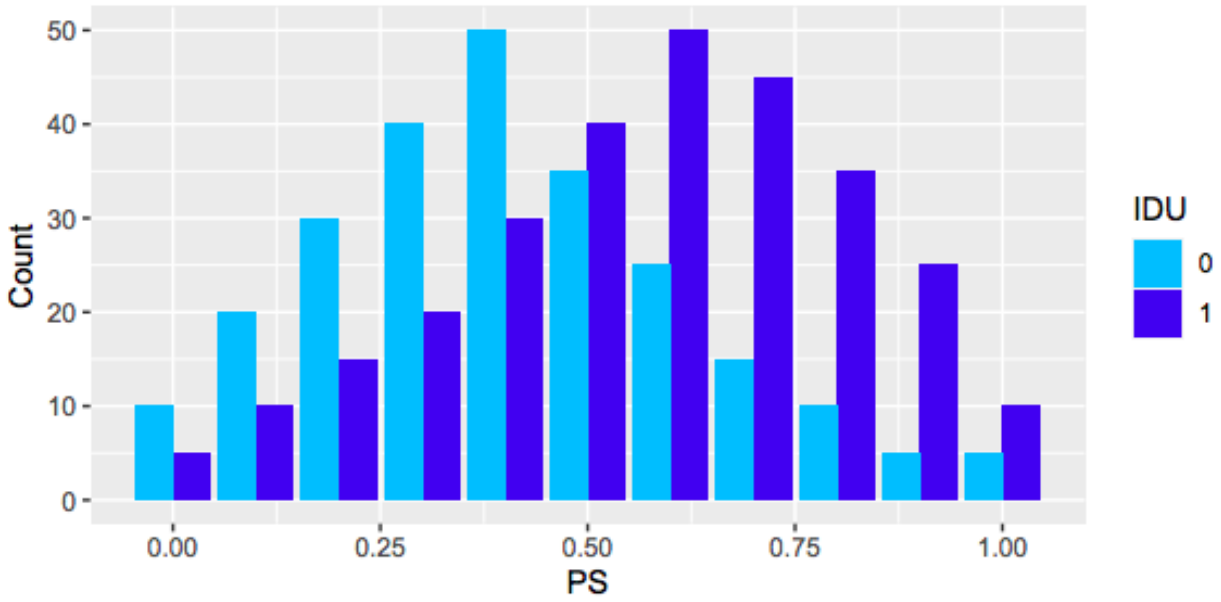
You are part of a team that is interested in the effect of injection-drug use (A) (IDU = 1, No-IDU = 0) on the risk of HIV infection (Y) (Infection = 1, No infection = 0). The data consists of individuals followed over 4 years with current IDU status measures in 2016 (t = 0) and 2018 (t = 1) as well as whether an individual was infected with HIV by 2020 (t = 2). You also have data on potential confounders year of birth (C) (continuous) and income (L) (continuous) measured in 2016 and 2018 (there are obviously many other confounders that may be considered, but we have reduced it to two for simplicity). Assume there is no measurement error and no loss to follow-up.

The first analysis your team pursues is the effect of baseline IDU in 2016 on HIV infection status in 2020. Use the following DAG for questions 1-4.

1. Name one advantage and one disadvantage of using propensity score methods over traditional multivariable outcome regression. (1 point)

2. Write a model to estimate the propensity score including all the covariates. How would you calculate the propensity score for an individual born in 1980 who made $50,000 in income in 2016? (2 points)

3. You obtain the following distributions of the propensity score in the treated and the untreated. Evaluate the region of common support. (1 point)

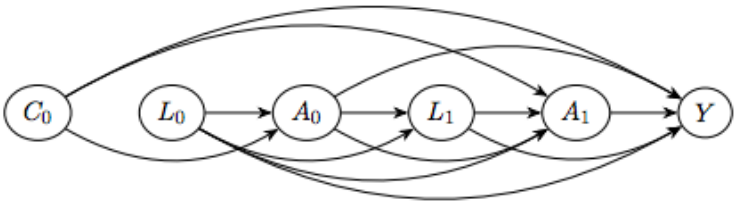


4. Some of your colleagues are arguing for stratification on the propensity score. Others want to use matching. Name one advantage and one disadvantage of using each of the two strategies. (2 points)

Time-varying Injection-drug Use and HIV Infection (7 points)

Your team would also like to leverage your longitudinal data on IDU in studying the effect on HIV infection. You are interested in the joint effect of IDU in 2016 and 2018 on HIV infection status by 2020. Year of birth is a time-invariant confounder, but the team is concerned about time-varying confounding of IDU status by income.

Use the following DAG for questions 5-8.



5. A colleague suggests using propensity score stratification using all the covariates (C_0, L_0, L_1) in a single propensity score to control for time-varying confounding to obtain the joint effect of interest. Explain why this method will fail in 3 sentences or less. (Hint: A DAG may be useful to illustrate your reasoning) (1 point)

6. Instead of stratification, your team decides to pursue inverse-probability of treatment weighting. Write an expression that estimates the overall stabilized weight for an individual who is an injection-drug user at both time points in terms of the probability of IDU at each time point. (Do not include any baseline confounders in the numerator) (2 points)

7. You fit the following marginal structural model in your weighted population and obtain the following estimates: $\widehat{\beta_0} = 2.5, \widehat{\beta_1} = 0.2, \widehat{\beta_2} = 0.5$. Calculate and provide an interpretation of the joint marginal effect estimate of IDU status at the two time points on HIV infection on the odds ratio scale. The counterfactual contrast of interest is between $Y_{a_0=1,a_1=1}$ and $Y_{a_0=0,a_1=0}$. (2 points)

$$logit(Pr(Y_{a_0 a_1} = 1)) = \beta_0 + \beta_1 A_0 + \beta_2 A_1$$

8. Your team is interested in effect modification of the effect of IDU at the two time points by year of birth. Write a new expression for appropriate stabilized weights for an injection-drug user at both time points and a new marginal structural model incorporating this effect modifier at both exposure times. (2 points)

Measurement Error (13 points)

According to a recent UNAIDS report, an emerging "blind spot" of the global HIV/AIDS response has been in outreach to men and boys, who are both less likely than women to know their HIV status and appear to be less likely to have access and successfully adhere to HIV treatment. Consequently—while women are disproportionately more affected than men by HIV—more men are likely to die of AIDS-related conditions than women.

You are working for a city health department that is interested in improving access to HIV testing among men. You learn that the city recently conducted seroprevalence surveys, which provide you with recent data on HIV prevalence among key populations—sub-groups in which HIV prevalence and risk of infection can be high. Assume that these data reflect the true prevalence of HIV in these populations.

The city has a stockpile of rapid HIV antibody tests that it distributes to clinics, support groups, and other partners in outreach as part of its testing program. The tests are saliva brush tests, which have been reported to have a sensitivity of 95% and a specificity of 80% in previous validation studies.

Male sex workers	Gay men and other men who have sex with men	Men who inject drugs	All men (aged 15 years and older)
24%	6%	3%	2%

1. One of the key partners for the testing program is a community-led civil society organization that promotes the health, safety, and well-being of male sex workers in your city. As part of their health promotion efforts, the organization helps administer the rapid HIV tests donated by the city to **male sex workers**, and then refers folks who test positive to the city health department for additional services. Based on the sensitivity and specificity of the test and the seroprevalence data above, what is the likelihood that a positive result on a test provided through this organization actually reflects a true case of HIV? (2 points)

2. You are chatting with a colleague from the city's harm reduction program, which also utilizes the city's antibody HIV tests. The program primarily serves **men who inject drugs**; HIV testing is offered alongside a needle exchange service and opioid substitution therapy. Your colleague mentions that she does not think the test is particularly useful, since most of her clients who test positive on the tests are later determined to have had false positive results when they are retested with more accurate PCR (polymerase chain reaction) tests. Based on the data above, what proportion of positive results on HIV tests administered by the harm reduction program are likely to be false positives? (2 points)

3. Your boss—who is not trained in epidemiology—notices that the share of positive results that are later shown to be false positives is different among tests administered through the harm reduction program compared to those administered through the male sex worker testing outreach partnership. "The test kits we send to one program must somehow be different from the others…or, worse, they are defective!" In 2-3

sentences, why might the positive predictive values be different across testing programs, even if the tests themselves are identical and working as expected? (2 points)

4. A professor from a local university contacts you; she is interested in conducting a case-control study examining the association between experiences of healthcare discrimination (a binary variable, coded as sometimes, usually, or always experienced **versus** never experienced) and HIV infection. In her study population, the respondents were tested for HIV using the extremely accurate PCR test, the gold standard for testing acute HIV infections. She has discovered, however, that a small error in her database has affected a random fraction of her observations. The error randomly re-coded the binary variable for reported experiences with healthcare discrimination, such that 1s were recoded as 0s and 0s were recoded as 1s. While she is confident that the error misclassified these values entirely by random chance, she cannot reverse the error.

    a. What type of misclassification error is this an example of? (1 point)

    b. Will this error result in bias? If so, is the bias towards the null, away from it, or is the direction of the bias unclear? If not, explain why the result is not biased. (2 points)

    c. Imagine that a test of no association between discrimination and HIV status is conducted using these data and that the p-value associated with this test comes back as 0.015. Based on your answers above, is the true likelihood of a type I error equal to, smaller than, or greater than 1.5%? (2 points)

Interaction and Mediation (12 points)

Suppose you are a lead investigator on a study that aims to assess the effect of participation in the Supplemental Nutrition Assistance Program (SNAP), on food security. Households that participate receive a monthly debit card that helps offset the cost of food purchased at SNAP-participating vendors such as grocery stores. As part of your study, you collect the following information:

$$A = \begin{cases} 0 & \text{if a household did not participate in SNAP in the past month} \\ 1 & \text{if a household participated in SNAP in the past month} \end{cases}$$

$$Y = \quad \text{Continuous household food security score, ranging from 0 (food insecure) to 50 (food secure)}$$

$$Z = \begin{cases} 0 & \text{if a household's financial assets were below the poverty level} \\ 1 & \text{if a household's financial assets were above the poverty level} \end{cases}$$

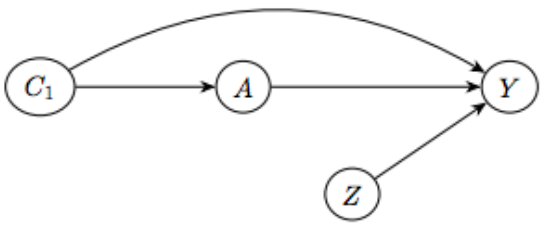
$$C_1 = \begin{cases} 0 & \text{if a household did not have access to a community pantry} \\ 1 & \text{if a household did have access to a community pantry} \end{cases}$$

Data from your study is summarized in tabular form below, where N is the number of households in each combination of the variables.

Table 1: SNAP Study Data

A	Z	C_1	N	$E(Y \mid A = a, Z = z, C_1 = c_1)$
0	0	0	100	10
1	0	0	200	15
0	1	0	100	15
1	1	0	200	25
0	0	1	100	25
1	0	1	100	30
0	1	1	100	30
1	1	1	100	40

DAG 1



1. Assume that DAG 1 above represents the **true** data generating mechanism. Note that because we have complete data, all the effects below are identified. For each of the following counterfactual effects for the group that has access to a community pantry ($C_1 = 1$): (4 points)

   i. State whether the effect describes effect heterogeneity or causal interaction. (1 point)

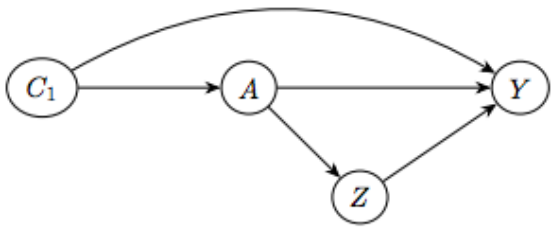
   ii. Provide an expression for the effect using only the *observed* variables, calculate it from the data in , and provide a 1-sentence interpretation. (1 point)

$$\text{(A)}: \left\{E(Y_{a=1,z=1} \mid C_1 = 1) - E(Y_{a=0,z=1} \mid C_1 = 1)\right\} - \left\{E(Y_{a=1,z=0} \mid C_1 = 1) - E(Y_{a=0,z=0} \mid C_1 = 1)\right\}$$

$$\text{(B)}: \frac{\left\{E(Y_{a=1} \mid C_1 = 1, Z = 1)/E(Y_{a=0} \mid C_1 = 1, Z = 1)\right\}}{\left\{E(Y_{a=1} \mid C_1 = 1, Z = 0)/E(Y_{a=0} \mid C_1 = 1, Z = 0)\right\}}$$

2. Now suppose that it is possible that SNAP participation may affect household financial assets as in DAG 2 below. One of your colleagues wonders whether the effect of SNAP participation on food security is mediated by the effect of SNAP participation on household financial assets. They suggest the popular Baron & Kenny product method to assess the direct and indirect effects of SNAP exposure. Given your assessment of interaction in question 1, is this a valid approach if your interest is in the causal mediation effects? Explain in 1-2 sentences. (2 points)

DAG 2



3. Using counterfactual notation, provide expressions for the controlled direct effect (intervening to set $M = 1$), the natural direct effect, and the natural indirect effect. Interpret these within the context of the study. (3 points)

4. When reviewing the data, you realize that you have not measured information on inherited family wealth, which may affect both household SNAP participation (A) and financial assets (Z). You had planned on using the parametric approach to estimate the CDE, NDE, and NIE based on the models:

$$E(Z \mid A, C_1) = \beta_0 + \beta_1 a + \beta_2 c_1$$
$$E(Y \mid A, Z, C_1) = \theta_0 + \theta_1 a + \theta_2 z + \theta_3 az + \theta_4 c_1$$

Which, if any, of the following effect estimates would still be unbiased? Explain in 2-3 sentences (You may provide a DAG if you wish, but it is not necessary) (3 points)

$$CDE(1) = \theta_1 + \theta_3$$
$$NDE = \theta_1 + \theta_3(\beta_0 + \beta_2 c_1)$$
$$NIE = \theta_2\beta_1 + \theta_3\beta_1$$

Sensitivity Analysis and Unmeasured Confounding (12 points)

You come across a recently-published paper assessing the impact of long-term exposure to ambient fine particulate matter (PM2.5) on mortality in young people. In their analysis, the authors used Cox Proportional Hazards Models to assess the association of interest. Models were stratified by four individual-level characteristics (age, gender, race, and Medicaid eligibility), and adjusted for area-level risk factors and meteorological variables based on participant zip code. The authors presented a hazard ratio of 1.10 (1.06, 1.13). For the following questions, you may assume that the outcome is **rare**.

1. Calculate the E-value for the main effect estimate and 95% confidence interval. Interpret both results. (2 points for calculations, 1 for interpretation)

2. You come across another study that uses the same method, same exposure, and same outcome (we'll call this "Study B", and the first study "Study A"). Study B adjusts for age, race, gender, and income. The E-value for the main effect estimate of Study B is 1.75. Based on this information, can you conclude which study (Study A or Study B) provides stronger evidence of robustness to confounding? Why or why not? Answer in no more than 3 sentences. (2 points)

3. In Study A, you believe that outdoor physical ativity may be an important confounder. Through a quick literature search, you see that moderate outdoor physical activity is associated with lower mortality, characterized by a risk ratio of 0.65. What would the least extreme risk ratio for the association between moderate outdoor physical activity and ambient PM2.5 exposure have to be in order to explain away the confidence interval of the observed results? (2 points for calculations, 1 for interpretation)

4. A colleague, upon doing a literature review, believes that the association between physical activity and ambient PM2.5 exposure is no more than 1.70. Given this and the risk ratio of 0.65 for the association between physical activity and mortality, obtain a corrected estimate and 95% CI for the most such a confounder could alter the observed estimates. (4 points)