EPI202: Fall 2022
Homework 1
To be uploaded as PDF to course website by 9:30AM ET on November 3, 2022

Please provide concise, precise answers.

We encourage collaborative learning in this course. You may discuss homework assignments with other students. However, all written work that you submit for grading must be your own, in your own words, reflecting your understanding of the homework assignment. Homework assignments should not be prepared by copying, paraphrasing, or summarizing someone else's work.

Proper notation should be used throughout the assignment and all calculations should be shown.

Note: for all questions, please include all relevant formulas and define any variables that you use and have not already defined in previous questions to receive full credit.

Part I. Crude analysis of person-time data

Olfson M, Walla M, Wang S, Crystal S, Blanco C. Risks of fatal opioid overdose during the first year following nonfatal overdose. Drug and Alcohol Dependence. 2018; 190:112-119 https://www.sciencedirect.com/science/article/abs/pii/S0376871618303466?via%3Dihub

1. Use the data from Table 3 to create the following table that examines the crude association between age (age 18-34 versus all other age groups combined) and the incidence of fatal overdose during follow-up.

	Number of cases	Total person-years at risk
Age 18-34	165	20257
Age 35-64*	605	46479

*combines information from age 35-44 and 45-64.

2. Using the data in your table, is there an association between age and the incidence of fatal overdose in the study?
   (a) State the null and alternative hypotheses
   (b) Calculate a valid test statistic
   (c) Calculate the p-value associated with the test statistic
   (d) Interpret the p-value in words
   (e) State the outcome of a Neyman-Pearson hypothesis test

Construct another table to describe the data:

	Age 18-34	Age 35-64	
Cases	$a = 165$	$b = 605$	$M_1 = 770$
PY	$N_1 = 20257$	$N_0 = 46479$	$T = 66736$
	$\widehat{IR}_1 = 8.145$ cases / 1,000 PY	$\widehat{IR}_2 = 13.017$ cases / 1,000 PY	

H_0: The incidence rate of fatal overdose does not differ between those ages 18-34 compared to those ages 35-64. ($IR_1 = IR_0$).

H_A: The incidence rate of fatal overdose differs between those ages 18-34 compared to those ages 35-64. ($IR_1 \neq IR_0$).

Let X = the number of exposed cases = $a = 165$

$E(X|H_0) = \frac{M_1 N_1}{T} = \frac{770*20257}{66736} = 233.725,\ Var(X|H_0) = \frac{M_1 N_1 N_0}{T^2} = \frac{770*20257*46479}{66736^2} = 162.781$

Then, test statistic: $Z^2 = \frac{[X-E(X|H_0)]^2}{Var(X|H_0)} = \frac{[165-233.725]^2}{162.781} = 29.02$

$Pr[\chi^2 > 29.02] < 0.0001$

The probability that a result as extreme or more than 29.02 would occur due to random variation if the null hypothesis was true.

We reject the null at the $\alpha = 0.05$ level. We conclude that there is statistically significant evidence for an association between age (age 18-34 compared to age 35-64) and the incidence of fatal overdose, assuming no confounding, selection bias, or information bias.

3. Calculate the crude incidence rate ratio (IRR) for the incidence of fatal overdose for ages 18-34 compared to all older participants.
    a. Interpret your numerical result in words.
    b. Calculate the 95% confidence interval for the IRR and interpret your numerical result in words.

$\widehat{IRR} = \frac{\widehat{IR_1}}{\widehat{IR_2}} = \frac{8.145}{13.017} = 0.6258$

Assuming no confounding, selection bias, or information bias, the incidence rate of fatal overdose for ages 18-34 is 0.6258 times the incidence rate of fatal overdose for ages 35-64.

The 95% confidence interval for the ln IRR:

$X \pm 1.96\sqrt{\widehat{Var}(X)} = ln\,\widehat{IRR} \pm 1.96\sqrt{\widehat{Var}(ln\,\widehat{IRR})} = ln\,\widehat{IRR} \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b}}$

$= ln(0.6258) \pm 1.96\sqrt{\frac{1}{165} + \frac{1}{605}} = (-0.6409,\ -0.2967)$

So the 95% confidence interval for IRR: $e^{(-0.6409,\ -0.2967)} = (0.5268,\ 0.7433)$

These data are consistent with incidence rate ratios ranging from 0.5268 to 0.7433 with 95% confidence for the association between age (age 18-34 compared to age 35-64) and the incidence of fatal overdose, assuming no confounding, selection bias, or information bias.

Part II. Crude analysis of count data.

In this section, you will analyze data from the Myocardial Infarction Onset Study. This study was a multicenter cohort study of myocardial infarction patients enrolled between 1989 to 1996 at the time of their myocardial infarction. All participants were followed through the National Death Index until death or a minimum of 10 years after enrollment. There was no loss to follow-up. In this analysis, you will evaluate the relationship between reporting having hypertension at the baseline interview and the cumulative incidence of death from any cause over the following 10 years.

The variables in this dataset are described below:

Variable Name	**Description**
Id	ID number
Age	Age (continuous, years)
age_cat	Age Category (1: <50yrs, 2: 50-64 yrs, 3: 65+ yrs)
Anxiety	Anxiety (1: yes; 2: no)
Female	Female (1: female, 0: male)
Married	Married (1: yes, 0:no)
Educ	Educational Attainment (1: <HS, 2: HS, 3: >HS)
Dm	Diabetes (1: yes, 0:no)
Htn	Hypertension (1: yes, 0:no)
phys_activity	Frequency of Physical Activity (0: <1/wk, 1: 1-3/wk, 2: 4+/wk)
Evermarj	Ever use marijuana (1: yes, 0:no)
follow_up	Duration of follow-up (years)
Dead	Death within 10 years (1: died, 0: survived)
Cvdeath	Death from cardiovascular causes (1: CVD death, 0: did not die of CVD)

The dataset name is MI_Onset_10 and is available for download from the course website in several file formats including CSV, R, SAS, and Stata. You may use SAS, STATA, R, the EPI202 calculator, or any other statistical analysis software package of your choosing.

1. Read the dataset into your preferred software and complete the following 2x2 table:

	Hypertension reported at baseline		
Death from any cause	Yes	No	Total
Yes	a=601	b=492	M_1=1093
No	c=1029	d=1590	M_0=2619
Total	N_1=1630	N_0=2082	T=3712

2. What is the 10-year cumulative incidence of death from any cause in the entire sample?
    a. Provide a 95% confidence interval for the cumulative incidence you calculated (HINT, use the formula for binomial proportion).

The 10-year cumulative incidence of death from any cause in the entire sample:

$\widehat{CI} = \frac{1093}{3712} = 0.2945$ cases over the 10-year follow-up period

The 95% confidence interval for the 10-year cumulative incidence of death from any cause in the entire sample:

$$X \pm 1.96\sqrt{\widehat{Var}(X)} = \widehat{CI} \pm 1.96\sqrt{\frac{\widehat{CI}*(1-\widehat{CI})}{T}} = 0.2945 \pm 1.96\sqrt{\frac{0.2945*(1-0.2945)}{3712}}$$

$= (0.2798,\ 0.3091)$ over the 10-year follow-up period

3. Calculate the crude cumulative incidence difference (CID) of death over the ten-year follow-up period comparing those who reported having hypertension at baseline to those who did not.
    a. Interpret your numerical result in words.

$\widehat{CID} = \widehat{CI}_1 - \widehat{CI}_0 = \frac{601}{1630} - \frac{492}{2082} = 0.1324$ cases over the 10-year follow-up period

The cumulative incidence of death for those who reported having hypertension at baseline is 0.1324 cases excess to the cumulative incidence of death for those who did not report having hypertension at baseline over the 10-year follow-up period, assuming no confounding, selection bias, or information bias.

4. Construct a 95% confidence interval for the crude CID that you calculated in the previous question.
    a. Interpret your numerical result in words.

The 95% confidence interval for the crude CID:

$$X \pm 1.96\sqrt{\widehat{Var}(X)} = \widehat{CID} \pm 1.96\sqrt{\widehat{Var}(\widehat{CID})} = \widehat{CID} \pm 1.96\sqrt{\frac{ac}{N_1^3} + \frac{bd}{N_0^3}}$$

$$= 0.1324 \pm 1.96\sqrt{\frac{601*1029}{1630^3} + \frac{492*1590}{2082^3}} = (0.1027,\ 0.1621)$$ over the 10-year follow-up period

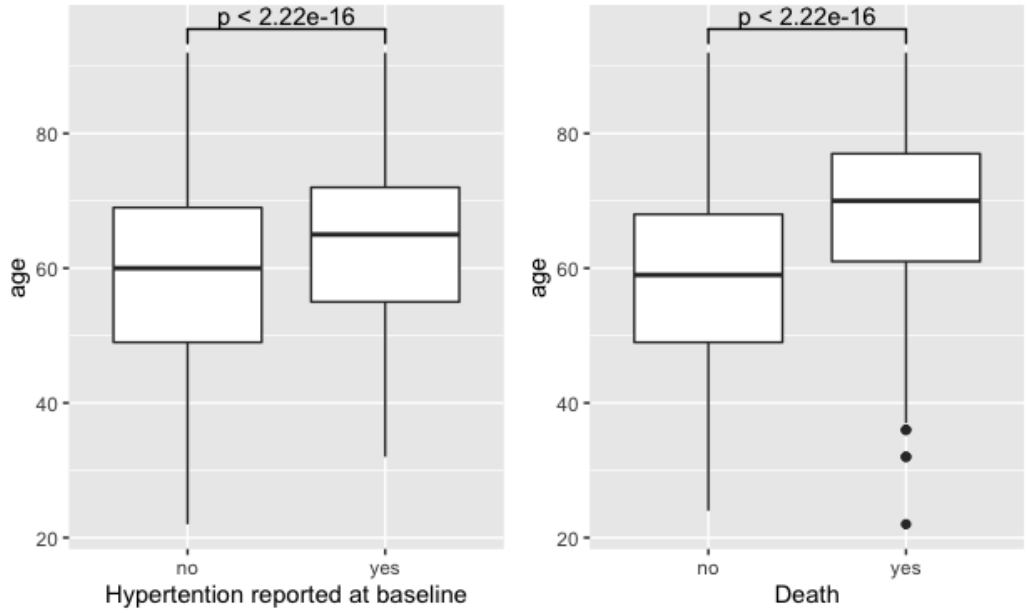
With 95% confidence, these data are consistent with cumulative incidence differences ranging from 0.1027 to 0.1621 over the 10-year follow-up period for the association between all-cause mortality and hypertension reported at baseline, assuming no confounding, selection bias, or information bias.

5. Do you think that the crude CID represents the causal cumulative incidence difference of the risk of hypertension on all-cause mortality? If not, why not?

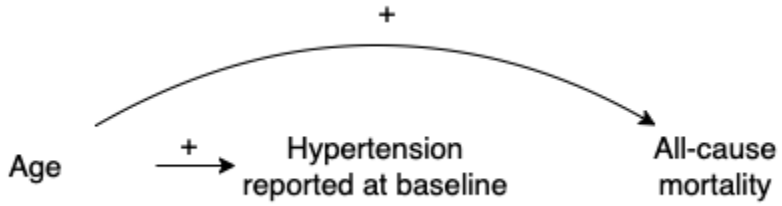
No, the crude CIR does not represent the causal CID because the crude CID does not control any confounders. For example, age is a confounder in the association between all-cause mortality and hypertension, then exchangeability does not hold.

NAME: ____Diego Liang_____

1. There is an association between age and hypertension reported at baseline in the study base, because the two-sample t-test performed shows there is statistically significant difference for age in different status of hypertension at baseline.
2. There is an association between age and all-cause mortality even in the unexposed (did not report hypertension at baseline), because the two-sample t-test performed shows there is statistically significant difference for age between whether death occurs.
3. Age is not the downstream consequence of hypertension reported at baseline or all-cause mortality.



6. Draw a Directed Acyclic Graph showing your assumptions about potential confounding by age. Use the (+) and/or (-) signs on the relevant arrows on your DAG to indicate the direction of the hypothesized associations.
    a. Based on your assumptions, what direction do you hypothesize confounding by age will bias the results of the crude analysis? Explain your answer.



The confounding by age will bias the results of the crude analysis upward, and the crude CID might be larger than the causal CID, because age is statistically significantly larger among those who reported hypertension at baseline and among those who die.

Part III. Crude analysis of case-control data

In this section, you will use data from a nested, density-sampled case-control study based upon the first 5 years of follow-up of participants in the Myocardial Infarction Onset Study. In this analysis, you will examine the relationship between having hypertension at the time of the baseline interview and death from cardiovascular causes within the first 5 years after enrollment. The variables in this dataset are described below:

Variable Name	Description
id	ID number
age	Age (continuous, years)
age_cat	Age Category (1: <50yrs, 2: 50-64 yrs, 3: 65+ yrs)
female	Female (1: female, 0: male)
married	Married (1: yes, 0:no)
educ	Educational Attainment (1: <HS, 2: HS, 3: >HS)
dm	Diabetes (1: yes, 0:no)
htn	Hypertension (1: yes, 0:no)
phys_activity	Frequency of Physical Activity (0: <1/wk, 1: 1-3/wk, 2: 4+/wk)
evermarj	Ever use marijuana (1: yes, 0:no)
case	Case-status (0: control, 1:case)

The dataset name is CV_death_Case_control and is available for download from the course website in several file formats including CSV, R, SAS and Stata. You may use SAS, STATA, R, the EPI202 calculator or any other statistical analysis software package of your choosing.

1. Read the dataset into your preferred software and complete the following 2x2 table:

	Diabetes at Baseline	
Death from Cardiovascular Causes	Yes	No
Cases	a=163	b=269
Controls	c=178	d=686

2. Is there a crude association between having diabetes at baseline and the incidence of death from cardiovascular causes in these case-control data?
   (a) State the null and alternative hypotheses
   (b) Calculate a valid test statistic
   (c) Calculate the p-value associated with the test statistic

$N_1 = a + c = 341$, $N_0 = b + d = 955$, $M_1 = 432$, $M_0 = 864$, $T = 1296$

H_0: There is no crude association between having diabetes at baseline and the incidence of death from cardiovascular causes. ($OR = 1$)

H_A: There is a crude association between having diabetes at baseline and the incidence of death from cardiovascular causes. ($OR \neq 1$)
Let X = number of exposed cases = $a = 163$

$$E(X|H_0) = \frac{N_1 M_1}{T} = \frac{341*432}{1296} = 113.667, Var(X|H_0) = \frac{M_1 M_0 N_1 N_0}{T^2(T-1)} = \frac{432*864*341*955}{1296^2(1296-1)} = 55.882$$

$$Z^2 = \frac{[X-E(X|H_0)]^2}{Var(X|H_0)} = \frac{(163-113.667)^2}{55.882} = 43.5517$$

$Pr(\chi^2 > 43.5517) < 0.0001$
We reject the null at the $\alpha = 0.05$ level. We conclude that there is statistically significant evidence for a crude association between having diabetes at baseline and the incidence of death from cardiovascular causes, assuming no confounding, selection bias, or information bias.

3. For the association between diabetes at baseline and cardiovascular mortality:
    i. Calculate the OR and interpret the numerical result in words
    ii. Calculate the 95% CI for the OR and interpret the numerical result in words

$\widehat{OR} = \frac{ad}{bc} = \frac{163*686}{269*178} = 2.3353$ over the 5-year follow-up period
The odds of death from cardiovascular causes for those having diabetes at baseline is 2.3353 times the odds of death from cardiovascular causes for those having no diabetes at baseline over the 5-year follow-up period, assuming no confounding, selection bias, or information bias.

The 95% confidence interval for the ln OR:

$$X \pm 1.96\sqrt{\widehat{Var}(X)} = ln(\widehat{OR}) \pm 1.96\sqrt{\widehat{Var}(ln(\widehat{OR}))} = ln(\widehat{OR}) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$= ln(2.3353) \pm 1.96\sqrt{\frac{1}{163}+\frac{1}{269}+\frac{1}{178}+\frac{1}{686}} = (0.5931,\ 1.1031)$$

So, the 95% confidence interval for the OR:

$$e^{(0.5931,\ 1.1031)} = (1.8096,\ 3.0136)$$

With 95% confidence, these data are consistent with odds ratios ranging from 1.8096 to 3.0136 for the association between diabetes at baseline and cardiovascular mortality over the 5-year follow-up period, assuming no confounding, selection bias, or information bias.