BIOST 536: Homework 3

Department of Biostatistics @ University of Washington

October 19, 2024

Designing a Descriptive Table

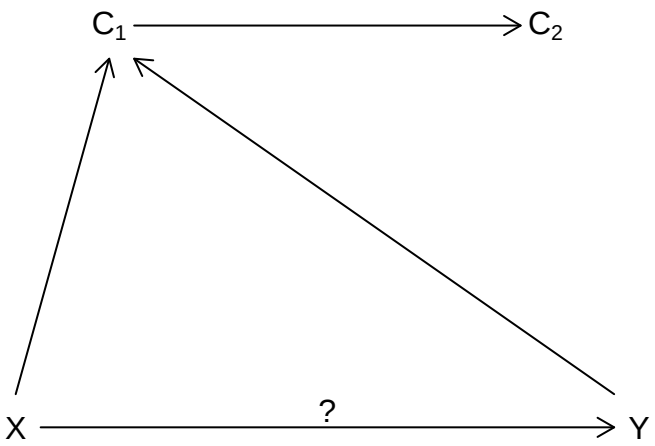
1. Your team has conducted a case-control study to examine the evidence that a (binary) exposure affects a disease, and you have analyzed study data. You are beginning to write a scientific article on your findings. The first table in your article will describe study participants by sharing descriptive statistics for key variables. How would you construct your table? Specifically, would you compute and report descriptive statistics on the whole sample (no stratification)? Or would you stratify on some variable and, if so, what variable? Write a short paragraph to explain your choice.

If it helps you to be more concrete, assume that D is a type of cancer, the exposure is a genetic variant, and the hypothesis is that having one or more copies of the genetic variant increases a person's risk of the cancer.

Characteristic	**Overall**, N = 10,000	**Disease**, N = 5,000	**No disease**, N = 5,000
E			
Exposed	3,548 (35%)	2,283 (46%)	1,265 (25%)
Unexposed	6,452 (65%)	2,717 (54%)	3,735 (75%)

Designing a DAG

2. Before it was understood that polycystic kidney disease (PKD) is a genetic disorder, Dr. Ott hypothesized that lead poisoning was a cause of PKD. In planning a study to collect evidence to study a possible effect of lead poisoning on PKD, Dr. Ott wonders whether glomerular filtration rate (GFR) is a confounder because prior work showed that GFR is associated with both lead poisoning and PKD. Suppose, in truth, lead poisoning is a cause of renal failure (the kidneys don't work as well as they should), affecting GFR. Similarly, PKD is a cause of kidney failure. Draw a DAG summarizing the information presented. Should Dr. Ott treat GFR as a confounder?



1	X	Lead poisoning
2	C_1	Kidney failure
3	C_2	GFR
4	Y	PKD

Case-control parameterization

3. Assume that $P(D)$ in a population is 10%. Investigators plan to conduct a case-control study (unmatched) to study associations between D and a binary exposure E and also collect data on a binary covariate C. Their analysis model will be

$$logit(p) = \beta_0 + \beta_1 \cdot C + \beta_E \cdot E + \beta^* \cdot C \times E$$

(a) The investigators will sample an equal number of cases and controls – 1:1 sampling. What is π, the ratio of sampling probabilities? What is the expected value of β_0 ? Write the expected value in terms of population parameters.

In a case-control setting, $\pi = \frac{P(Z=1|D=1)}{P(Z=1|D=0}$ defines a measure of how much more likely a case is to be sampled than a control.

(b) For every sampled case, investigators will sample 9 controls – 1:9 sampling. What is π, the ratio of sampling probabilities? What is the expected value of $\hat{\beta}_0$? Write the expected value in terms of population parameters. Comment on whether this expected value surprises you, or if you can make sense of the difference from (a).

4. Suppose you have data for 90 people on a continuous explanatory variable X and a binary outcome D, summarized in the following table.

X	D	Frequency
0	no	20
0	yes	10
1	no	15
1	yes	15
2	no	10
2	yes	20

(a) (Optional) Although it is not very informative, make a scatterplot of X and D (coded as no=0 and yes=1). Do your best to make the scatterplot informative (for example, "jitter" to show overlapping points). Also, make the plot comparable with the plot you will make for Q6A (e.g. same axes scales).

(b) Fit a simple logistic regression model of D on X. Write the fitted model.

(c) According to the fitted model, what is the probability of D for individuals with X=0? Why does this make sense?

(d) According to the fitted model, what is the probability of D for individuals with X=1? Why does this make sense?

(e) According to the fitted model, what is the probability of D for individuals with X=2? Why does this make sense?

(f) For each of X=0,1,2, estimate p, the probability of D, using the table above (not the logistic model). Complete the following table:

(g) Plot logit(p) against X. What do you notice?

5. Suppose you have data for 900 people on a continuous explanatory variable X and a binary outcome D, summarized in the following table:

X	D	Frequency
0	no	200
0	yes	100
1	no	150

X	D	Frequency
1	yes	150
2	no	100
2	yes	200

Fit a simple logistic model to these data. Compare and contrast the results with your results from Q4. Your comparison should include regression parameter estimates, standard errors, and confidence intervals.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.5	0.111	-6.25	0	0.401	0.62
X	2.0	0.087	8.00	0	1.690	2.37

We estimate that the odds of D is 2 times greater for X=1 compared to X=0 (95% CI for odds ratio: 1.69-2.374).

6. Suppose you have data for 150 people on an explanatory, continuous variable X and a binary outcome D, summarized in the following table:

X	D	Frequency
-1	no	30
0	no	20
0	yes	10
1	no	15
1	yes	15
2	no	10
2	yes	20
3	yes	30

(a) (Optional) Although it is not very informative, make a scatterplot of X and D (using standard 0/1 coding as in Q4). Do your best to make the scatterplot informative (for example, "jitter" to show overlapping points). Also, make your plot comparable with the plot from Q4A.

(b) Fit a simple logistic regression model of D on X. Write the fitted model.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.26	0.299	-4.50	0	0.139	0.451
X	3.84	0.204	6.61	0	2.658	5.946

We estimate that the odds of D is 3.843 times greater for X=1 compared to X=0 (95% CI for odds ratio: 2.658-5.946).

(c) According to the fitted model, what is the probability of D for individuals with X=0? Why is this different from Q4C?

(d) According to the fitted model, what is the probability of D for individuals with X=1? How does this compare with Q4D? Why does this make sense?

End of report. Code appendix begins on the next page.

Code Appendix

```
# clear environment
rm(list=ls())

# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-')
ndigits = 3
options(digits = ndigits)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# load relevant packages
library(dplyr)      # data frame manipulation
library(gtsummary)  # table summaries
library(dagR)       # DAG simulation
library(broom)      # model coefficient table
library(knitr)      # pretty tables

# load data

set.seed(1)
n <- 5000
pE <- 0.15  # (baseline) probability of exposure
data <- data.frame(D = sample(rep(0:1, n)))   # 1:1 case-control
data$E <- rbinom(n, 1, pE+(0.4*data$D))       # P(E|case) > P(E|control)

data %>%
  mutate(D = ifelse(D==1, "Disease", "No disease"),
         E = ifelse(E==1, "Exposed", "Unexposed")) %>%
  gtsummary::tbl_summary(by = D) %>%
  add_overall() %>%
  bold_labels()
dag <- dagR::dag.init(covs = c(1,1), arcs = c(0, 1, 1, 2, 3, 1),
                      y.name = "PKD", x.name = "Lead poisoning",
                      cov.names = list("Kidney failure","GFR")) %>%
  dag.draw()
data.frame(X = c(0,0,1,1,2,2),
           D = rep(c("no","yes"),3),
           Frequency = c(20,10,15,15,10,20)) %>%
  knitr::kable()
# data.frame(X = 0:2,
           # 'P(D)' = c(),
           # 'odds(D)' = c(),
           # 'log odds(D)' = c()) %>%
  # kable()
q5data <- data.frame(X = c(0,0,1,1,2,2),
                     D = rep(c("no","yes"),3),
                     Frequency = 10*c(20,10,15,15,10,20))
kable(q5data)
mod.lr <- glm(data=q5data, as.factor(D)~X, family=binomial(), weights=Frequency) %>%
  broom::tidy(exponentiate = TRUE, conf.int = TRUE) %>%
  mutate_if(is.numeric, round, ndigits)
```

```r
mod.lr %>% kable()
q6data <- data.frame(X = c(-1,0,0,1,1,2,2,3),
                     D = c("no", rep(c("no","yes"),3), "yes"),
                     Frequency = c(30,20,10,15,15,10,20,30))

q6data %>% kable()
mod.lr2 <- glm(as.factor(D)~X, binomial(), q6data, weights=Frequency) %>%
  broom::tidy(exponentiate = TRUE, conf.int = TRUE) %>%
  mutate_if(is.numeric, round, ndigits)

mod.lr2 %>% kable()
```

End of document.