Ch 6: Samples and Observational Studies

Disclaimer

There are sentences in this chapter I disagree with or think are poorly described. In Edition 4:

- pg 142: Observation vs. Experiment textbox and following paragraph
- pg 143: Definition of confounding and conceptual diagram of confounding
- pg 147: Description of bias

For these sections, the material presented in lecture takes precendence

Learning objectives for today

- To determine when treatment/exposure is manipulated or observed by an investigator
  - Experimental vs observational designs
- To define the population of interest
  - Target population
  - Study population
- To determine how samples are drawn from populations
  - Complete sample (census)
  - Types of random samples
  - Convenience samples
  - Volunteer samples

Observation vs. Experimentation

- Recall the three types of questions we can try to answer:
  - Description
  - Prediction
  - Causation/Etiology
- The book argue that experiments are "the only source of fully convincing data" to study causes and effect. We disagree.
- Epidemiologists, economists, and others often use observational data to study cause and effect and have developed a careful theory of causal inference that is sometimes less well understood by others.

Observation vs. Experimentation

- A study is observational if the investigator **observes** what happens and does not control treatment or exposure.
- A study is experimental if the investigator **manipulates** who is exposed or treated and who is not.

Observation vs. Experimentation

- Baldi & Moore present an example of observational data being misleading in the study of the causal effect of hormone replacement therapy on cardiovascular disease. However, these observational data, analysed in a different way yield the same conclusion as the randomized controlled trial. Link to more information.

Lurking variable a.k.a confounding variable

- Lurking variables/confounders are only important if you want to estimate whether some factor (often called the exposure or treatment) causes some outcome
- **There are no confounders for predictive or descriptive studies.** The concept is not applicable.

Lurking variable a.k.a confounding variable

- Suppose you are interested in the effect of coffee drinking on pancreatic cancer.
- You observe a positive association between coffee drinking and pancreatic cancer in a sample of individuals.
- What other factors might lead to pancreatic cancer? Are any of those factors associated with coffee drinking?
- If so, a crude comparison of coffee drinking and pancreatic cancer is **confounded** by those other factors and the observed relationship is biased.
- The estimate of the relationship is biased away from the *true causal estimate*

Confounding

- The association between an exposure and an outcome is confounded if there exist one or more variables that are causes of the outcome that are also associated with the exposure of interest.
- Example: Coffee drinking and pancreatic cancer
- Example: Community water flouridation and cavities among children
- Example from book: Alcohol and health response. What is wrong with the illustration?

Population of interest

Target Population

- The entire group of individuals about which we want information, and to whom we would like to apply our estimates and conclusions.
- Identifying the target population is part of setting up your "problem" in the PPDAC framework

Study population

- The part of the population from which we can actually select/recruit individuals and collect information.
- Sometimes, the study population is not a part of the target population.

Study sample

- Individuals who were selected/recruited and whom we gathered data about.
- We use a sample to draw conclusions about the entire population.

Example: Predictors of longevity

Brandts L, van den Brandt PA. Body size, non-occupational physical activity and the chance of reaching longevity in men and women: Findings from the Netherlands Cohort Study. *J Epidemiol Community Health*. Published Online First: 21 January 2019.

Introduction The rising number of obese and/or physically inactive individuals might negatively impact human lifespan. This study assessed the association between height, body mass index (BMI) and non-occupational physical activity and the likelihood of reaching 90 years of age, in both sexes separately.

Methods Analyses were conducted using data from the Netherlands Cohort Study. The NLCS was initiated in 1986 as a large prospective cohort study and included 120,852 men and women aged 55–69 years from 204 Dutch municipalities. Participants born in 1916–1917 (n=7807) completed a questionnaire in 1986 (at age 68–70 years) and were followed up for vital status information until the age of 90 years (2006–2007).

Example: Predictors of longevity

- Who is in the target population?
- Who is in the study population?
- To whom might we be trying to generalize these results?

Study design: Sampling

Study design: Sampling

- How a sample is chosen from the population.
- This is part of the "Plan" part of PPDAC
- When you are designing a study you need to decide how you will sample individuals (observations):
  - Who belongs to the target population?
  - How will you identify the study population?
  - How will you take a sample of the target population? Or, can you assess everybody (i.e, take a census)
  - How many people will you sample?
- Think about what you want your data frame to look like.

Representative(ness)

- Does the sample represent the population? Can we make conclusions based on the sample that will apply to the population as a whole?
- Represemtativeness is also called **external validity**.

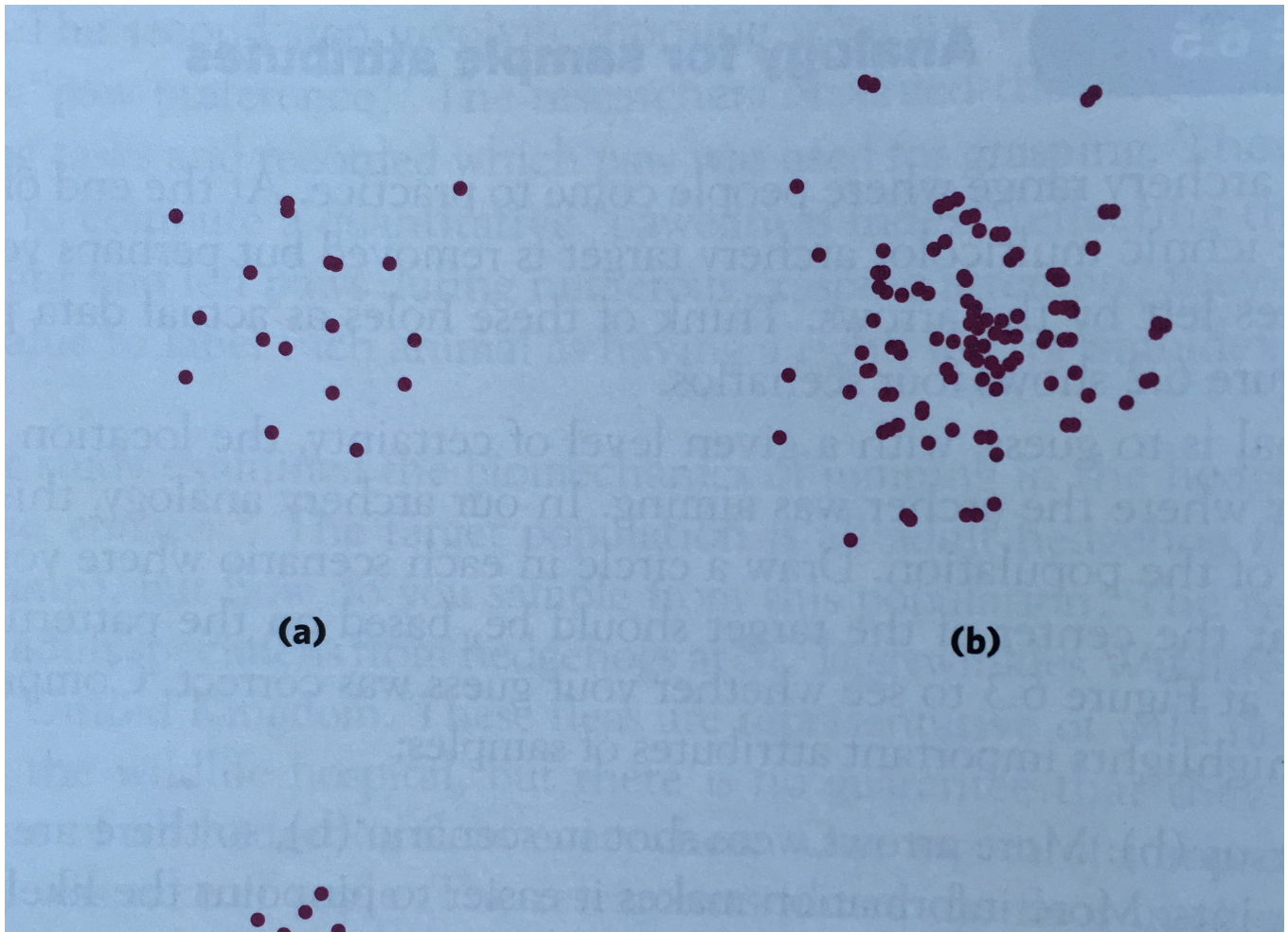
Inference

The process of drawing conclusions about a population based on a sample

Where is the target?

- Four samples are shown. Imagine guessing the location of the center of the population from the cetner of the sample distribution

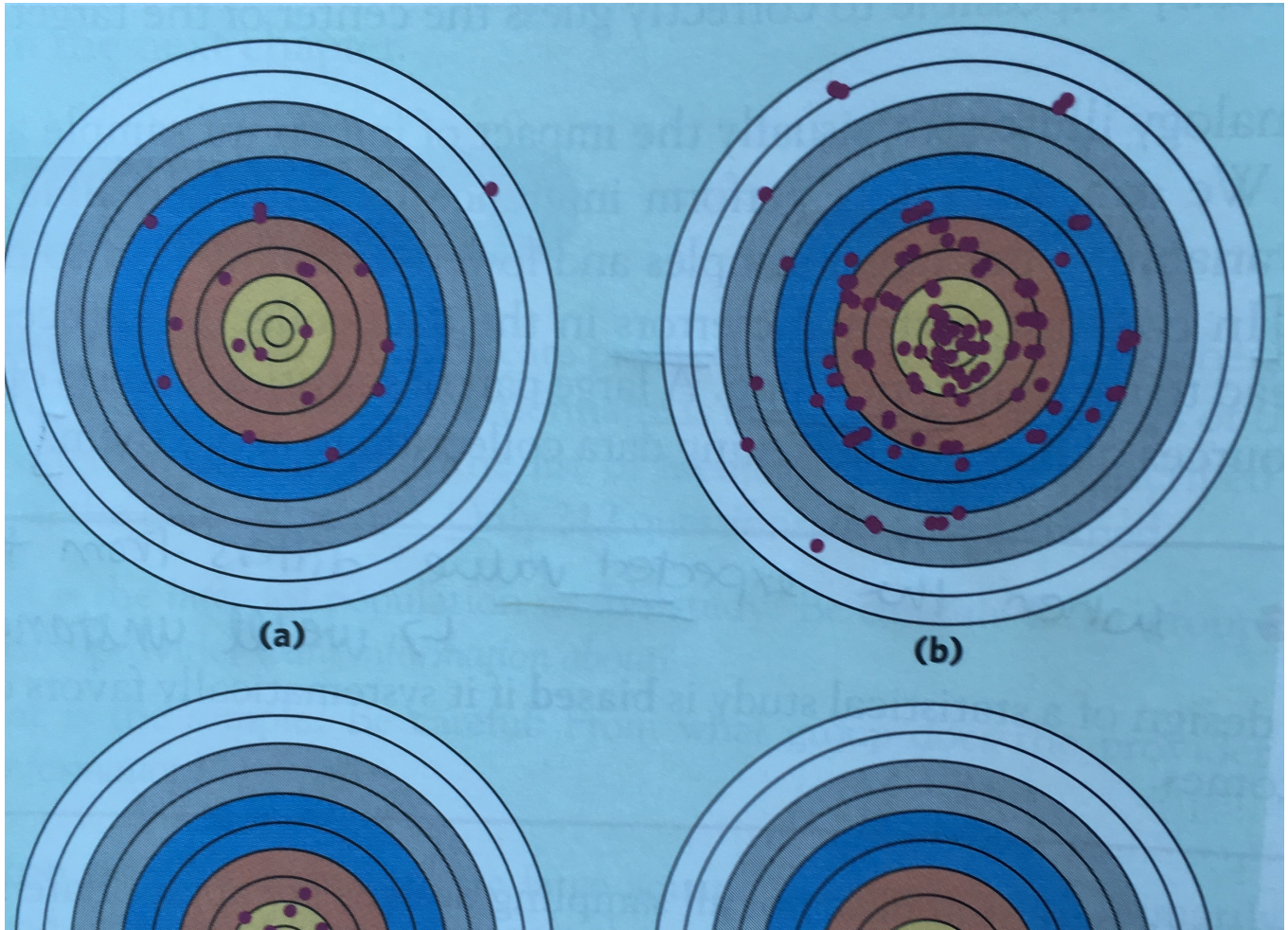
```
knitr::include_graphics("Ch6_random.jpg")
```



Where is the target?

- The samples are now superimposed on archery targets representing the target population. Sample (d) in this analogy is a biased sample

```
knitr::include_graphics("Ch6_target.jpg")
```



Where is the target?

Key takeaways from the previous figures:

1. A vs. B: More arrows make it easier to pinpoint the center of the target (assuming the archer has good aim)
2. A vs. C: Less variability makes it easier to pinpoint the center of the target
3. A vs. D: Systematic downward misses in scenario D bias our best guess of where the target is.

In real life we do not know where the target actually is, so we can't be sure whether our guesses are biased or not. But we can discuss the possibility of bias, and quantify how much bias would be required to negate a conclusion

Statistical bias

When the expected value based on a sample differs from the true underlying parameter value. We will understand this more as we learn more about what a parameter is.

Poor study designs

- Convenience samples
- Voluntary response samples
- Generally, non-probability designs

Good designs: Probability samples

- Simple random sample (SRS): A sample chosen by chance, where each individual in the data set has the same chance of being selected.
- We can easily choose SRS of data frames using R

Example of SRS in R

- First read in the hospital cesarean data
- What is the unit of analysis here?

```
library(readxl)
library(dplyr)
CS_data <- read_xlsx("Ch02_Kozhimannil_Ex_Cesarean.xlsx", s
CS_data <- CS_data %>% mutate(ID = row_number())
head(CS_data)
```

```
## # A tibble: 6 x 8
##   Births HOSP_BEDSIZE cesarean_rate lowrisk_cesarea~ ...
##    <dbl>         <dbl>          <dbl>           <dbl> <lg
## 1    767             1          0.344          0.107  NA
## 2    183             1          0.454          0.186  NA
## 3    668             1          0.430          0.195  NA
## 4    154             1          0.279          0.0844 NA
## 5    327             1          0.306          0.119  NA
## 6   2356             1          0.301          0.0662 NA
## # ... with 2 more variables: `Low Risk Cearean rate*100`
```

Example of SRS in R

```
CS_100_1 <- CS_data %>% sample_n(100)
CS_100_2 <- CS_data %>% sample_n(100)
```

Do you expect `head(CS_100_1)` to equal `head(CS_100_1)`?

Example of SRS in R

```
head(CS_100_1 %>% select(Births, HOSP_BEDSIZE, cesarean_rat
```

```
## # A tibble: 6 x 4
##   Births HOSP_BEDSIZE cesarean_rate    ID
##    <dbl>         <dbl>          <dbl> <int>
## 1   3783             3          0.287   468
## 2    664             3          0.324   359
## 3    216             1          0.231    38
## 4    657             3          0.311   311
## 5    346             3          0.329   494
## 6    431             2          0.281   192
```

```
head(CS_100_2 %>% select(Births, HOSP_BEDSIZE, cesarean_rat
```

```
## # A tibble: 6 x 4
##   Births HOSP_BEDSIZE cesarean_rate    ID
##    <dbl>         <dbl>          <dbl> <int>
## 1   2133             2          0.306   181
## 2    713             1          0.163   114
```

Example of SRS in R

Why are these first six lines different?

Anytime you do something *randomly* in R, the results will be different. This is a good thing! This allows you to pick many different random samples. In future weeks we will do this a lot.

Example of SRS in R

What if you want to ensure that you pick the same SRS as a friend?

Then you need to specify the same seed in the `set.seed()` function:

```
set.seed(123)
CS_100_1 <- CS_data %>% sample_n(100)

set.seed(123)
CS_100_2 <- CS_data %>% sample_n(100)

identical(CS_100_1, CS_100_2)
```

```
## [1] TRUE
```

SRS a fraction in R

```
CS_5percent <- CS_data %>% sample_frac(0.05)
```

Proportionate Stratified sampling in R

- Group *and then* sample the same fraction from each group

```
CS_10percent_grouped <- CS_data %>%
  group_by(HOSP_BEDSIZE) %>%
  sample_frac(0.1)
```

- Proportionate stratified SRS assembles a sample that maintains the relative proportions of `HOSP_BEDSIZE` in the chosen sample compared to the population

Proportionate Stratified sampling in R

How to check you really did sample 10% of each HOSP_BEDSIZE group?

First see how many hospitals fall into each category

```
CS_data %>% group_by(HOSP_BEDSIZE) %>% tally()
```

```
## # A tibble: 3 x 2
##   HOSP_BEDSIZE     n
##          <dbl> <int>
## 1            1   131
## 2            2   179
## 3            3   270
```

Then calculate 10% of each tally and calculate the sum to see if it is the same as the sample size you end up with after you ran the previous code

```
CS_data %>% group_by(HOSP_BEDSIZE) %>%
  tally() %>%
  mutate(ten_per = n * 0.1) %>% # how did I know this was
```

Disproportionate Stratified sampling in R

- When might you want to over represent certain groups?
- Example: Estimating infant mortality by race/ethnicity when some race/ethnic groups are very small (e.g., indigenous groups in U.S./Canada)
- Then, you may want to over sample certain groups so you can better estimate infant mortality in those groups than if you sampled proportionately

Multistage sampling

- SRSs inside SRSs
- For example, sampling schools using a SRS, and then sampling students within those schools.

Sample surveys, cohorts, and case-control studies

- We won't cover this information in class, but you should read these pages of the textbook (pg 152-9) or learn the concepts in your epidemiology course.