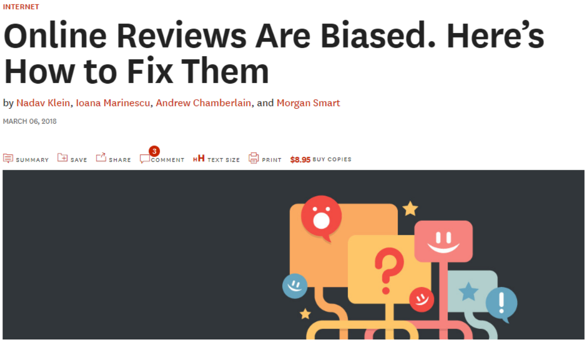
Chapter 6: Samples and Observational studies

Mi-Suk Kang Dufour

September 14, 2020

Statistics is Everywhere

Statistics is Everywhere



INTERNET

Online Reviews Are Biased. Here's How to Fix Them

by Nadav Klein, Ioana Marinescu, Andrew Chamberlain, and Morgan Smart

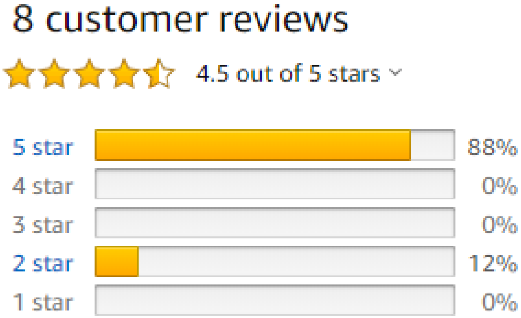
MARCH 06, 2018

Statistics is Everywhere

- from *Harvard Business Review*, March 2018:
  Online reviews are a powerful tool for sharing information at scale. But it's important to remember the source — many online reviews today are from those who've voluntarily decided to share opinions, giving a distorted view of products, services and companies.

Statistics is Everywhere

- What happens to the distribution of responses when all observations are from people who volunteer to participate?
- What kind of a variable is being diplayed here?
- What kind of a visualization is this?



8 customer reviews

★★★★☆ 4.5 out of 5 stars

5 star 88%

4 star 0%

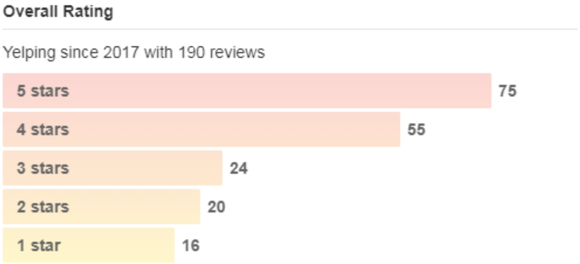
3 star 0%

2 star 12%

1 star 0%

Statistics is Everywhere

What is the sample size here?
What do we notice about this distribution?

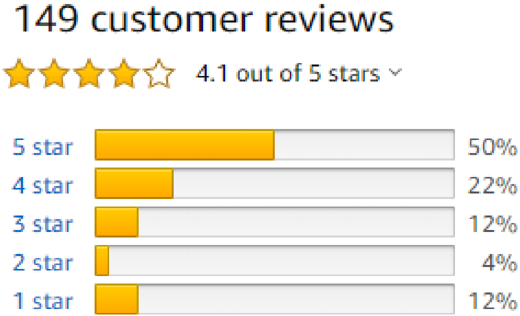


Overall Rating

Yelping since 2017 with 190 reviews

5 stars	75
4 stars	55
3 stars	24
2 stars	20
1 star	16

Statistics is Everywhere



149 customer reviews

★★★★☆ 4.1 out of 5 stars

5 star	50%
4 star	22%
3 star	12%
2 star	4%
1 star	12%

Statistics is Everywhere

- More from the article:
  "That problem generalizes to most online reviews. Research shows many of today's most popular online review platforms — including Yelp business reviews, and Amazon product reviews — have a distribution of opinion that is highly polarized, with many extreme positive and/or negative reviews, and few moderate opinions. This creates a "bi-modal" or "J-shaped" distribution of online product reviews that has been well-documented in the academic literature. This makes it hard to learn about true quality from online reviews.""

Important Note

There are several definitions and descriptions in the book chapter that I disagree with or think are poorly described. Particularly 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 the concepts around sampling and observational studies and study design in general, for these sections, and for the definitions of causality and confounding, the material presented in lecture takes precendence

Learning objectives for today

Defining a study by:

- Whether the treatment or exposure is controlled by an investigator
  - Experimental vs observational designs
- The population of interest
  - Target population
  - Study population
- How the sample was drawn from the population
  - Complete sample (census)
  - Random sampling
  - Convenience sampling
  - Volunteer sampling
- Was selection conditional on exposure or outcome?
- Method and timing of data collection

Who or what controls exposure?

Types of problems: from our PPDAC framework

- Recall the three types of problems we can attempt 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, sociologists, political scientists, 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

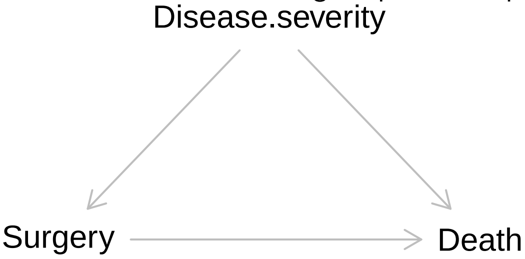
- A study is observational if we are observing what happens and do not have control over the treatments or exposures.
- A study is generally considered experimental if the investigator is experimenting by controlling who is getting the exposure(treatment) and who is not.

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.

Observation vs. Experimentation

- Suppose we are interested in whether a certain surgical procedure prolongs life in cancer patients:



- What is disease severity in this DAG?
- How does this DAG change if the surgical procedure is randomly assigned?

Lurking variables a.k.a. confounders

- Lurking variables/confounders are only important if you are asking a causal/etiologic question.
- There are no confounders for predictive studies

Other ways exposure is assigned

- Pre and post designs
- Randomized roll out (stepped wedge)
- Natural experiments

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 our PPDAC framework

Study population

The part of the population from which we can actually select/recruit individuals and collect information.

Study sample

Individuals we were able to select/recruit and gather data from.

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

Think about this abstract:

- what is the target population?
- What is 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 you will identify the study population (sometimes called identifying a sampling frame)
  - How will you take a sample of the target population. Or whether you can assess everybody (a census)
  - How many 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?
- Representativeness is also called external validity.

Study sampling designs that generally do not give representative samples

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

More representative 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

```
## # A tibble: 6 x 8
##   Births HOSP_BEDSIZE cesarean_rate lowrisk_cesarea~ ...5  `Cesarean rate
##    <dbl>        <dbl>         <dbl>            <dbl> <lgl>            <dbl
## 1    767            1         0.344           0.107  NA                34.
## 2    183            1         0.454           0.186  NA                45.
## 3    668            1         0.430           0.195  NA                43.
## 4    154            1         0.279           0.0844 NA                27.
## 5    327            1         0.306           0.119  NA                30.
## 6   2356            1         0.301           0.0662 NA                30.
## # ... with 2 more variables: `Low Risk Cearean rate*100` <dbl>, ID <int>
```

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_2)`?

Example of SRS in R

```r
head(CS_100_1 %>% select(Births, HOSP_BEDSIZE, cesarean_rate, ID))
```

```
## # A tibble: 6 x 4
##   Births HOSP_BEDSIZE cesarean_rate    ID
##    <dbl>        <dbl>         <dbl> <int>
## 1    370            2         0.324   243
## 2    505            3         0.384   419
## 3    392            1         0.245    85
## 4    300            2         0.36    216
## 5    381            1         0.362    15
## 6    302            2         0.245   271
```

Example of SRS in R

```r
head(CS_100_2 %>% select(Births, HOSP_BEDSIZE, cesarean_rate, ID))
```

```
## # A tibble: 6 x 4
##   Births HOSP_BEDSIZE cesarean_rate    ID
##    <dbl>        <dbl>         <dbl> <int>
## 1   1341            3         0.301   478
## 2    109            1         0.385    13
## 3    100            1         0.37     39
## 4   1100            2         0.472   147
## 5    427            1         0.464    99
## 6   4810            3         0.254   328
```

Example of SRS in R

```r
identical(CS_100_1, CS_100_2)
```

```
## [1] FALSE
```

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 use `set.seed()`:

```r
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

```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)
dim(CS_10percent_grouped)
```

```
## [1] 58  8
```

- 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 in the original data

```
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
```

Proportionate Stratified sampling in R

then in the sample

```
CS_10percent_grouped%>%group_by(HOSP_BEDSIZE) %>% tally()
```

```
## # A tibble: 3 x 2
##   HOSP_BEDSIZE     n
##          <dbl> <int>
## 1            1    13
## 2            2    18
## 3            3    27
```

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.

Conditional selection?

- Was selection conditional on exposure or outcome?
  - If we are choosing people to participate in our study based on their exposure status this is generally a cohort design
  - If we are choosing people to participate in our study based on their outcome status this is generally a case-control design
  - Matched pairs are a special case of study design where participants are chosen conditional on multiple factors
- We will not go into much detail here about these types of designs, however you should know that if we selected participants conditional on exposure then the marginal distribution of exposure is not meaningful, likewise if we selected participant conditional on outcome then the marginal distribution of outcome is not meaningful.

Study design: Data collection

Study design: Data collection

- Part of your study design is also about how you will collect the variables:
  - By survey? (self reported information)
  - Measured by device?
  - Collected from health records?
  - Google search?
- What are the types of variables and levels of each variable?
  - Continuous?
  - Categorical? How many categories? Ordinal or nominal?
- Think about what you want your data frame to look like.

Temporality

Temporality

- Method and timing of data collection
  - One timepoint only (cross sectional)
  - Multiple timepoints (longitudinal)
  - Moving forward in time (prospective)
  - Collecting data from previous times (retrospective)

Putting it together

Reading an article title:
What does the title tell you about the design? About what the target population might be? What would you ask about the methods?

From the American Journal of Epidemiology Volume 188, Issue 2, February 2019

- Associations of a Healthy Lifestyle Index With the Risks of Endometrial and Ovarian Cancer Among Women in the Women's Health Initiative Study
- Death and Chronic Disease Risk Associated With Poor Life Satisfaction: A Population-Based Cohort Study

and from New England Journal of Medicine 2019; 380:415-424

-Partial Oral versus Intravenous Antibiotic Treatment of Endocarditis a randomized, noninferiority, multicenter trial

Parting humor

From PhDcomics.com