Section 4: Evaluating the quality of data and their statistics

STOR 155.02, Spring '21

updated 2021-02-09

What you will learn

- data collection
- experimental vs. observational data
- correlation, causality and 'confounding/lurking' variables
- working with surveys
- summarizing subcategories

Resources

- Textbook ch 1.3-1.5, 1.10 again for categorical variable concepts

Heads up

This is kind of a long lecture

with a bit more to cover than usual so far

I will go faster

but stop me for questions

and be prepared to look back at the slides and video

you can do it

Data collection, and quality of statistical statements

Review: Big picture

Our perspective: using data to learn about some aspect of the world

Often trying to answer a specific question

Examples from class so far:

distribution of likes for fake news facebook posts?

highly liked fake news posts common or rare?

land-grant universities still profiting from Morrill Act?

trends in female life expectancy?

Review: Some terms

population: group of people/places/things you hope to learn about

national economies, elderly U.S. residents, north atlantic right whales, cat pictures online ...

sample: subset for which you have data.

sometimes sample = population!

observation: unit of a single element in sample data

These distinctions aren't always as clear and neat as the textbook would like you to think.

Examples from class so far:

Question: distribution of likes for fake news facebook posts?

Population: fake news posts on facebook

Sample: dataset of posts Buzzfeed identified as fake news in 2018

Question: trends in women's health / wellbeing in selected countries?

Population: selected countries

Sample: female life expectancy at birth from 1960-2018 in those countries

Goal is to evaluate:

a. How well do the sample observations represent the population?

b. Does the sample data support my conclusions, or is there something else to explain what I see?

c. Are there problems with my data, such as missing values, that prevent me from adequately answering the question?

Focus on three types of problems:

1. sample data inappropriate: sample does not represent the population, data inadequate to answer the question at hand

2. missing or problematic values

3. confouding/lurking variables

Example: The story of a statistic

> We hear Democrats talking about the middle class and workers ... [but] if only 39 percent of this country can afford a thousand-dollar emergency, we must use the word poverty.
>
> ------ Rev. William Barber II on NPR, Jan. 31 2021

Is this true?

or at least as close to truth you can get in a soundbite?

What is it based on?

First step: What's the reference?

closest I found to what Barber could be talking about was

> A similar statistic came out of the 2018 Federal Reserve Report on the Economic Well-Being of U.S. Households, finding that **40% of adults in this country could not afford a $400 emergency**
>
> ----- Numbers behind the Moral Budget, Repairers of the Breach blog post (a Barber-affiliated advocacy group)

citation:

2018 Federal Reserve Report on the Economic Well-Being of U.S. Households

Step two: Get the data the stat is based on

About the data

report based on the Fed's Survey of Household Economics and Decisionmaking (SHED)

it's a long questionnaire asking about households' feelings about the economy, work, financial stability

What I got

2019 vintage of the survey

the most recent year with complete results

What is the population? (you answer this)

What is the sample?

Good surveys will have a detailed description of how the sample of respondents (whose responses make up the dataset) was collected

In this case, the sample of about 12,000 respondents is 'designed to be representative of adults age 18 and older living in the United States.'

to understand the details, we need to defer to the next section when we discuss randomness.

The Fed gives details here.

How you get your sample matters

[one obs. … … … … … … … … … … whole pop.]

WORST BEST

this isn't very helpful though

Rules of thumb for sample data

GOOD

controlled experiment --- e.g. data from measurements in a lab

'random' --- e.g. observations determined by rolling a (many-sided) die

details on what randomness means next section

BAD

'anecdotal' sample

e.g. ask my friends about fake news they've seen on facebook

'biased' sample where observations are chosen (or self-select) because of certain characteristics

e.g. *all online review systems* such as yelp, RMP, Amazon reviews...

good if the reviewer shares your values, bad for calculating statistics like averages

SHED survey: What are the variables?

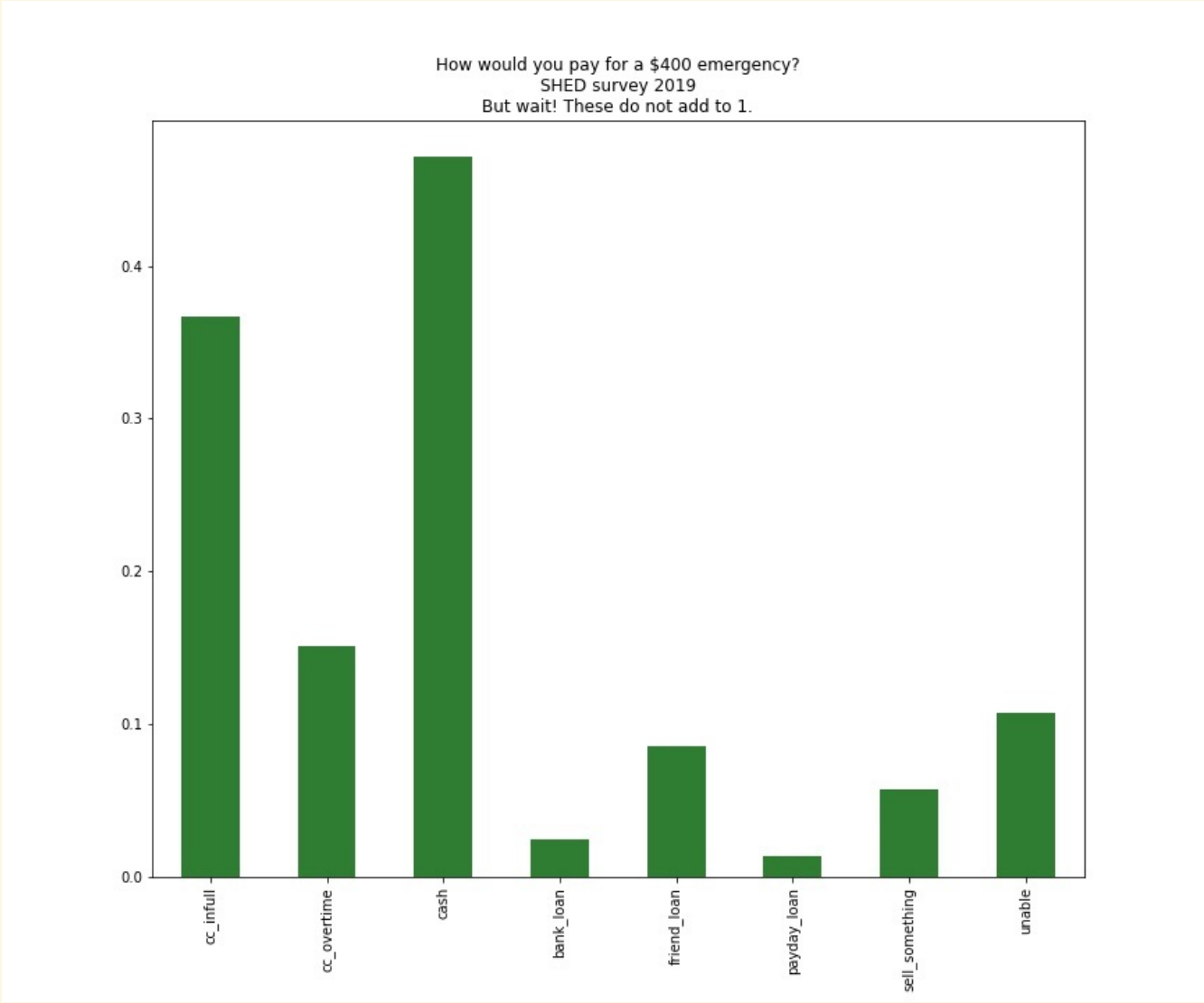
Full list in the codebook

Class dataset picks out only a few. See the related notebook on github

Today, interested in responses to

> Suppose that you have an emergency expenses that costs $400. Based on your current financial situation, how would you pay for this expense?

'afford' to pay it if using cash or card paid off next month

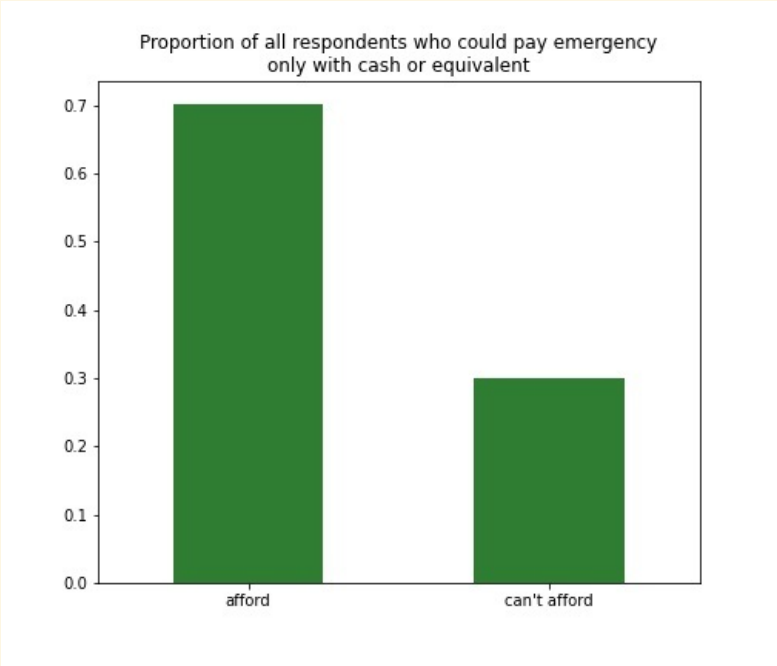

How would you pay for a $400 emergency?
SHED survey 2019
But wait! These do not add to 1.
0.0
0.1
0.2
0.3
0.4
cc_infull
cc_overtime
cash
bank_loan
friend_loan
payday_loan
sell_something
unable

From the plot, it looks like almot 90 percent *can* afford to pay

but

Data quality wrinkle: Some people give multiple answers for how they would pay.

What should we do?


Proportion of all respondents who could pay emergency only with cash or equivalent
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
afford
can't afford

Takeaways

about 30 percent could not 'afford' a $400 emergency in 2019

but when the Fed included data from April 2020, the number rose to 37 percent

in my *opinion*, data do support Barber's basic point

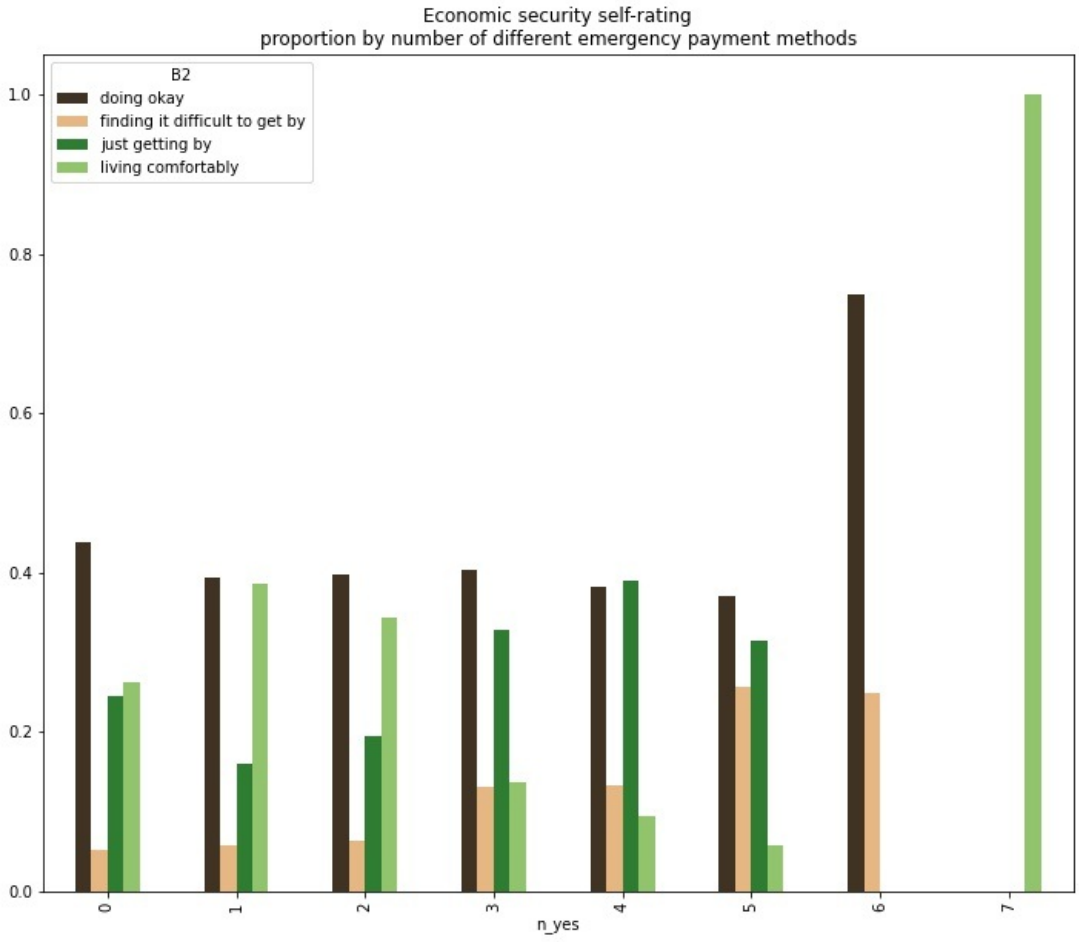
that about a third of Americans would need to go into debt for even a small financial emergency

and this is consistent with the Fed's own interpretation of this data

Why not just remove observations that list more than one way to pay?

this sub-sample of respondents differs from the original sample in a meaningful way

people rating themselves as worse off make up larger proportions of people paying multiple ways


Economic security self-rating
proportion by number of different emergency payment methods
B2
doing okay
finding it difficult to get by
just getting by
living comfortably
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
n_yes

These ideas are most relevant for

surveys (e.g. Census),

scientific observation (sample collected in the field)

and **medical studies**

Doesn't make too much sense for talking about the life expectancy data, for example

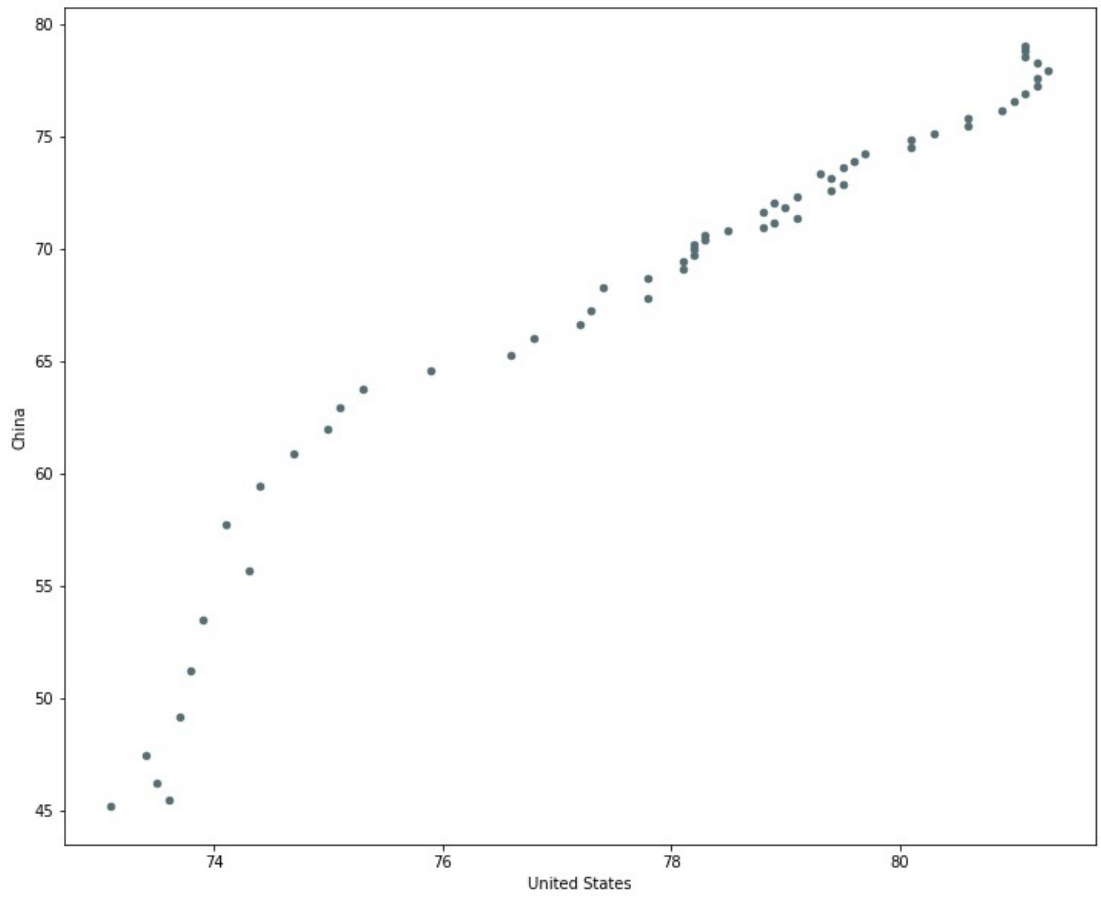
But it always is relevant to think about how your data were generated!

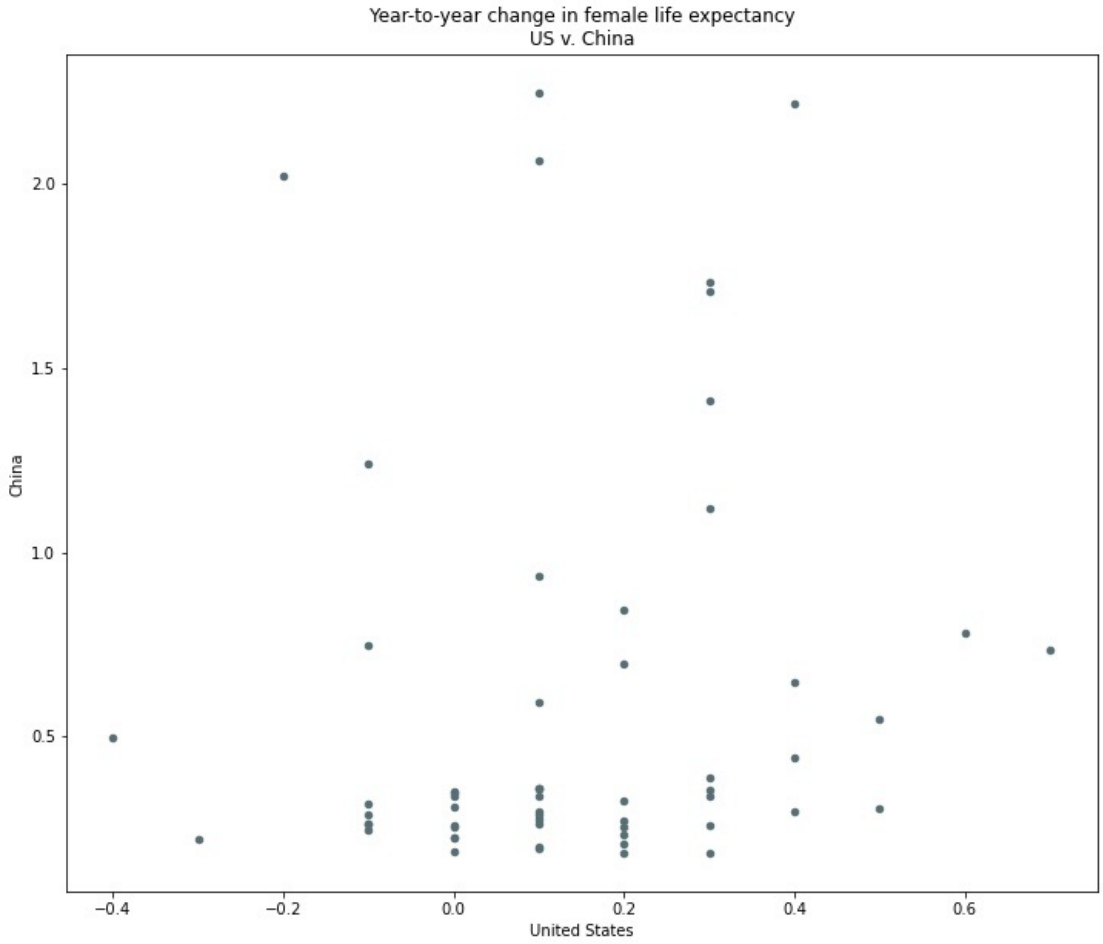
Confounding variables

Remember the plot of US and Chinese female life expectancy from last time / homework

r = Corr(US, China) was almost perfect!

but the correlation in year over year changes in the US and China had correlation near zero!


80
75
70
65
60
55
50
45
China
74
76
78
80
United States


Year-to-year change in female life expectancy
US v. China
China
United States
2.0
1.5
1.0
0.5
−0.4
−0.2
0.0
0.2
0.4
0.6

This happens because both US and China life expectancies are highly positively correlated with time!

Confounding variable

Z is counfounding for evaluating the relationship between X and Y

if Z is strongly correlated with both X and Y

so that correlation between X and Y

is really just because of their correlation with Z

e.g. time is confounding

for China and US life expectancies

in this case can see this by removing the effect of time with year-to-year changes

This is why when evaluating relationships between variables with correlation

it's important to talk about positive/negative relationships

and not about one variable *causing* another to increase/decrease

PollEv.com/brendanbrown849

poll closes at _____

don't be late!


Five more minutes