In-class Discussion: Understanding Datasets

DSPC7514 Data Analysis for Policy Research

Instructor: Harold Stolper

Contents

Describing datasets

How was the data produced?

Some common types of observational data:

- **Survey data:** survey responses from a sample of individuals
- **Administrative data:** administrative records from a government agency or organization

Cross-sectional vs panel data

A **cross-section** of data is just a set of observations for one period of time. For example, a set of evaluations from students for a particular course in a given semester. You can imagine a second cross-section of data: a second set of evaluations from a different group of students for the same course but in a different semester.

Multiple cross-sections of data can be combined into a single dataset (repeated cross-sectional data), but that is different from **panel data** (also called longitudinal data) where the *same individuals* or entities are observed in each period.

‘Unit of observation’ and ‘population represented by the sample’

It might seem obvious if you’re working with familiar data, but in many cases it may not be immediately obvious: what is the **unit of observation** or **unit of analysis** in the dataset you’re working with, and what population (or sampling frame) are observations drawn from? In other words, what information distinguishes one row from the next: does each row represent a unique person, household, province, country, etc.? Or maybe a household-month or state-year pair, as examples of the unit of observation for hypothetical panel datasets.

Many large government surveys are based on a sample drawn at random from some broader population (e.g. a sample of working adults). Survey data generally requires you to apply **sampling weights** in order to compute statistics that are representative of a broader population of interest (e.g. all adults in China aged 18+). This is because the likelihood of a subject responding to a survey is not random, but in fact depends on other variables such as age, gender, education, etc.

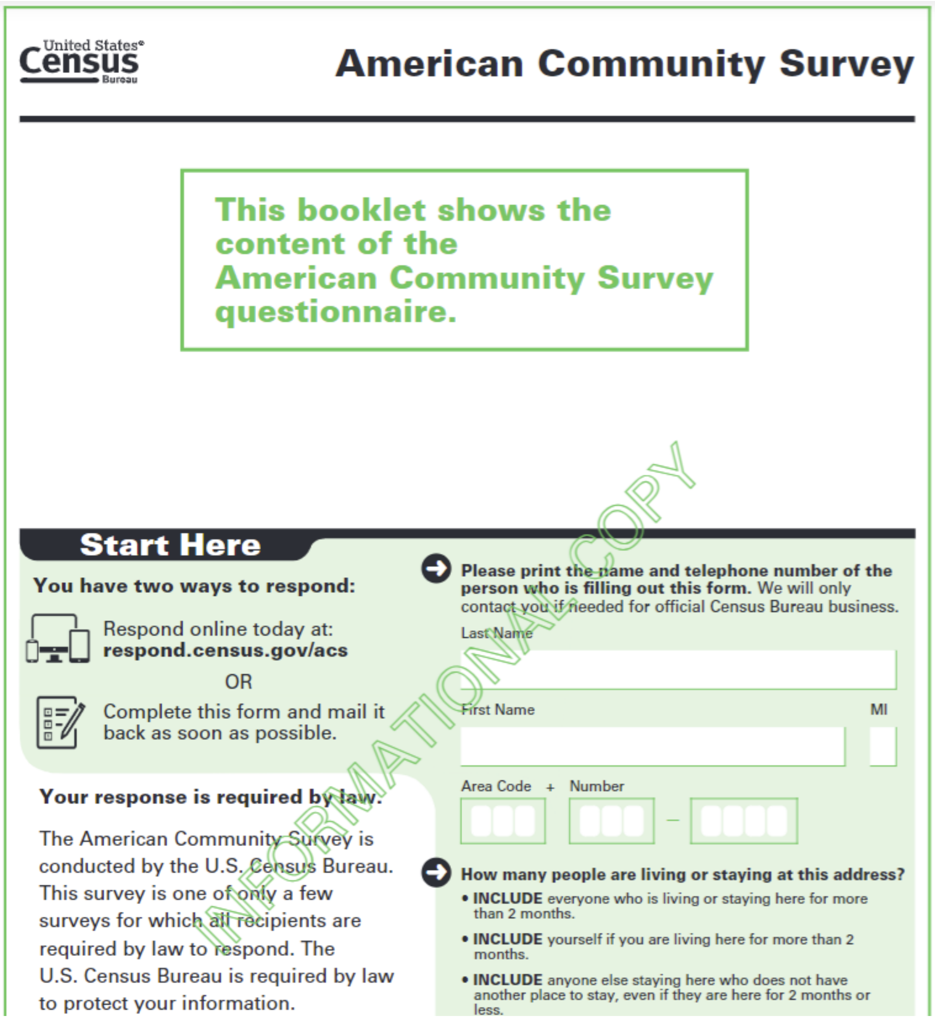
Sampling weights serve to re-weight the data so people who were more likely to respond count less when computing statistics (and people who were less likely to respond count more). For survey data, you typically must use weights or else the sample won’t be representative of the true population of interest in terms of age, race, gender, etc.

Whether or not weights are provided/necessary, you should make sure you can describe the **population represented by the sample** of data that you are working with.

Discussion question: *Which respondents are more likely to respond to a telephone survey?*

Application: American Community Survey

The American Community Survey is a nationwide survey conducted by the US Census Bureau that gathers information on social, economic, housing, and demographic characteristics of the population. The Census Bureau contacts over 3.5 million households each year to participate in the survey. The resulting data is used to track changes in communities and inform policy decisions about the allocation of resources. Let’s take a closer look at the some of the documentation to better understand this data source and what it can tell us.



United States® Census Bureau

American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

Start Here

You have two ways to respond:

Respond online today at: **respond.census.gov/acs**

OR

Complete this form and mail it back as soon as possible.

Your response is required by law.

The American Community Survey is conducted by the U.S. Census Bureau. This survey is one of only a few surveys for which all recipients are required by law to respond. The U.S. Census Bureau is required by law to protect your information.

Please print the name and telephone number of the person who is filling out this form. We will only contact you if needed for official Census Bureau business.

Last Name

First Name MI

Area Code + Number

How many people are living or staying at this address?

- **INCLUDE** everyone who is living or staying here for more than 2 months.
- **INCLUDE** yourself if you are living here for more than 2 months.
- **INCLUDE** anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.

Sampling frame

The universe for the ACS consists of all valid, residential housing unit addresses in all county and county equivalents in the 50 states, including the District of Columbia that are eligible for data collection. Beginning with the 2018 sample, we restricted the universe of eligible addresses further to exclude a small proportion of addresses that do not meet a set of minimum address criteria.

The Master Address File (MAF) is a database maintained by the Census Bureau containing a listing of residential, group quarters, and commercial addresses in the U.S. and Puerto Rico. (Source)

Discussion question: based on the above description, can you think of any families/individuals that are not represented by this sample?

Sample design

The Census Bureau uses a two-phase, two-stage sample design:

Phase 1: The Census Bureau selects two separate address samples, Period 1 and Period 2, at different times.

> Phase 2: The Census Bureau randomly assigns most of the sample addresses to one of the 12 months of the sample year. Addresses in remote Alaska are assigned to either January or September. The Census Bureau then selects the CAPI sample.
>
> The Census Bureau mails questionnaires to about 295,000 addresses each month, giving each address a 1-in-480 chance of being selected. Addresses are not selected more than once every five years. If a household doesn't respond within six weeks, the Census Bureau may try to contact them by phone. If that doesn't work, they may visit a sample of the remaining addresses for an in-person interview. There is a separate process for people who live in group quarters, such as college dorms, nursing homes, or military barracks. (Source)

Discussion question: based on the above description, does it sound like there is random sampling from the broader population (sampling frame)? How would you describe the population represented by the ACS sample?

Weighting overview

> The basic estimation approach is a ratio estimation procedure that results in the assignment of two sets of weights: a weight to each sample person record, both household and group quarters persons, and a weight to each sample housing unit (HU) record. Ratio estimation is a method that takes advantage of auxiliary information (in this case, population estimates by sex, age, race, and Hispanic origin, and estimates of total HUs) to increase the precision of the estimates as well as correcting for differential coverage by geography and demographic detail. This method also produces ACS estimates consistent with the population estimates from the Population Estimates Program (PEP) of the Census Bureau by these characteristics and the estimates of total HUs for each county in the United States.
>
> For any given tabulation area, a characteristic total is estimated by summing the weights assigned to the people, households, families, or HUs possessing the characteristic. Estimates of population characteristics are based on the person weight. Estimates of family, household, and HU characteristics are based on the HU weight. As with most household surveys, weights are used to bring the characteristics of the sample more into agreement with those of the full population by compensating for differences in sampling rates across areas, differences between the full sample and the interviewed sample, and differences between the sample and independent estimates of basic demographic characteristics (Alexander, Dahl, & Weidman, 1997). (Source)

Sampling error

> Sampling error is the difference between an estimate based on a sample and the corresponding value that would be obtained if the entire population were surveyed (as for a census). Note that sample-based estimates will vary depending on the particular sample selected from the population. Measures of the magnitude of sampling error reflect the variation in the estimates over all possible samples that could have been selected from the population using the same sampling methodology.
>
> Estimates of the magnitude of sampling errors – in the form of margins of error – are provided with all published ACS data. The Census Bureau recommends that data users incorporate margins of error into their analyses, as sampling error in survey estimates could impact the conclusions drawn from the results. (Source)

The margin of error (MOE) conveys the amount of sampling error in the results of a survey (for a given variable). At a certain confidence level (e.g. 95 percent), the MOE is equal to half the confidence interval. Below is the formula for the margin of error with 95 percent confidence. Note how it depends on the standard normal distribution to obtain the appropriate z-value (here, 1.96).

$$MOE_{95} = 1.96 * \sqrt{\sigma_n/n}$$

Application: Palestinian opinion polling

People stand amid rubble in the Jabalia refugee camp, northern Gaza Strip, on July 21.

Mahmoud Zaki/Xinhua News Agency via Getty Images

The Palestinian Center for Policy and Survey Research (PCPSR) conducts periodic opinion polls to promote "a better understanding of Palestinian domestic and international environment." Located in Ramallah in the West Bank, their polls "examine public opinion in both the Gaza Strip and the West Bank, focusing on three main issues: governance in Palestinian society, confidence in the two-state solution and Palestinians' attitudes toward violent struggle against Israel" (see this recent NPR article).

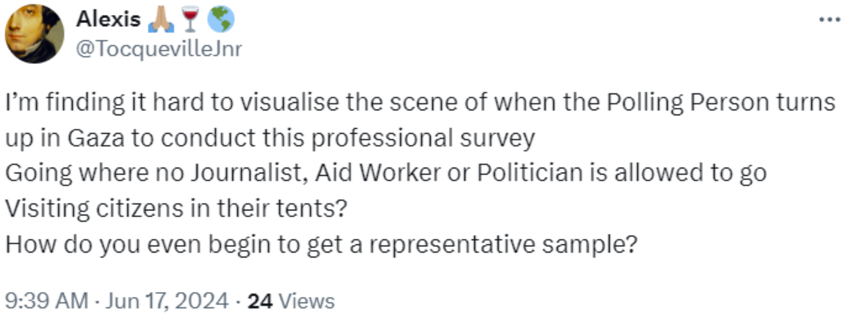
Below are excerpts from PCPSR about their survey methodology for PCPSR's data collection in Gaza:

> As we did in our previous poll three months ago, 75 communities were selected from residents of Rafah, Khan Younis, Al-Mawasi, Deir al-Balah and other areas in the central Gaza Strip and from the displaced people who were sheltering in those areas under the instructions of the Israeli army, so that these communities were either "counting areas," according to the classification of the Palestinian Bureau of Statistics, as was done in Rafah, some areas of Khan Younis and the central Gaza Strip, or displaced communities in built-up shelters, which are schools and other institutions affiliated with the government or UNRWA, or tent gatherings located in the areas of Rafah, Khan Younis, Al-Mawasi and the central Gaza Strip. The sample was drawn according to the following methodology:
>
> 1) In the **"counting areas"** specified by the Palestinian Bureau of Statistics, where the number of these areas reached 29.
> 2) In the **built-up shelters**, a regular random sample was withdrawn from the lists of these centers that were obtained, representing all the shelter centers in western Rafah, Deir al-Balah and other areas in central Gaza Strip, Rafah and Khan Younis areas, and the number of these areas reached 20.
> 3) In the **tent gatherings** in the areas of Rafah, Khan Younis, Al-Mawasi and the central Gaza Strip, where satellite maps showing the locations of these communities were relied upon. These areas were divided into blocks and a regular random sample of 26 blocks was drawn.

> In each “counting area”, built-up shelter, or tent gathering, 10 people were randomly selected for interviews while taking into account gender and age distribution. Refusal to conduct interviews was 9%.
>
> It is worth noting that 51% of the public in the Gaza Strip say they were displaced to their current location, where they were interviewed, because of the Israeli invasion of Rafah starting on May 6, while the remaining 49% say they were not displaced to their current location because of that particular attack. (Source)

> To ensure the safety of our data collectors in the Gaza Strip, interviews were conducted with residents in specific areas where no active combat was present. The areas covered included parts of the Rafah and Khan Younis areas and the central Gaza Strip and all shelters therein, but not the northern besieged enclave and other areas of combat in the central Gaza Strip and in the eastern area of Rafah. . .
>
> The sample size of this poll was 1570 adults, of whom 760 were interviewed face-to-face in the West Bank (in 76 residential locations) and 750 in the Gaza Strip (in 75 locations). Due to the uncertainty about the exact population size and distribution at that moment in the Gaza Strip, we almost doubled the sample size in that area in order to reduce the margin of error. The total sample was reweighted to reflect the actual relative size of the population in the two Palestinian areas. Thus, the sample used is representative of the entire populations of the two regions. The margin of error stands at +/-3%. (Source)

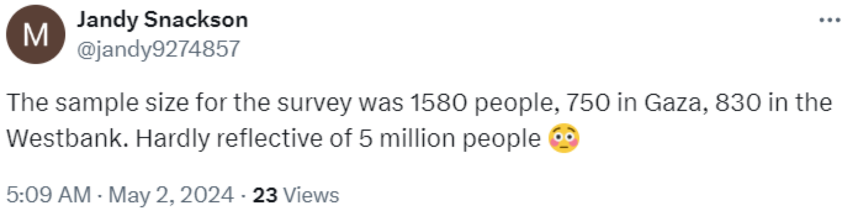
Discussion question: Below are some comments from twitter relating to the representativeness and utility of the survey. Based on what you’ve read and what you know about sampling and inference, what are your reactions to these tweets? Where do you agree or disagree?



Alexis 🙏🍷🌎
@TocquevilleJnr

I’m finding it hard to visualise the scene of when the Polling Person turns up in Gaza to conduct this professional survey
Going where no Journalist, Aid Worker or Politician is allowed to go
Visiting citizens in their tents?
How do you even begin to get a representative sample?

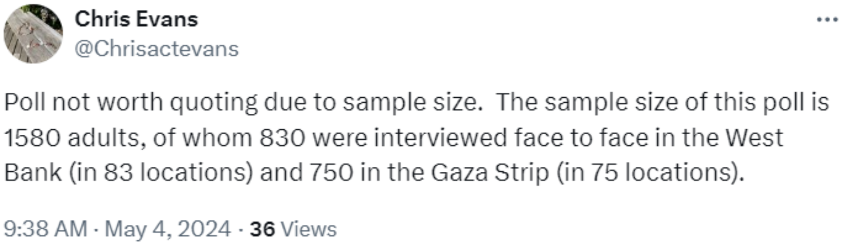
9:39 AM · Jun 17, 2024 · **24** Views



Jandy Snackson
@jandy9274857

The sample size for the survey was 1580 people, 750 in Gaza, 830 in the Westbank. Hardly reflective of 5 million people 😳

5:09 AM · May 2, 2024 · **23** Views



Chris Evans
@Chrisactevans

Poll not worth quoting due to sample size. The sample size of this poll is 1580 adults, of whom 830 were interviewed face to face in the West Bank (in 83 locations) and 750 in the Gaza Strip (in 75 locations).

9:38 AM · May 4, 2024 · **36** Views