Census and Survey Data and Introduction to Databases

Jen Helsby and Eric Potash

Computation for Public Policy

Lecture 1: January 5, 2016

computationforpolicy.github.io

Outline

- Difference between census and survey data?
- Surveys
    - Sample bias
    - Methods
- Sample Reweighting and Non-sample error
- U.S. Census Products

Census and Surveys

- What’s the difference between a census and a survey?

Census and Surveys

- What’s the difference between a census and a survey?
- A survey collects data about a sample (subset) of the population.

Census and Surveys

- What’s the difference between a census and a survey?
- A survey collects data about a sample (subset) of the population.
- A census collects (or at least attempts) to collect data about the entire population.

Random Samples

- When the individuals in a sample are (uniformly) randomly sampled from the population, the survey statistics approximate the population.
  - Law of Large Numbers, Central Limit Theorem

Random Samples

- When the individuals in a sample are (uniformly) randomly sampled from the population, the survey statistics approximate the population.
    - Law of Large Numbers, Central Limit Theorem
- Unfortunately survey samples are almost never truly uniformly random.

Survey Bias

- The difference between the true population and the sample population is called *bias.*

Survey Bias

- The difference between the true population and the sample population is called *bias.*
- In a survey where the design selects the sample there are two kinds of bias:

Survey Bias

- The difference between the true population and the sample population is called *bias.*
- In a survey where the design selects the sample there are two kinds of bias:
    - **undercoverage**: the when sample is not representative of the population of interest (e.g. individuals are not documented).
    - **non-response**, when individuals choose not to respond to the survey

Census

- "Representatives and direct Taxes shall be apportioned among the several States ... according to their respective Numbers ... . The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years."
- 2010 Census cost $13 billion ($42 / capita)

Census

- 2010 Census cost $13 billion ($42 / capita)
- Still suffers from non-response
  - Fine of up to $100 for not participating
  - But can be validated (by surveys!). Found 2010 census to be within about .01%

Non-sampling errors

“Non-sampling errors in surveys may be attributed to a variety of sources, such as how the survey was designed, how respondents interpret questions, how able and willing respondents are to provide correct answers, and how accurately the answers are coded and classified.”

Survey Methods

- Variety of survey methods
    - Door-to-door
    - Mail
    - Phone
    - Internet
- Advantages and disadvantages to each

Sample Reweighting

- Build a model of response (e.g. “Who would you vote for president?”) as a function of covariates (e.g. age, education, race, income)

Sample Reweighting

- Build a model of response (e.g. “Who would you vote for president?”) as a function of covariates (e.g. age, education, race, income)
- Weight the *sample* responses based on the probability distribution of the covariates in the *population*.

Sample Reweighting Example

- If in an online survey of presidential election preferences, males between the ages of 20 and 35 are *over*-represented, they will be *under*-weighted in the model.

Sample Reweighting Problems

- High margin of error (variance) for underrepresented groups.
- Depends on a good estimate of distribution of covariates in the population (e.g. census, but not always available).
- Sample reweighting assumes that we observe the necessary covariates to explain the question of interest.

Sample Reweighting Problems

- High margin of error (variance) for underrepresented groups.

Sample Reweighting Problems

- High margin of error (variance) for underrepresented groups.
- Depends on a good estimate of distribution of covariates in the population (e.g. census, but not always available).

Sample Reweighting Problems

- High margin of error (variance) for underrepresented groups.
- Depends on a good estimate of distribution of covariates in the population (e.g. census, but not always available).
- Sample reweighting assumes that we observe the necessary covariates to explain the question of interest.

American Community Survey

- “The American Community Survey (ACS) is an ongoing statistical survey by the U.S. Census Bureau. It regularly gathers information previously contained only in the long form of the decennial census, such as ancestry, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics.”

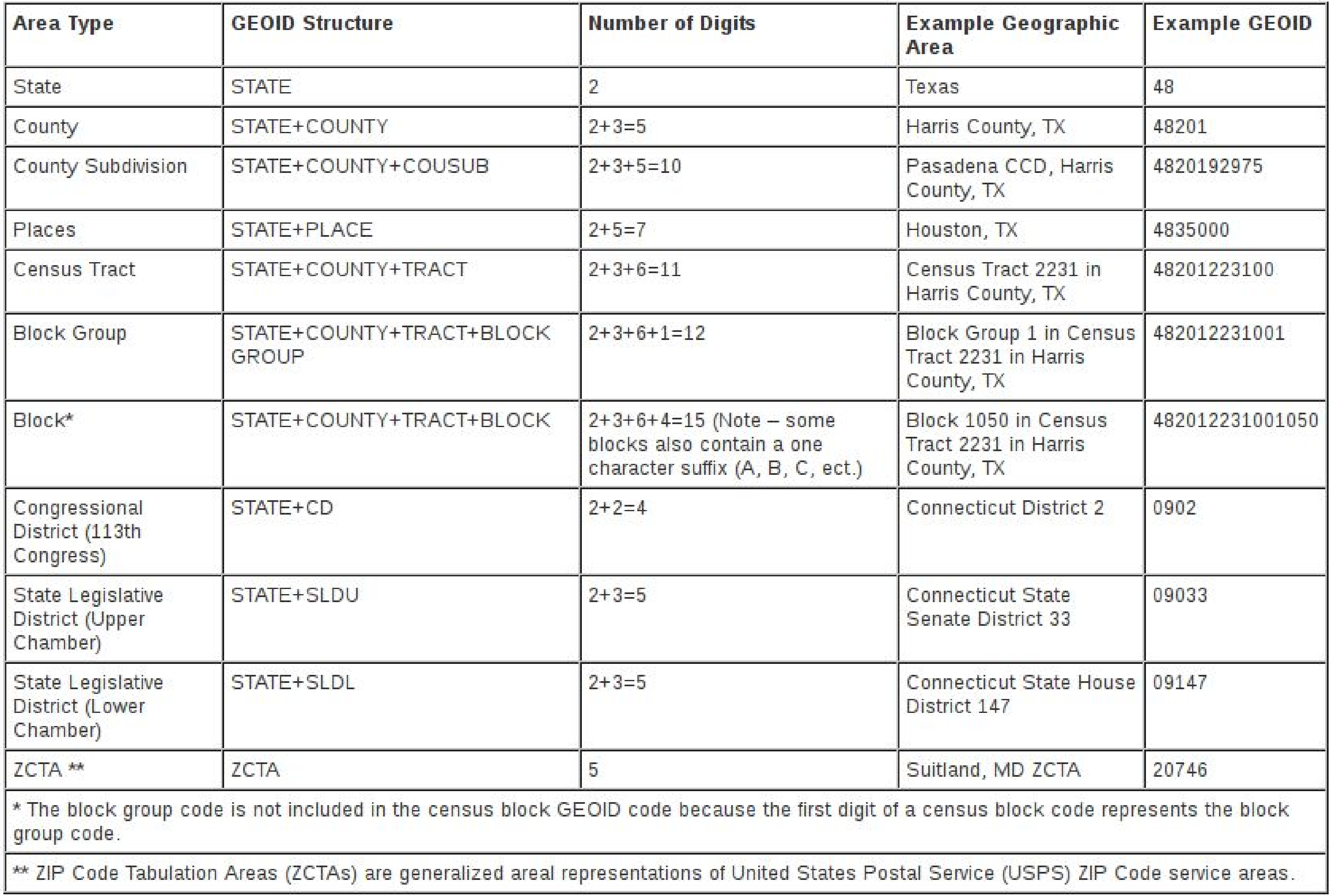
- Sent to approximately 295,000 addresses monthly

Microdata

- Individual survey responses are called *microdata*.
- For privacy reasons, the Census does not release the complete microdata.
- Instead releases:
    - Public Use Microdata (PUMS)
    - Summaries

ACS Summary Data

- Summary data is aggregated to different geographic and temporal levels:
  - 1-year estimates: > 65,000 people
    - All states, ~26% of counties
  - 3-year estimates: > 20,000 people
    - Discontinued due to budget constraints
  - 5-year estimates: census block group
    - ~600-3000 people
    - ~200k block groups in the country



Area Type	GEOID Structure	Number of Digits	Example Geographic Area	Example GEOID
State	STATE	2	Texas	48
County	STATE+COUNTY	2+3=5	Harris County, TX	48201
County Subdivision	STATE+COUNTY+COUSUB	2+3+5=10	Pasadena CCD, Harris County, TX	4820192975
Places	STATE+PLACE	2+5=7	Houston, TX	4835000
Census Tract	STATE+COUNTY+TRACT	2+3+6=11	Census Tract 2231 in Harris County, TX	48201223100
Block Group	STATE+COUNTY+TRACT+BLOCK GROUP	2+3+6+1=12	Block Group 1 in Census Tract 2231 in Harris County, TX	482012231001
Block*	STATE+COUNTY+TRACT+BLOCK	2+3+6+4=15 (Note – some blocks also contain a one character suffix (A, B, C, ect.)	Block 1050 in Census Tract 2231 in Harris County, TX	482012231001050
Congressional District (113th Congress)	STATE+CD	2+2=4	Connecticut District 2	0902
State Legislative District (Upper Chamber)	STATE+SLDU	2+3=5	Connecticut State Senate District 33	09033
State Legislative District (Lower Chamber)	STATE+SLDL	2+3=5	Connecticut State House District 147	09147
ZCTA **	ZCTA	5	Suitland, MD ZCTA	20746

* The block group code is not included in the census block GEOID code because the first digit of a census block code represents the block group code.

** ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas.

American Fact Finder

- Could download ACS Summary File for a state or the country: Summary File Lookup

- American Fact Finder is a website that provides Data Profiles and other tables containing popular and useful data.

Third Parties

- IPUMS
- Census Reporter
- Social Explorer

Databases

- A database is a computer program that stores and retrieves data.

- Common type of database is a *relational database*.

- Most common way of accessing its data is using a language called SQL (Structured Query Language).

Why Databases?

- Reliability
- Concurrency
- Scalability
- Performance
- SQL

Why Not Databases?

- Not all data is relational (there are databases for that!).

- Not all tasks are database queries (but *many* are):
    - File input/output
    - Images (charts)
    - Networking (websites)

Database Concepts

- Tables
- Variables
  - Types
- Constraints
  - Unique, Not-null
  - Relational
    - Primary Key
    - Foreign Key

Database Schema

- We refer to the collection of tables, their variables and constraints as a *schema*.

SQL

- A “natural” language for querying relational data.

```
SELECT * FROM employees employees

SELECT first_name, last_name from employees

SELECT count(*) from employees
```

Database Clients

- The database is a *server* that accepts queries from a *client* and returns results.
- There are many database servers:
    - PostgreSQL, MySQL, Oracle, etc.
- There are many PostgreSQL clients
    - psql (command line)
    - pgadmin (graphical)
    - python

GROUP BY

```
SELECT department_id, count(*)

FROM employees

GROUP BY department_id
```

Aggregate Functoins

```
SELECT department_id, count(*), max(salary), min(salary)

FROM employees

GROUP BY department_id
```

JOIN

```
SELECT employees.department_id, count(*),
    departments.department_name
FROM employees
JOIN departments
    ON employees.department_id = departments.id
GROUP BY employees.department_id
```