

Understanding, Finding, and Using Data

17.871
Spring 2015

Goals for Today

- Understanding research datasets
- Identifying potential data sources
- Hands-on exercises to download data for your upcoming problem set
- Searching for research data from ICPSR

Today's Research Topics

1. The percentage of people with driver's licenses, by age and state
2. Patterns of conflicts among countries
 - How can we find the right data? Look for:
 - Data on certain topics
 - During certain time periods
 - In particular geographic areas
 - Data can be challenging to find—start early

Social Science Data Services

- Support for finding data
- Libraries' research guide:
<http://libguides.mit.edu/ssds>
 - find data on any subject
 - find data in subject-specific resources
 - use data
- GIS Services: <http://libraries.mit.edu/gis>

Data File Characteristics

- Structure: e.g., fixed-field vs. delimited
- Unit of analysis: e.g., person, household, administrative unit, event
- Types of variables: e.g., continuous vs. discrete/categorical
- The codebook tells you:
 - study overview and how produced
 - structure of the data files
 - variable information

Searching for Data

- Who has the time, funds, and authority to collect the data?
- Find data on any by subject: top sources
- Data Centers
 - ICPSR
 - Harvard-MIT Data Center (HMDC)
- Subject-specific sources: political science
- Tip on locating data: [search the literature](#)

Statistical Abstract of the U.S.

- Key source for identifying sources of data on various topics
- Contains tables of data and footnotes that link to sources of the data and more detail
- Search or browse (table of contents or index)
- Looking for: the number of licensed drivers by age group and state
- <http://libraries.mit.edu/get/stat-abstract>
- American Fact Finder demonstration

Exercise: Part 1

- Accessing data sources noted in the Statistical Abstract
- From table 1117 navigate to its original source
- In the original source of data, find and download the 2012 data (most recent year available), broken down by state (as well as age)
- Write down for yourself the source of this data file you have just downloaded—be precise!

Exercise: Part 2

- Find a table giving total population by state, from the 2012 Population Estimates program, from American Fact Finder:
<http://factfinder2.census.gov/>
- Use the Advanced Search to select:
 - topics > people > basic count > population total
 - topics > people > age & sex > age
 - geographies > name > geographic type: state > all states
- Note the options of:
 - Population estimates program
 - estimates from the ACS (American Community Survey)
 - data from the 2010 Census (2010 Demographic Profile)
 - *Select: 2012 PEPAGESEX: Annual Estimates of the Resident Population for Selected Age Groups by Sex for the U.S. ...
- Download the table (+documentation) to your Athena locker as a .csv file (with or without annotations is fine)
- Write down the source of the data

Accessing Research Data: ICPSR

- World's largest social science data archive
- Collects and disseminates data sets
- Training in quantitative analysis
- Summer internship and research paper competition
- Note: responsible use and citing data
- MyData registration system
- Data also accessible via the Harvard Dataverse Network: <http://libraries.mit.edu/get/dvn>

ICPSR Demonstration

- <http://libraries.mit.edu/get/icpsr>
- Sample topic: patterns of conflicts among countries
- Main search
- Browsing by topic
- Bibliography of Data-Related Literature
- Reviewing a study
- Downloading data

Exercise: Part 3

- In ICPSR: <http://libraries.mit.edu/get/icpsr>
- Search for the following study: #24386, Correlates of War Project: Militarized Interstate Dispute (MID) Data, 1816-2001
- Download to your Athena directory:
 - DS2: MIDB: Participants in Each of the Disputes in MIDA (Version 3.10) (should come with all relevant documentation)

Format Data

- Import the data into your statistical software package
- Can use setup files to import the data or Stat-transfer to convert among formats
- Extract variables of interest
- Subset observations of interest
- In addition to Stata: GIS Services:
<http://libguides.mit.edu/gis>

Additional Support for Data Analysis

- Guide to Statistical Software:
<http://libguides.mit.edu/stat>
- Research Technology Consulting:
<http://projects.iq.harvard.edu/rtc>
 - Learn or troubleshoot statistical software
 - Data analysis support and programming advice
 - Statistical methodology questions
 - Research project planning and guidance
 - Use of research technology (e.g., screen scraping, social network analysis, and more)

Conclusion

- Feel free to contact me for help:
Katherine McNeill
mcneillh@mit.edu
- Ask Us! <http://libraries.mit.edu/ask>
- Thanks for coming!
- More questions?