Goals for Today

- Overview of Data Research Process
- Understanding research datasets
- Resources available to you at MIT
- Hands-on exercises
Social Science Data Services

• Support for finding and managing data
• Libraries’ subject guide: http://libraries.mit.edu/guides/subjects/data
  – Data Access
  – Training
  – Hardware & Software
• Political Science & Government Subject Guide: http://libraries.mit.edu/guides/subjects/polisci/
Data Research Process

1. Define research question
2. Define type of data needed
3. Identify potential data sources
4. Determine usefulness of data
5. Access the data
6. Format the data
7. Analyze the data
Define Type of Data Needed

• Questions to ask yourself (variables, geography, feasibility, etc.)
• Statistics or data sets
• Data sets:
  – Primary research data
  – Organized into data files
  – Can ask your own questions of the data
  – Require analysis using statistical software
Data File Structure

- Fixed-field vs. delimited
- Rectangular/LRECL vs. card image vs. hierarchical
- Unit of analysis (e.g. person, household, administrative unit, event)
## File Formats

<table>
<thead>
<tr>
<th>Type</th>
<th>Stata</th>
<th>SPSS</th>
<th>SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Application-specific data file</td>
<td>.dta</td>
<td>.sav</td>
<td>.sas7bdat</td>
</tr>
<tr>
<td></td>
<td>.por</td>
<td></td>
<td>.trans</td>
</tr>
<tr>
<td>2. Application-specific program file</td>
<td>.do</td>
<td>.sps</td>
<td>.sas</td>
</tr>
<tr>
<td></td>
<td>.dct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Delimited file</td>
<td></td>
<td>Tab-delimited (.dat, .txt)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comma-delimited (.csv)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Space-delimited (.dat, .txt)</td>
<td></td>
</tr>
<tr>
<td>4. Fixed-field file</td>
<td></td>
<td>.dat</td>
<td></td>
</tr>
<tr>
<td>5. Codebook</td>
<td></td>
<td>.pdf, .txt</td>
<td></td>
</tr>
</tbody>
</table>
Types of Variables

- Alpha/character vs. numeric
- Continuous vs. discrete/categorical
- Micro-level vs. summary-level/aggregate
- Weight variables

Tutorials on Understanding Data Files: [http://libraries.mit.edu/guides/subjects/data/training/tutorials/understanding.html](http://libraries.mit.edu/guides/subjects/data/training/tutorials/understanding.html)
Sample Codebook

TABLE OF CONTENTS

INTRODUCTORY MATERIALS

GENERAL INTRODUCTION

STUDY DESCRIPTION

STUDY DESIGN, CONTENT AND ADMINISTRATION

NATIONAL ELECTION STUDY SAMPLE DESIGN

STUDY POPULATION

DUAL FRAME SAMPLE DESIGN

PES SAMPLE DESIGN - MULTI-STAGE AREA PROBABILITY

AREA SAMPLE DESIGN ASSUMPTIONS, SPECIFICATIONS AND OUTCOMES

NES RDD (RANDOM DIGIT DIAL) SAMPLE

NES RDD SAMPLE DESIGN ASSUMPTIONS, SPECIFICATIONS AND OUTCOMES

NURSES POST-ELECTION STUDY SAMPLE DESIGN

NURSES DATA - WEIGHTED ANALYSIS

NURSES ANALYSIS WEIGHTS - CONSTRUCTION

NURSES PROCEDURES FOR SAMPLING ERROR ESTIMATION

NOTES ON CONFIDENTIAL VARIABLES

FILE STRUCTURE AND NOTE ON "DATASET NUMBER" AND "VERSION NUMBER"

CODEBOOK INFORMATION

PROCESSING INFORMATION

VARIABLE DESCRIPTION LIST

VARIABLE DOCUMENTATION

V000001 - V000003 Identification and weights
V000004 - V000125 Pre administrative, sampling, etc.
V000126 - V000213 Post administrative, candidate, etc.
V000214 - V000227 Pre: Survey variables
V000228 - V000279 Pre: Demographic variables
V000380 - V000441 Pre: Interviewer observation
V000442 - V000563 Pre: Randomization description
V000564 - V000718 Post: Survey variables
V000719 - V000751 Post: Interviewer observation
V000752 - V000810 Post randomization description

APPENDICES

NOTES ON SAMPLING VARIABLES

CENSUS DEFINITIONS

TYPE OF RACE
Identify Potential Data Sources

• Data is usually expensive and time-consuming to collect, store, and publish
• Who has the time, funds, and authority to collect the data?
• Why might someone want the data?
• Who is responsible for collecting or managing the data?
• Who might be external stakeholders?
Searching for Data

• Data Access by subject
• Multi-subject sources
  – Example: Lexis-Nexis Statistical
• Data Centers
  – Harvard-MIT Data Center (HMDC)
  – ICPSR
  – How to search a data center catalog
• Tips on locating data
  – Search the political science literature
Exercise 1

• Search for data on your topic in either:
  • Social Science Data Services subject guide
  • or
  • Lexis-Nexis Statistical
Determine Usefulness of the Data

• Will the data answer your research question?
• Consider: sample design, method of data collection, measures and units of analysis, variables, file structure.
• If you’re using a data file, consult the codebook.
Access the Data: ICPSR

- ICPSR is the world’s largest social science data archive
- Collects and disseminates data sets
- Training in quantitative analysis
- For undergraduates: summer internship and research paper competition
- Note: responsible use and citing data
- MyData registration system
ICPSR Demonstration

- [http://libraries.mit.edu/get/icpsr](http://libraries.mit.edu/get/icpsr)
- Basic or advanced search
- Bibliography of data-related literature
- Browse for data
  - Special Topic Archives
  - Series
  - Thesaurus
- Description and download
About HMDC

• Based at Harvard, MIT has a membership
• Provides a data repository and support for using and analyzing data
• Data repository contains data:
  – purchased/licensed by the MIT Libraries (including ICPSR data)
  – produced by MIT faculty members
  – from select government agencies
• [http://libraries.mit.edu/get/hmdc](http://libraries.mit.edu/get/hmdc)
Search and Download from HMDC

- Basic or Advanced search
- Variable information search
- Cataloging Information
- Documentation, Data and Analysis
- Subsetting and Analysis
Exercise 2

• Search for data sets on your topic in either:
  • ICPSR
  • or
  • Harvard-MIT Data Center
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
Support for Data Analysis

• Software information: http://libraries.mit.edu/guides/subjects/data/software

• Your department

• HMDC Statistical Consulting Service
  – help getting started with statistical software
  – advice on setting up statistical analyses
  – general statistical advice

• ICPSR Summer Program

• GIS Services
Conclusion

• Feel free to contact me for help:
  Katherine McNeill-Harman
  mcneillh@mit.edu

• http://libraries.mit.edu/ask-us/

• Please fill out the evaluation form.

• Thanks for coming!