MITLibraries

Understanding, Finding, and Using Data

17.871

Spring 2008

Goals for Today

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

MITLibraries

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)

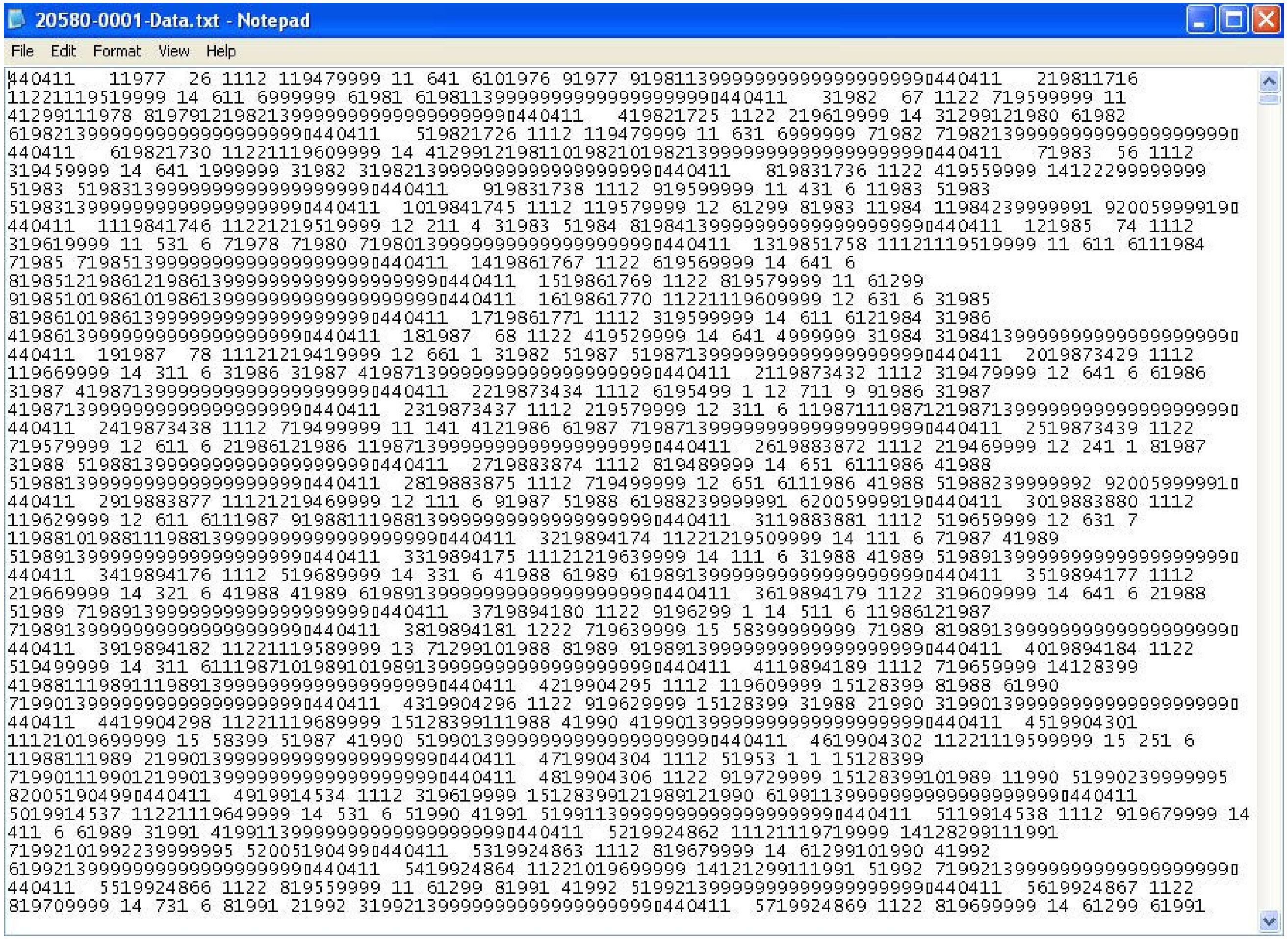
File Formats

Type	Stata	SPSS	SAS
1. Application-specific data file	.dta	.sav .por	.sas7bdat .trans
2. Application-specific program file	.do .dct	.sps	.sas
3. Delimited file	Tab-delimited (.dat, .txt) Comma-delimited (.csv) Space-delimited (.dat, .txt)		
4. Fixed-field file	.dat		
5. Codebook	.pdf, .txt		

Types of Variables

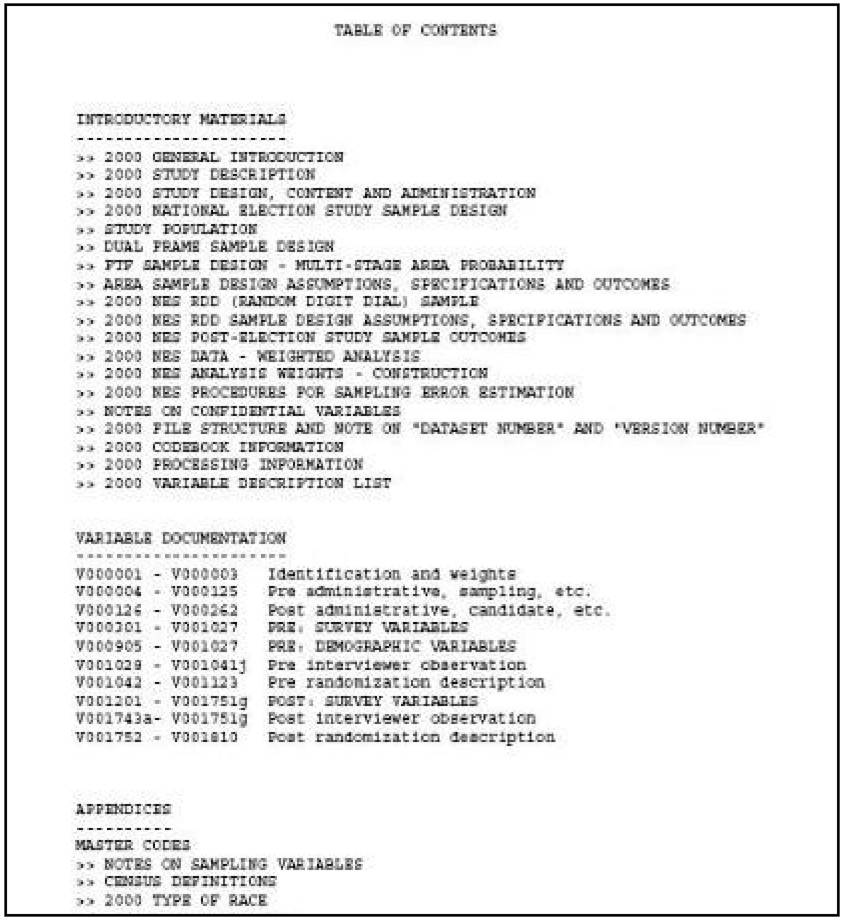
- Alpha/character vs. numeric
- Continuous vs. discrete/categorical
- Micro-level vs. summary-level/aggregate
- Weight variables
- Tutorials on Understanding Data Files: http://libraries.mit.edu/guides/subjects/data/training/tutorials/understanding.html

Sample Data File



```
440411   11977  26 1112 119479999 11 641 6101976 91977 91981139999999999999999999▯440411   219811716
11221119519999 14 611 6999999 61981 61981139999999999999999999▯440411   31982  67 1122 719599999 11
41299111978 81979121982139999999999999999999▯440411   419821725 1122 219619999 14 31299121980 61982
61982139999999999999999999▯440411   519821726 1112 119479999 11 631 6999999 71982 71982139999999999999999999▯
440411   619821730 11221119609999 14 41299121981101982101982139999999999999999999▯440411   71983  56 1112
319459999 14 641 1999999 31982 31982139999999999999999999▯440411   819831736 1122 419559999 14122299999999
51983 51983139999999999999999999▯440411   919831738 1112 919599999 11 431 6 11983 51983
51983139999999999999999999▯440411  1019841745 1112 119579999 12 61299 81983 11984 11984239999991 92005999919▯
440411  1119841746 11221219519999 12 211 4 31983 51984 81984139999999999999999999▯440411  121985  74 1112
319619999 11 531 6 71978 71980 71980139999999999999999999▯440411  1319851758 11121119519999 11 611 6111984
71985 71985139999999999999999999▯440411  1419861767 1122 619569999 14 641 6
81985121986121986139999999999999999999▯440411  1519861769 1122 819579999 11 61299
91985101986101986139999999999999999999▯440411  1619861770 11221119609999 12 631 6 31985
81986101986139999999999999999999▯440411  1719861771 1112 319599999 14 611 6121984 31986
41986139999999999999999999▯440411  181987  68 1122 419529999 14 641 4999999 31984 31984139999999999999999999▯
440411  191987  78 11121219419999 12 661 1 31982 51987 51987139999999999999999999▯440411  2019873429 1112
119669999 14 311 6 31986 31987 41987139999999999999999999▯440411  2119873432 1112 319479999 12 641 6 61986
31987 41987139999999999999999999▯440411  2219873434 1112 6195499 1 12 711 9 91986 31987
41987139999999999999999999▯440411  2319873437 1112 219579999 12 311 6 11987111987121987139999999999999999999▯
440411  2419873438 1112 719499999 11 141 4121986 61987 71987139999999999999999999▯440411  2519873439 1122
719579999 12 611 6 21986121986 11987139999999999999999999▯440411  2619883872 1112 219469999 12 241 1 81987
31988 51988139999999999999999999▯440411  2719883874 1112 819489999 14 651 6111986 41988
51988139999999999999999999▯440411  2819883875 1112 719499999 12 651 6111986 41988 51988239999992 92005999991▯
440411  2919883877 11121219469999 12 111 6 91987 51988 61988239999991 62005999919▯440411  3019883880 1112
119629999 12 611 6111987 91988111988139999999999999999999▯440411  3119883881 1112 519659999 12 631 7
11988101988111988139999999999999999999▯440411  3219894174 11221219509999 14 111 6 71987 41989
51989139999999999999999999▯440411  3319894175 11121219639999 14 111 6 31988 41989 51989139999999999999999999▯
440411  3419894176 1112 519689999 14 331 6 41988 61989 61989139999999999999999999▯440411  3519894177 1112
219669999 14 321 6 41988 41989 61989139999999999999999999▯440411  3619894179 1122 319609999 14 641 6 21988
51989 71989139999999999999999999▯440411  3719894180 1122 9196299 1 14 511 6 11986121987
71989139999999999999999999▯440411  3819894181 1222 719639999 15 58399999999 71989 81989139999999999999999999▯
440411  3919894182 11221119589999 13 71299101988 81989 91989139999999999999999999▯440411  4019894184 1122
519499999 14 311 6111987101989101989139999999999999999999▯440411  4119894189 1112 719659999 14128399
41988111989111989139999999999999999999▯440411  4219904295 1112 119609999 15128399 81988 61990
71990139999999999999999999▯440411  4319904296 1122 919629999 15128399 31988 21990 31990139999999999999999999▯
440411  4419904298 11221119689999 15128399111988 41990 41990139999999999999999999▯440411  4519904301
11121019699999 15 58399 51987 41990 51990139999999999999999999▯440411  4619904302 11221119599999 15 251 6
11988111989 21990139999999999999999999▯440411  4719904304 1112 51953 1 1 15128399
71990111990121990139999999999999999999▯440411  4819904306 1122 919729999 15128399101989 11990 51990239999995
82005190499▯440411  4919914534 1112 319619999 15128399121989121990 61991139999999999999999999▯440411
5019914537 11221119649999 14 531 6 51990 41991 51991139999999999999999999▯440411  5119914538 1112 919679999 14
411 6 61989 31991 41991139999999999999999999▯440411  5219924862 11121119719999 14128299111991
71992101992239999995 52005190499▯440411  5319924863 1112 819679999 14 61299101990 41992
61992139999999999999999999▯440411  5419924864 11221019699999 14121299111991 51992 71992139999999999999999999▯
440411  5519924866 1122 819559999 11 61299 81991 41992 51992139999999999999999999▯440411  5619924867 1122
819709999 14 731 6 81991 21992 31992139999999999999999999▯440411  5719924869 1122 819699999 14 61299 61991
```

Sample Codebook



```
                    TABLE OF CONTENTS


INTRODUCTORY MATERIALS
----------------------
>> 2000 GENERAL INTRODUCTION
>> 2000 STUDY DESCRIPTION
>> 2000 STUDY DESIGN, CONTENT AND ADMINISTRATION
>> 2000 NATIONAL ELECTION STUDY SAMPLE DESIGN
>> STUDY POPULATION
>> DUAL FRAME SAMPLE DESIGN
>> FTF SAMPLE DESIGN - MULTI-STAGE AREA PROBABILITY
>> AREA SAMPLE DESIGN ASSUMPTIONS, SPECIFICATIONS AND OUTCOMES
>> 2000 NES RDD (RANDOM DIGIT DIAL) SAMPLE
>> 2000 NES RDD SAMPLE DESIGN ASSUMPTIONS, SPECIFICATIONS AND OUTCOMES
>> 2000 NES POST-ELECTION STUDY SAMPLE OUTCOMES
>> 2000 NES DATA - WEIGHTED ANALYSIS
>> 2000 NES ANALYSIS WEIGHTS - CONSTRUCTION
>> 2000 NES PROCEDURES FOR SAMPLING ERROR ESTIMATION
>> NOTES ON CONFIDENTIAL VARIABLES
>> 2000 FILE STRUCTURE AND NOTE ON "DATASET NUMBER" AND "VERSION NUMBER"
>> 2000 CODEBOOK INFORMATION
>> 2000 PROCESSING INFORMATION
>> 2000 VARIABLE DESCRIPTION LIST


VARIABLE DOCUMENTATION
----------------------
V000001 - V000003   Identification and weights
V000004 - V000125   Pre administrative, sampling, etc.
V000126 - V000262   Post administrative, candidate, etc.
V000301 - V001027   PRE: SURVEY VARIABLES
V000905 - V001027   PRE: DEMOGRAPHIC VARIABLES
V001028 - V001041j  Pre interviewer observation
V001042 - V001123   Pre randomization description
V001201 - V001751g  POST: SURVEY VARIABLES
V001743a- V001751g  Post interviewer observation
V001752 - V001810   Post randomization description



APPENDICES
----------
MASTER CODES
>> NOTES ON SAMPLING VARIABLES
>> CENSUS DEFINITIONS
>> 2000 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
- 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

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!