Secondary Analysis and Public Use Data Sets

Paul E. Johnson[1], B. Wade Garrison[2]

[1]Center for Research Methods and Data Analysis
[2]Center for Faculty Initiatives and Engagement
University of Kansas

April 30, 2014

Using Other Peoples' Data

This happens A LOT to SOME PEOPLE in SOME FIELDS

- Advisor says "go get Roper Poll from 1996 on death penalty"
- Student goes looking, finds a "fixed field" text file with numbers like:

```
12134362614451116561 1
12145256236223523422 5
12153342577373326783 9
12167457432533211457 1
12177467321567876868 6
```

- Student needs a magical decoder ring
  - Definition of information in each column
  - Codebook that explains what the columns mean

No Commonly Accepted Format Exists

Even after "all this time"

- There is still no "mandatory" data storage format that will always be accepted
- Efforts to create generic, commonly accepted formats have been resisted by commercial software developers
- Consumers are generally "oblivious" until it is too late to escape
- Hence, anybody can use anything and expect other people to learn to use it
- Expect lots of jargon about "metadata," "compressed storage," and so forth

Some Examples We Encounter Regularly

- American National Election Study (ANES)
- General Social Survey (GSS)
- American Social Capital Community Benchmark Survey
- Survey of Income Program Participation (SIPP)
- National Longitudinal Survey

Sections are "open", "public", unrestricted
Some material closely held.

- DOE projects: NLTS (National Longitudinal Transition Study)
- DOD projects: Surveys in Afghanistan, Iraq, etc

"Secret" Confidential data that is very closely guarded

Web Access For "Taste Testing"

- http://sda.berkeley.edu "SDA: Survey Documentation and Analysis"

SDA: Survey Documentation and Analysis

Home | Archive | Community | Documentation | Projects | News

SDA is a set of programs for the documentation and Web-based analysis of survey data. SDA is developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley. CSM also develops the CASES software package.

To see how it all works, test-drive SDA at our demonstration SDA Archive. Browse the documentation for a survey and get *fast* data analysis results. The SDA Archive includes several datasets, including the **General Social Survey (GSS)** and the **American National Election Study (ANES)**. You can also look at some other archives that use SDA software.

SDA Features

Documentation:

- **Codebooks**: SDA can produce both HTML and print-format codebooks. The documentation for each study contains a full description of each variable, indexes to the variables, and links to study-level information.
- **DDI (Data Documentation Initiative) compatibility**: SDA programs can produce DDI-format metadata from SDA datasets and from other metadata formats. SDA also provides an online utility that converts DDI metadata to SDA's own metadata format (DDL).

Analysis:

- **Various analysis types are available**: frequencies and crosstabulation, comparison of means, correlation matrix, comparison of correlations, multiple regression, logit/probit regression.
- **Fast results**: SDA was designed to produce analysis results *very quickly* -- within seconds -- even for large datasets with millions of cases and thousands of variables. Although many of our users assume we are using some sort of super computer to achieve these speeds, the secret lies solely in the method of storing the data and the design of the programs. The SDA Archive on our site runs on a low-cost (Intel) Linux server -- although versions of SDA are also available for Windows and (Sparc-based) Solaris.

SDA (cont.)

- Several large general purpose datasets have been put into an SDA format for interactive exploration.
- Probably not sufficient to do a whole project, but sufficient to explore and decide if you really want to use this data.