

On-Line Analysis of Data from Complex Samples

Jon Cohen and Paul Planchon

Working with the National Center for Education Statistics, the American Institutes for Research has designed and deployed an on-line data analysis tool that:

- Enables non-technical users to access custom-run tabulations (with appropriate standard errors) over the internet;
- Provides a range of data warehousing functions that enable disparate data sets to be merged on the fly, and the appropriate variance estimation procedures selected for the resulting merged data;
- Pilots a new analytic layer of metadata that encapsulates some of the substantive knowledge and analytic skill that data analysts typically bring to their tasks, in addition to embodying knowledge about the survey data itself.
- Implements a completely modular design that will support multiple user interfaces, new statistical capabilities (e.g., graphics, or running regressions or other procedures over the web), and multiple simultaneous data formats.

The Demonstration

The demonstration will highlight the non-technical nature of the current user interface. Early trials have shown that people who are comfortable reading tables are comfortable using the system. When users encounter the system over the internet, they find a familiar table of contents organized by general topic area. They can "click" to expand branches of the table of contents until they find the particular kernel of information in which they are interested.

Upon selecting a topic, a "virtual table" appears—a representation of a table that has a title, column headings, source and population footnotes, and a single row labeled "Total." Along with the virtual table the user finds a list box describing the characteristics that might form rows of the table. Text instructs them to select as many of the characteristics as they like, which will be crossed to form the rows of the table.

In the first instance, this is it—the user presses a button and the table is generated from micro-data and delivered to their browser. However, the user is also offered the option of filtering the data to represent specific subpopulations, and to change how variables are categorized and labeled.

Resulting tables include margins of error calculated in a manner appropriate to the table requested (which may have required the merging of multiple data sets) footnotes describing the population to which the table is applicable, and descriptions of the sources of the data.

Where supported by the data, the system enables users to merge data from multiple data bases to form tables. These operations are handled entirely by the system, so that boundaries between data sets are transparent to the user.

Presenter Name: Dr. Jon Cohen

Affiliation: American Institutes for Research

Pelavin Research Center

1000 Thomas Jefferson Street, NW

Suite 400

Washington, DC 20007

(202) 944-5300 Ext. 5420

(202) 944-5454 (FAX)

jcohen@air-dc.org

Paul Planchon

National Center for Education Statistics

U.S. Department of Education

555 New Jersey Avenue, NW

Washington, DC 20208

(202) 219-1616

Paul_Planchon@ed.gov