Connect Fall 1996 STATISTICS AND THE SOCIAL SCIENCES

Statisticians' Lib: Using Scanners and OMR Software for Affordable Data Input

Frank LoPresti with Zvia Segal Naphtali

Falling prices for reliable scanners, along with improved software, should make life simpler for researchers in the social sciences. While programs like like SPSS and SAS satisfy most needs for statistical analysis, they depend on the existence of good, clean datasets. You can now create such datasets using Remark Office OMR (optical mark recognition) software and a low-priced scanner. I have been using Remark with a two-year-old Pentium-based computer and an HP 4P scanner for affordable, robust, versatile input of questionnaire data to file.

Before I describe how it's done, let me give a short history of data input. That will give you an idea of just how big an improvement this style of input is.

The Data Bottleneck

For years, people who do statistical analysis have been designing questionnaires, getting them filled out by respondents or interviewers, and then somehow wrestling the data into a computer. The questionnaires might be a department's student evaluations, a perfume company's evaluaton of its packaging and ad campaign, or a PhD candidate's thesis research. But always, the most problematic aspect of data collection has been getting the data from the questionnaire into the computer.

With the first computers, much of the data was input by creating decks of punched cards. While this process allowed one to create the necessary computer files, it was subject to input error and thus had to be verified. To verify the data, one created two hypothetically identical decks and compared them to each other; any discrepancy indicated an error. One then went back to the questionnaire to find the right value -- and then had to punch a new card to correct the mistake. Although some keypunch machines allowed you to copy 79 of the 80 fields and edit the one incorrect punch, the process was still time-consuming and therefore expensive. Only well-funded researchers could afford to verify; others merely "checked for errors with a friend."

Over the years, we have seen developments to aid in the task. Keypunch sessions were replaced by input sessions at mainframe terminals. Using primitive text editors on mainframe computers, researchers typed line after line of questionnaire responses to create a file. Correcting an error now meant revising a single line, not an entire card. A great improvement came with personal computers. With spreadsheets and wordprocessors, data could be entered with relative ease, with or without verification. I was excited when laptop PCs came out and packages like DE/SPSS were developed. DE (Data Entry) allowed one to create screens with active fields for data to be typed in by the user. The data ended up in SPSS (Statistical Products and Sevice Solutions) files.

Unfortunately, this method of data collection has a fairly high overhead -- it requires a PC for each collector of data. It is useful in many controlled situations in a social-science lab, where one machine can serve for many experimental subjects, one after another. My mentor taught me that the task of creating a clean data set from questionnaires should be budgeted the lion's share of a project's resources. In the early days of computing, data entry was a substantial and ugly task.

Automated Data Entry

At about the same time that punch-cards were developed, another technology came into use that allows multiple-choice forms to be read. These forms are probably familiar to most people from tests like the SAT. I took tests on these forms back in 1955.

Originally, the answers could be "read" by the computer because the soft pencil lead that filled in the bubbles or squares conducted electricity between matched pairs of minute electrodes. More recently, the marked forms have been scanned optically; the technology is still expensive and difficult to use. Hardware dedicated to this limited task costs several thousands of dollars. It is used along with software that allows us to design our own forms -- a difficult skill to learn. Otherwise we pay specialists to design the forms or we buy generic answer sheets that can be used along with a separate questionnaire. Such a solution might be satisfactory for a large academic or administrative department, but it is of limited flexibility and is beyond the means of most individual researchers and small groups.

Since the data go from the form directly to the computer, traditional data verification isn't needed. And in testing situations such as the SAT, the subject can usually be held responsible for poorly filled out forms. But in other situations, ambiguities resulting from partially filled-in bubbles, partial erasures, and crossed-out answers have to be resolved by an operator, who cleans up the form with eraser and pencil -- an often costly intervention. Over the years, this technology has made few inroads at NYU as a data-input method.

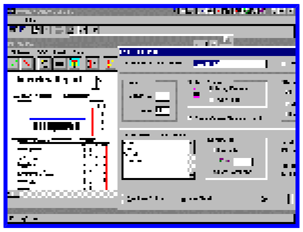
Will OCR Help? Or Maybe OMR?

Optical character recognition (OCR) is a technology that works quite well on uniformly typed or typeset material, though the user must still be prepared to correct several errors per page. But on handwritten material, it is nearly useless; though it can be "trained" to recognize a given person's handwriting, that trainability would be no help in deciphering a series of short questionnaires filled out in different hands. Nor is it intended for reading forms and questionnaires. Thus, OCR wouldn't be useful to survey researchers.

Traditional OMR (optical mark reading) software is too expensive, and current OCR isn't much help. The Remark Office OMR software for Windows 3.1 fills the gap by recognizing filled-in bubbles (and even barcodes) on user-defined forms, and by allowing an operator to complete the data files by typing in the data from handwritten responses.

How It's Done

Once the questions have been developed -- a major undertaking not discusssed here -- the reseacher can use any wordprocessor to design the questionnaire itself. Remark allows several types of answer fields to be used in the design. Many questions can be fully automated using traditional bubble OMR fields -- the mainstay of automated forms, allowing easy marking for true-or-false responses, multiple-choice questions, or a rating of ranges, as from best to worst. (The "bubbles" can be virtually any shape or size.) Barcodes can also be used -- most likely for ID numbers to be pasted in, but possibly attached to images that could be inserted. Additionally, open fields -- blanks -- may be used. These would allow a respondent to write in a name or country of birth, say, rather than trying to encode a 10- or 15-letter name in a field of bubbles 26 high.



There's considerable freedom in the design of the questionnaire for Remark, since the researcher can essentially tell the software where on the form to look for answers, and what to look for. The "bubbles" to be filled in can be of any shape desired and in any array; the blanks can be of any length. You can also place tick marks at the corners of each page to help orient the scanner.

The next step is to create the template -- to tell the software what to look for. First, you scan in a clean, blank copy of the entire questionnaire. Then, you analyze each page: using the mouse, you draw a rectangle around each field to be read, and indicate whether the answer is recorded in OMR bubbles, written on a blank line, or pasted in as a barcode. At this point, you assign the meaning for each array of bubbles -- numeric range (Likert scale) or categoric (true-false or multiple-choice). You also point out the tick marks if you've used them, which helps the software to align the scanned form properly.

After the forms have been filled out, we are ready to get the data into the computer. Basically what happens now is that the forms are scanned and the software extracts the data. The scanner (preferably with an automatic page feed) scans in around four pages per minute. After the physical forms are scanned in, the operator normally doesn't handle them again; they filed and, emergencies aside, forgotten.

On the computer monitor, the operator sees a window like a spreadheet, with each row representing the answers from one questionnaire, and each column representing a variable. (Each question may produce one or more variables, depending on the design.) If a cell is highlighted and has no value, the operator needs to intervene -- either because there was a problem reading a bubble field (too many bubbles marked, or a messy scan) or because a handwritten answer needs to be typed in. Using the mouse, the operator double-clicks on the highlighted cell, an action window opens, showing an image of the unresolved answer. In the case of an image field, the scanned handwritten response is displayed and the operator types it in; if a bubble field is at issue, the operator is shown a scanned image of the problem answer and can then resolve the problem.

The Bottom Line

The cost for my Hewlett-Packard 4P scanner with a fifty-sheet automatic feed was about $1000. (Although I originally purchased it for automating data input, I now use for faxing and image-scanning as well.) Since it can cost a dollar a page to have data entered and verified by hand, even a single research project can pay for the scanner, plus the roughly $400 for the software.

Remark Office OMR 3.0 is made by Principia Products, Inc., of West Chester, Pa.; for more information, call them at 610/429-1359, or visit their Web site at http://www.principiaproducts.com/ .

Frank LoPresti will be giving a demonstration on using this software on Friday, November 15, from 2 to 4 in room 313, Warren Weaver Hall. Zvia Naphtali will be including an introduction to this software in a course this fall.

Professor Naphtali teaches statistics, survey research, and GIS at NYU's Wagner Graduate School of Public Service. Frank Lopresti heads the ACF Statistics and Social Science Group.
frank.lopresti@nyu.edu
naphtali@is2.nyu.edu

Posted 24 September 1996

Search Archives
Connect Home