A Glimpse at the Future of Social Science Statistical Data: new forms of data analysis, new types of access, and new issues for data providers

by ***Stephen E. Fienberg****

Abstract

The world of computing and communications is in enormous ferment and statisticians and statistical data users need to pay attention. The Internet has changed forever how we think about the issue of access. At the same time, new forms of data are emerging as alternatives to the traditional numerical responses that survey methodologists have dutifully encoded for use in statistical analyses. Survey data sets of the future may well consist, either through direct collection or forms of record linkage, of combinations of traditional numbers, text, images, sound, and even symbolic summaries. New statistical methods will need to deal with such mixed media, and the new data and methods will raise new issues with regard to such topics as confidentiality. As a consequence, the ways in which organizations provide and individuals access statistical data bases are surely going to change in a radical fashion. This presentation offers some thoughts and speculations on these topics.

Introduction

I am pleased to join you here today. Those involved with social science data archives around the world are providing an extremely important service to highly varied professional communities in support of what I take to be a fundamental principle: the access to research data. There are those who still believe that we should restrict access to data, both government data and those collected by researchers in universities, for confidentiality and other reasons. I believe such thinking is fundamentally flawed, and that the new world of communication and computing will ultimately lead us in the direction of unrestricted access. In fact this is one of the themes of my talk.

The past fifteen years has produced a remarkable revolution in computing and communications, and we are regularly told about the radical changes we can expect to see in the near future. It is that future I plan to comment on this morning, and the likely impact on the way all of us do our work. My message is that "the future is now" and, if we work to take advantage of what it offers, we will change what we do in a radical fashion. My talk consists of brief glimpses at the past and present, and a longer look at the near future. I'll end with some thoughts and speculation on how social science data archives and government statistical agencies can improve the accessibility of the data they collect and produce, and how statistical users should be thinking about the new environments for databases.

The Past

I stand before you as someone whose professional career began in a different era, with the "so-called" mainframe computers of the 1960s, and I remember quite vividly my first exposure to government statistical data bases at the University of Minnesota in the early 1970s. In the building next to us, colleagues had established one of a handful of computer centers nationwide, about seven I believe, devoted solely to the analysis of U.S. census public use tapes. Those of us outside the center would not have dreamed of having direct access to any of these data, except in the summary form released in print by the Bureau. Instead, a full-time center staff responded, after a day or so delay, to requests for summaries or cross tabulations, provided that the data released did not violate what were then the U.S. Census Bureau's confidentiality guidelines.

Statistical agencies have often been in the vanguard of technological change and they have used innovations in computing to change the tasks of their employees and the accessibility of their data. For example, the U.S. Census Bureau employed, some 50 years ago, the first large computer outside of the military. Today, however, changes in computing and communications are occurring at a dizzying pace, and few government agencies can afford to lead let alone follow.

The Present

The world of computing and statistical databases has changed quite dramatically since the early years of computing to which I have been referring, most notably with the rise of the personal computer and the development of distributed computing environments linked through networks that make resources that are thousands of miles away seem as if they are across the campus or down the hall. For example, the results of the 1990 U.S. decennial census are available on compact disc, and anyone with a PC and CD-ROM reader has the physical capacity to carry out innovative analyses of official statistical data, and large chunks of the data are also accessible from the Internet in various forms. And we are slowly catching up with the

missed opportunities of the past. For example, at the University of Minnesota, historians are assembling a WWW site with PUMS (public use microdata files) for all censuses extending back almost to the US Civil War.

Today, the U.S. Census Bureau is, at long last, beginning to think about the possibility of direct World Wide Web (WWW) access beginning with files from the 2000 census of population and housing. But what such unrestricted and unfettered access means for confidentiality may require new considerations. How can data archives and government statistical agencies plan for the uncertain future ahead? At a minimum, they must try to anticipate the types of demands that statistical users will make, and think in terms of leading once again.

The Future of Computer Networking and Collaborative Statistical Research

I now turn to the future, cast largely in terms of activities that I engage in as a statistical researcher. You will need to translate components into a framework more directly relevant to your personal interests. But do not think that what I will describe is a fantasy; each piece of it exists today in a usable or at least a semi-usable form. It is simply that most of us have not yet had occasion to assemble all of the pieces together in a single place to exploit for our own work.

To begin, I describe my environment at Carnegie Mellon University. On my office desk sits a computer workstation with multiple processors; it has far greater capacity than the largest computer available only a decade ago. This workstation is part of a department local area network consisting of about 50 similar workstations and a variety of servers, as well as a graphics lab with several more powerful graphics devices and a terabyte of rapidly accessible memory, all of which are capable of being linked for complex computing and simulation tasks. Our LAN in turn is part of a campus-wide network which has an access point in every classroom and every faculty and staff office on campus as well as in the dormitories and a number of special student computer laboratories. This allows me access to specially programmed and configured machines for language translation and text processing. We link directly to the joint CMU/University of Pittsburgh Supercomputer Center, which is part of the high speed backbone for the U.S. Internet infrastructure. This allows me direct links to resources from businesses, government statistical offices, and universities all over the world. Attached to my computer is a laser printer, a scanner, a compact disk player, stereo speakers and a number of other electronic devices, not all of which are depicted here. My vision is about how statisticians can expect to use this equipment and related technology to facilitate their day-to-day work and collaborations with colleagues around the globe.

It is 9 a.m. on a bright May day, in 1998, and I have just entered my office at Carnegie Mellon University. My workstation is already on and I sign in with my name and password. I open several windows on the screen including my WWW browser, with an applications menu that includes the department's electronic mail system. As I open my electronic mailbox, a message arrives from a colleague, Guido, at the Catholic University of Chile in Santiago. The message includes a draft section for a paper we are working on with a colleague at Statistics Netherlands, Leon and another document whose format I do not recognize. A covering note from Guido describes the additional file, which includes a set of 50 variables which he would like to merge into our data base for estimating the size of the population of several countries using multiple sources. He explains that he acquired the data only yesterday for the U.S., Chile, and five of the EC countries from the web homepages of their statistics agencies, and the datafile consisting of several gigabytes of data was included in the unidentified compressed file I found a few moments ago.

Our paper deals with population size estimation using new statistical techniques for multiple-recapture analysis, and it is based in part on a new probabilistic matching approach which extends the widely-used Fellegi-Sunter method of record linkage. We now have four sources of data for samples from each of the several countries. In his message, Guido explains how he has developed a new way to produce the posterior probabilities of matches for our analysis and that he has also prepared a new program to display the results with dynamic graphics. He suggests that we arrange an interactive video session with Leon in The Netherlands, during which we can experiment with this new program and fine-tune our methods. I send an e-mail message to Leon, who is working late at his office preparing his section of the paper which utilizes text information recorded in survey interviews to supplement responses to a series of questions on race and ethnicity. He responds in minutes suggesting that we begin immediately.

Through customized menus on my WWW browser, I now activate the our current draft manuscript, the results of the data analyses, and the video teleconferencing system. Displayed on each of our screens are live video images of the three participants, in this case Guido, Leon, and myself, and a joint interactive "whiteboard" workspace that we can each manipulate. Guido begins to demonstrate the matching algorithms and his newly developed dynamic graphics tools, but Leon and I occasionally intervene and adjust the procedures and the settings. As we watch the revised program execute we discuss how to alter the text of the paper describing the graphical tools and the data summaries. Leon explains how he proposes to use the text information on race and ethnicity to reclassify respondents by social group, and to correct the probabilistic matching algorithm. We then create a video of the dynamic

graphical display of our probabilistic matching results for The Netherlands. For our method of population estimation, we include all of the extra variables in the database that are not part of the matching algorithm as part of a covariance adjustment for heterogeneity.

The paper we are working on is being submitted to a new electronic statistical journal begun by statisticians at the *University of Stockholm* in collaboration with *Statistics Sweden.* The data base and video have actually become part of the paper we hope will be "published" in the journal.

I then do an electronic library bibliographic search using a local file of statistical titles and key words to locate additional references for the paper with Guido and Leon, and I e-mail them copies of what I find. Carnegie Mellon's campus library is part of an electronic service providing computer access to over 1200 journals, and I can receive faxes of requested articles within 24 hours for those journals that are not electronically archived. I can copy them into my files and share them with my colleagues who do not yet have direct access to this service. Leon in particular is hampered by access, because of the limited technical library at Statistics Netherlands, and the unwise agency policy on restricted access to the WWW.

Before leaving the office I open another web page and locate a weather summary for Odense where I have a meeting the next day. And I move the files from my workstation onto my portable computer to use for my live presentation, as I have done today.

The Future is Today

Who can predict the future with any certainty? As a statistician, I know that such forecasts are fraught with error. In about 1940, the British mathematician G.H. Hardy was asked to comment on the usefulness of the fruits of his research on number theory. He replied that the work had little likelihood of any real world application. This turned out to be incorrect and the past decade has seen a series of major new applications of Hardy's work in number theory. Similarly, there is an apocryphal story about the applied mathematician John von Neumann, who helped to develop the modern computer. At an early stage in this work he suggested that one very large computer is all that would be needed to solve the entire world's computational problems. He would be truly astonished by the new world of computing and communication that we enjoy and take for granted today.

Were I wrong in describing the future of computing and telecommunications to you, I would have placed myself in superlative company by invoking the names of Hardy and von Neumann. But the statistical tools of prediction are not what I have used in preparation for this talk. Rather I have simply told you about things that I have already done myself, some that I have actually seen but not used, and others that I have at least read about. My seemingly futuristic description includes several activities that I actually engaged in over the past month or so, not just contemplated for use on that day in May 1998, when I hope to return to Odense for another meeting. For example, a computer-based video teleconferencing system similar to the one I described exists, but I cannot afford to use it in my collaborations with Guido and Leon. Both of them exist too, by the way, and we collaborate and interact electronically, although not as a trio. Moreover, WWW access to large scale government statistical databases exists but it is limited at best, as is the harmonization of data that would allow comparable analyses across all EC countries. Here are some further caveats to the story I have told:

• The machine translation programs and text processing machines I described in passing was borrowed from my colleague and Dean of Computer Science at Carnegie Mellon University, Raj Reddy, who for several years has been working with others on just such a translation program project.

• The shortcomings of commercial telecommunications systems and the limited channel capacity of several links on the Internet make the kind of real-time quality interactive video/computer teleconferencing I described prohibitively expensive. The transmission of digitized video and audio requires advances in compression technology and alternate modes of transmission. But cable television systems in the U.S. and here in Europe are now providing Internet access to selected areas on a trial basis and in essence the capacity to do the very things I've speculated about.

• Statistics Sweden does not yet sponsor an on-line electronic statistics journal, although it does publish a traditionally-printed journal of very high quality, *The Journal of Official Statistics*. But electronic journals do exist and they are the focal point of the publication strategies for many professional societies, and some have just begun to include videos and interactive programs. The type of on-line access to electronic journals that I described is the goal of many of these groups, and it complements the electronic access to data.

• Most statistical methods still deal solely with numerical data , usually in the form of an $n \times p$ array, or perhaps hierarchically structured. This remains the focus of most data archives as well. But a number of researchers now work with the analysis of images, and others are interested in text data. In fact, one of the goals of a new interdisciplinary Center for Automated Learning and Discovery at Carnegie Mellon university is the development of tools for the analysis of mixed media data, including numerical data, images, text, sound, and symbols, etc. Learning to work with mixed media

presents new challenges to data archives.

• Then there is my own research on multiple-recapture methods (e.g., see Darroch et al., 1993), especially with probabilistic matching (Ding and Fienberg, 1996), but which has yet to progress to the kind of implementation involving text data that I represented. Nonetheless, I predict with some confidence that someone present here today will be working with such statistical tools before the end of the decade, in a fashion not unlike that which I have described.

Making Data Maximally Accessible to Meet Diverse Statistical Needs

As I have tried to suggest, the ways in which we provide and access databases are surely going to change in a radical fashion in the next couple of years. How should the providers and the users be thinking about such changes? My answer is via a new kind of accessibility. For me, accessibility involves a number of different dimensions, and I'll end by addressing three of these briefly: physical access; software reformatting data to meet user specifications; preserving confidentiality of individual responses.

Virtually every statistician and social scientist's desktop has on it a computer with substantial capacity to analyze data using one or more statistical packages that allow for exploratory data analysis as well as more formal statistical methods and graphical diagnostics. This capability is available today to university students and policy analysts in business and in local and state government, not simply specialists in statistics. What is not present for all, but what will be very soon, is the communications and networking capacity I described this morning. Thus I argue that government statisticians now must face the reality of what the users of their data would actually like to have in the way of data access and do in the way of statistical analyses. Statistical agencies can no longer simply release selected cross-classifications or complex files that require the agencies' own computer programs for analysis; they need to facilitate data linkage across data-bases, and provide data in a form suitable for analysis using causal models and prediction equations. And users will demand careful on-line documentation as well as access to virtually complete databases, formatted in ways to facilitate their analysis.

Because of changes in data access, agencies and data archives also need to develop software to function at the interface between on-line files and the kinds of analysis files users require. On the WWW, such interfaces need a transparency not present in current approaches and they will involve programs for formatting, extraction, and even for statistical analyses, especially if specialized analyses are required for proper use of survey data.

Perhaps the most challenging task facing statistical agencies in this new environment is the preservation of confidentiality of respondents. Never before will so many have had access to so much data. In such circumstances, the opportunities for malfeasance will inevitably grow. The simple solutions for data disclosure limitation may no longer be as effective, given an intruder with unlimited access to other databases, on the same or similar respondents and extensive processing capacity for record linkage and matching. The challenge for data archives and government statistical agencies will be to provide "complete" data on samples of respondents, but in a form that makes difficult (although not impossible) the task of an intruder attempting to gain information of specific individuals or enterprises. This requires new ways of thinking about public-use statistical microdata files that are consonant with modern statistical theory - - the focus of much of my current research (e.g., see Fienberg, Steele, and Makov, 1996; Fienberg, 1997). But this is a story for another day.

References

Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., and Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.

Ding, Y. and Fienberg, S.E. (1996). Multiple sample estimation of population and census undercount in the presence of matching errors. *Survey Methodology*, 22, 55-64.

Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and loglinear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference.* U.S. Bureau of the Census, Washington , DC, 87-105,

Fienberg, S.E. (1997). Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research. Background Paper prepared for the Committee on National Statistics, Washington DC.

* Prepared for presentation at the plenary session on "Statistical data Producers," IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science at Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A. He is currently a visiting researcher at Statistics Netherlands.
This work was supported in part through a contract from the U.S. Bureau of the Census with Westat, Inc.