

DESIGN AND ANALYSIS TECHNIQUES FOR LARGE DATA FILES: THE CODAP SYSTEM

Eduardo N. Siguel, Ph.D. and Sidford F. Sand National Institute on Drug Abuse, Rockville, MD 20852

ABSTRACT

This paper describes issues related to the design and analysis of large data files, and indicates how one set of large data files, the Client Oriented Data Acquisition Process (CODAP), is currently maintained and analyzed.

Key words: Client Oriented Data Acquisition Process; CODAP; data collection; large data files; statistical analysis; statistical issues; systems design.

1. SYSTEM DESIGN CONSIDERATIONS

The major steps in the design of a data system are: (1) determine objectives, (2) decide what kinds of issues one wishes to deal with, (3) describe the questions one wants to answer, and (4) design a data system so that it will provide the answers (research or system design). Ideally, data systems should be designed with specific objectives, and the data to be collected should be able to meet those objectives. Unfortunately, these objectives are rarely met when large data systems are designed. In practice, one may find that the design of a large data system is characterized by: (1) general, non-specific objectives such as "support of management decisions", (2) general issues, such as "We want to improve planning, management, evaluation, etc...", (3) failure to define in

advance of the system design effort the questions which are to be asked, and (4) system design being executed on the basis of what seem to be "interesting" questions, subject to constraints imposed by money, time, administrative "clearance" requirements, and the willingness of respondents to provide the information.

If one may assume that objectives were clearly stated, that issues and questions were defined in operational terms, and that the data elements to be collected are necessary and sufficient to answer the questions posed, then it is useful to consider the area of system design which directly effects the analyst's ultimate products: data collection. (For purposes of this discussion, availability of internal data control and processing resources which are adequate to handle collected raw data is also assumed.) There are two major aspects to consider when designing the data collection instruments and processes:

A. Substantive Attributes: (1) The complexity of the questions asked and the ease of formulation and expression of the answers. (2) The likely availability of respondents' knowledge and informational materials (records, logs, interviewees, etc.) which facilitate determination of correct answers. (3) The degree of interrelatedness of questions and answers, and the "intensity" of the requirement that answers be internally consistent. (4) A host of environmental and attitudinal aspects which inevitably influence all of the above. The amount of self-discipline which the data acquisition process imposes may be realistic or absurd depending on attitudinal and role factors. The most important determinant of the

success of the respondents' activities is usually the answer to his/her question: "What's in it for me?"

B. Attributes of Form Design and Instructions: A wide variety of "structural" techniques for increasing the viability of a form are available. Usually attributes such as arrangement of items on the page and the coding structures employed are belabored at length. Then a professional forms preparer adds a few additional niceties such as compatibility with typewriter spacing, different printing fonts, and color or shading for emphasis.

Most frequently overlooked or badly rendered are aspects having to do with "data control", such as use of carbon copies, preprinted serial numbers, aids to batching, logging, transmitting and filing of forms, turnaround and feedback documents or printouts, and machine-sensible forms. Even if the data to be acquired is easily encoded and training of respondents is sound, data control is crucial. In a large system, one frequently deals with a geographically distributed population of respondents who vary greatly in education and motivation and whose internal record-keeping arrangements vary from immaculate to nonexistent. More difficult are problems of "followup" in systems which "track" an activity of some kind in which a second, third, or nth transmission of data is related to previous data transmissions and provides additional information or corrects or updates previous data. Here problems of missing or duplicate items in a series of transmissions, failures to properly associate a transmission with its related predecessors, incorrect "transaction types" and resulting imbalances between types of transactions

can result in buildups of records which cannot be disposed of properly within the rules that govern the system.

For every such problem there are potential solutions. These may include manual and computerized logging, validity and consistency checks and a variety of feedback mechanisms, "turnaround" documents and a host of other techniques. The problems which defy solution usually stem from human factors of motivation, staff turnover and conflicting priorities or are problems whose genesis is a flawed, unreasonable or obsolete aspect of the system design itself. In the former situation the respondent and his motives, methods, and priorities are at least partially beyond the reach of the system maintainers, and even where the respondent's errors, inconsistencies, and omissions can be identified, usually only the respondent himself can provide the correct answers. Since the respondent's performance fell short the first time, the chances that he will ignore or compound the errors are quite high. There is thus a considerable difference between being able to detect errors and being able to get them corrected. The latter situation frequently stems from the indiscipline, alluded to above, of the systems designers themselves. Such problems may ultimately destroy the system itself by the simple process of yielding a data base of questionable usefulness. The cost in human terms of a system based on flawed concepts is immeasurable, and serves to reemphasize the importance of formulation of the system's basic concepts and objectives.

C. Compromises Between Substantive and Technical Issues: In the final analysis, for each system a balance is struck between substantive and technical issues. Each has a limiting effect on the other. The most

perfect, elegant expression of the designer's data "needs" will probably require a respondent population of psychic Ph.D.'s and a 20-page input form, while the data processing technician can easily design an almost infallible form and instructions, but one whose infantile oversimplifications and omissions will yield data which is clean, complete, and of almost no use to a statistician or program manager.

During the system design and testing process a large number of compromises are reached to ensure that, firstly, the data gathered will actually answer most of the important questions it is designed to answer in a meaningful, relatively undistorted manner. Secondly, the information must be obtainable and expressible for the respondent, and the form to be completed must make rendering of such answers as easy as possible.

When the mechanics of the information-gathering process are finally defined, a variety of training requirements and strategies will have been identified and instructional materials prepared, usually including manuals which tell a respondent how to fill out the forms involved. The strategies will reflect the designers' emphasis of various factors: minimization of errors in specific items, minimizing the time required to fill out the form, restrictions on coding space, simplification of questions and instructions, increased probability of legibility or successful transmission of completed forms, etc.

2. STATISTICAL ISSUES IN THE ANALYSIS OF LARGE DATA FILES

There are a wide variety of issues to be considered when one attempts to analyze the data in a given file. We will name a few of the most important ones that have particular impact upon large data files.

The first step in data analysis is to define the problem and the model or framework used to consider it. The objectives, issues, problem or question under consideration must be stated in operational terms, and phrased in the form of questions or hypotheses to be tested. In addition, there must be a model which serves as a framework within which to answer questions and a context within which to test hypotheses. Data, by itself, has no meaning, and must be interpreted within the context of a model. Therefore, design, issues and questions make sense only within the framework of a model of the situation under consideration. The statistician's role is to define the model which best describes the issues. Within the model, the statistician must phrase the questions in such a manner that a researchable, objective answer is possible.

Once the problem and the model are operationally defined, a methodology is developed which takes into account the nature of the data. Factors which the statistician may consider include: (1) How the data were collected. (2) The nature of errors. Usually emphasis is placed on sampling errors, but non-sampling errors may actually be much larger than sampling errors. Non-sampling errors include such errors as respondent errors, poor instrument reliability, measurement errors of other kinds, transmission errors, data processing errors, etc. (3) Methods useful in the analysis of the data. There are a variety of multivariate methods available. When large amounts of data are involved, efficient use of

computer time becomes a necessity. Computer efficiency begins with the use of efficient software and proper file design. Unnecessarily large record sizes or inadequately grouped records may greatly increase computer processing time. When one uses standard software packages, such as SPSS or BMDP, and not all cases are to be considered (for example, when one instructs the program to consider only females 18-20 years old), it is important to phrase a complex sequence of conditional statements in such a manner that conditions are tested according to the likelihood that they will fail, conditions with a higher probability of failing being tested first. This procedure reduces processing time because fewer records need to be processed completely. (4) Interpretation of the results. It is important to distinguish between statistically significant differences and differences that are not large enough to be meaningful in terms of policy and program decisions, management issues, etc. One often finds that relationships between two variables, X and Y (or the difference between X and Y) are analyzed testing for no relationship (or no difference between two distributions) using a chi square statistic (or similar statistic). With a large data file, a cross tabulation of almost any two variables is likely to have a very high chi square value. Two empirical distributions are likely to be found different even though the differences between them may be very small. Two alternative approaches can be used: (a) report the data with an appropriate confidence interval, or (b) determine, "a priori", a particular relationship that is meaningful (or a particular difference that is meaningful) and then test the hypothesis that the difference is greater than the pre-established value (rather than the null

hypothesis), or that the relationship is stronger than the preestablished value (using non-central chi square).

Another aspect requiring consideration involves the complications arising from the use of many variables. A relationship between two variables may change direction when a third variable is used as a control variable. When the data file consists of many observations (cases) and many variables, it is possible to obtain apparently contradictory findings according to which variables are included in the analysis. Inclusion or exclusion of subpopulations may change relationships. The availability of many cases and many variables encourages alternative approaches to data analysis and potential apparent inconsistencies in the interpretation of findings.

3. CODAP -- AN EXAMPLE OF A LARGE DATA FILE

A. Description of CODAP: The Client Oriented Data Acquisition Process (CODAP) is a data collection system developed and operated by the National Institute on Drug Abuse (NIDA) in treatment facilities (clinics) that receive federal funds. Its purpose is to provide current information which describes clients and the treatment provided to them in order to aid in planning, management and evaluation activities. Reports from between 1,500 and 1,800 clinics are received each month. Fifty states participate in data collection. About 40,000 admission and discharge reports describing clients admitted to and discharged from treatment are processed each month.

B. How CODAP Data are Analyzed: A large data file coupled with many demands for analysis requires automated procedures for table generation

and a variety of approaches to satisfy user demands. The Division of Scientific and Program Information, NIDA, has developed several approaches: (1) Periodic, usually quarterly, reports are prepared which present close to 100 tables. (2) Special issues are addressed in the Statistical Series, which describes applications to management of drug abuse programs, evaluation of treatment outcomes, and studies of patterns and factors associated with the development of drug abuse (epidemiology of drug abuse). (3) Data files are available less than five months after the data are collected. These files are provided to the Single State Agencies which coordinate drug abuse programs, and to an outside organization which in turn makes the files available to requestors or prepares tables upon request (at cost). (4) Technical assistance is provided to the states on how to use CODAP data. (5) Reports unique to each clinic/program are sent to those clinics/programs, together with comparable state/national data and suggestions for interpreting the data. (6) Special analyses are prepared upon request from federal government agencies.

In order to handle the large amounts of data involved, special analytic software has been developed which allows the following tasks to be performed automatically: (1) SPSS output is sent, via magnetic tape, to disk files for manipulation by text-editing software which produces camera-ready copy of tables. (2) Tables with a large number of variables (of the form A vs. B vs. C vs. D vs _.) are stored on magnetic tape. Another program reads those tables and produces summaries (collapsed over the categories of a given variable). In addition, for continuous, time-related variables, the output of both programs can be plotted using a

CALCOMP plotter. (3) Depending on the nature of the analysis, users can utilize extract files consisting of 20% and 1% samples of the data file, and also special subpopulations (such as daily heroin users) which have been found to be of specific interest. (4) A file of all tables computed from several of the larger files (such as the 100% sample) is kept as a reference. Requests are often answered from that reference system at a considerable savings in time and money.

ACKNOWLEDGEMENTS

We are grateful for the assistance and support of Dr. William H. Spillane and Mr. Neil Sampson, and the valuable comments provided by the staff of the Division of Scientific and Program Information, NIDA.

REFERENCES

- SIGUEL, E. N. and SPILLANE, W. H. (1976). The epidemiology of drug abuse; a new data base, new techniques and new findings. Proceedings of the Social Statistics Section of the American Statistical Association.
- SIGUEL, E. N. and SPILLANE, W. H. (1977). The client oriented data acquisition process (CODAP-77). Amer. J. Drug & Alcohol Abuse, IV(2).
- NIDA Statistical Series (1975 through 1977). A series of reports which primarily concern admission and discharge activity, client characteristics, types of drugs abused, and patterns of drug abuse; these variables are examined in relation to each other and to calendar quarter of admission, size of SMSA, and geographical region.

BIOGRAPHIES

E. Siguel received his Ph. D. in Mathematical Psychology from the University of Michigan. He has masters degrees in Physics and in Computers & Statistics and more than 16 years in statistical analysis with emphasis on large data files. His major interest is mathematical models applied to biological and social systems and he has written extensively on health care delivery and the epidemiology of drug abuse.

S. Sand received his B.A. in Political Science and Government from the American University. He has 9 years of systems analysis experience in mental health and drug abuse related data processing. His major interests are data base maintenance and documentation standards and techniques.