U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

DOT HS 808 597

July 1997

NHTSA Technical Report

Some Dimensions of Data Quality in Statistical Systems

This document is available to the public from the National Technical Information Service, Springfield, Virginia 22161.

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings and conclusions expressed in this publication are those of the author(s) and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturer's name or products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Technical Report Documentation Page

1. Report No.

DOT HS 808 597

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Some Dimensions of Data Quality in Statistical Systems

5. Report Date

July 1997

6. Performing Organization Code

7. Authors

Carl E. Pierchala, Ph.D.

8. Performing Organization Report No.

9. Performing Organization Name and Address

Mathematical Analysis Division
National Center for Statistics and Analysis
400 7th Street, SW
Washington, DC 20590

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency Name and Address

Research and Development
National Highway Traffic Safety Administration
400 7th Street, SW
Washington, DC 20590

13. Type of Report and Period Covered

NHTSA Technical Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

An important objective of a statistical data system is to enable users of the data to recommend (and organizations to take) rational action for solving problems or for improving quality of service or manufactured product. With this view in mind, this paper gives a list of some desired characteristics, or dimensions, of data quality, to be considered when building and maintaining statistical data systems. Some discussion of possible conflicts between different dimensions is given. Finally, data users and decision makers are encouraged to demand data systems of high quality, and system developers are urged to produce such systems.

17. Key Words

Quality, data quality, quality control, data quality characteristics, data system, statistical system, statistical data system, action

18. Distribution Statement

This document is available to the public through the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161

19. Security Classif. (of the report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. Of Pages

11

22. Price

ACKNOWLEDGMENTS

The author thanks Terry Klein, Dennis Utter, and especially Nancy Bondy, all of the National Highway Traffic Safety Administration (NHTSA), for suggestions and comments on previous work upon which this paper is based. Thanks also go to various NHTSA staff and to Michael P. Cohen for comments on drafts of the current paper.

1. INTRODUCTION

Somewhat surprisingly, when attending talks or reading papers that at least in part deal with data quality, I sometimes find that the authors say very little about what they mean by data quality. And often, perhaps only implicitly, data quality is taken to be the accuracy of the values in a computerized data set or database. Naus (1975) strongly emphasizes data editing as a way to eliminate errors in the values in a database, although he does also discuss methods used to deal with difficulties that occur in data collection and handling. He points out that some of these difficulties, such as coverage error, cannot be detected by editing. There are many interesting articles in the book edited by Liepins and Uppuluri (1990), but there is no explicit discussion of the meaning of the term "data quality." Falter (1981) gives a valuable discussion of data quality assurance, highlighting its importance in ultimately producing reproducible and valid scientific results. But he only implicitly defines "data quality". The purpose of this paper is to encourage the builders and users of data systems to take a broad view of the meaning and dimensions of data quality and to take action so that data systems of high quality are developed, maintained and appropriately utilized.

In his classic book on quality control, Shewhart (1931) devotes the fourth chapter to the meaning of the term "quality". After discussing the intuitive use of the term by the lay public as the goodness of a product, Shewhart goes on to give a more detailed discussion of the term. In essence, quality has to do with the characteristics that a product needs to have in order to have value to individuals. Thus, identifying the desired characteristics, or qualities, needed in the product is a foundation for the *control* of the "quality" of the product. The producer needs to define such qualities in measurable terms.

Some authors are careful when discussing the term "data quality". For example, O'Day (1993) explicitly gives eight dimensions for the quality of data on motor vehicle traffic crashes. Specifically, these are ascertainment (or completeness of coverage), consistency of coverage, missing data, consistency of interpretation, the right data, appropriate level of detail, correct entry procedures and freedom from response error. Savage (1976) identified documentation, timeliness, relevancy and balance in coverage (loosely, the tradeoff of cost vs. numbers of groups, variables and individuals studied). Dalenius (1983) used the term 'quality measure vector', in which he included as a minimum such things as accuracy, cost, privacy protection, relevance, timeliness and wealth of detail. Spencer (1985) states there is no unique metric for data quality, suggesting most generally the use of a risk function of an estimator. He distinguishes between data quality and the quality of a data program: the latter includes data *quantity*, detail, relevance, timeliness, analysis, dissemination and documentation. Huh, Keller, Redman and Watkins (1990) give four dimensions of data quality: accuracy, completeness, consistency and currency. Redman (1992), from a very broad and general information management system perspective, gives 27 dimensions of data quality.

Many authors address the issues of nonsampling error and total survey error. Bailar (1983) discussed the notion of error profile, which she characterized as "a systematic and comprehensive account of survey operations that yield survey results." She went on to suggest that the name could be changed from "error profile" to "quality profile". While an error profile indeed is an important aspect of data quality, keep in mind that it primarily deals with the dimension of accuracy and precision in estimates. Lessler and Kulka (1983) listed many sources of error in measurement. Again, most of these have to do with accuracy and precision.

The National Center for Statistics and Analysis (NCSA) of the U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) addressed the issue of the dimensions of data quality of a statistical data system in its draft quality plan (NCSA, 1993). This was done largely from the point of view of a survey or a census, but many of the dimensions discussed clearly apply to experimental or observational studies as well. Most of these dimensions will also be relevant to the statistical data systems and data sets of organizations other than Federal agencies.

One key point of view in developing NCSA's list of data quality attributes came from Deming (1975), who pointed out the importance of considering how to use statistics so that they help in taking action. Thus, important dimensions of data quality are those that ultimately help in using the data to take action.

In the next section, we give a number of characteristics, or dimensions, of data quality. Most of these are taken from the NCSA quality plan. Also given is a discussion of some potential conflicts between some of the characteristics. The reader may think of other data quality dimensions or other possible conflicts between dimensions.

2. DIMENSIONS OF DATA QUALITY

By the term "statistical data system" is meant a set of subsystems to collect, store, document, and disseminate data that are primarily intended to produce statistics to help in taking rational action to accomplish well-defined objectives usually related to solving some problem(s) or improving quality of products or services. A statistical data system does not necessarily reside on a computer system, although usually it will because of various advantages thus achieved. Indeed, computerization of the data is characterized below as a dimension of data quality.

To be most useful to help in producing information for taking action, statistical data systems need to possess a variety of characteristics, or dimensions. The extent to which these are present defines the quality of a given data system. As implied in the definition above, the term "statistical data system" is being used in a very broad sense. In part, this includes the system objectives, the sampling/study design and the measurement, recording, coding, editing, processing, storage, retrieval, analysis and reporting of the data. To help in providing a basis for evaluation and action to assure the quality of a data system, a broad set of fairly generalized characteristics is delineated below in terms of user needs.

2.1 A List of Some Data Quality Dimensions

A user wants most if not all of the following characteristics in a data system:

1. The data are appropriate for addressing issues of interest.

    - The appropriate population(s) or experimental material(s) are studied.

    - Variables appropriate for the issues of concern are collected.

- Appropriate selection, measurement, recording and coding techniques are used in collecting the data.
- When needed, new variables are added to the data, and measurement methods are revised.
- Changes/modifications are done in such a way that the ability to analyze the data is not jeopardized.

2. The data accurately represent reality.

- Sampling variability in estimates is reasonably low.
- There are minimal biases in estimates because of sampling methods, research design and other sources.
- There are little if any missing or incomplete data.
- Observation, recording, coding and entry into electronic storage are reasonably accurate in most if not all cases, and edit checks are effective. Thus, errors in the data are relatively small and/or infrequent.
- Standard operating procedures are followed in all phases of building the data system.
- Definitions are clearly delineated and correctly interpreted.
- The data are logically consistent.
- For trend analysis, there should be consistency over time in data collection and coding.
- In distributing the data to local and external users, care is taken to precisely identify the data release and to avoid the introduction of corrupt copies of the data.
- Valid and appropriate statistical methods are utilized.

3. The data are computerized.

- Generally speaking, the user wants the data to be on a computer system for ease and speed of processing and analysis.
- The structure of the computerized databases should allow for great flexibility of analysis.
- Appropriate and flexible statistical and data management software are available and used for analysis and processing of the data.

- Help is available in selecting appropriate analyses.

- The data can be easily distributed to other computer installations.

4. The data are easy to use.

   - The data structures on the computer are relatively simple.

   - The software for data manipulation is easy to use.

   - Data coding is designed for ease of use.

   - Data are easily and quickly accessible.

5. Documentation is readily available and easy to use.

   - The user does not have to guess or make assumptions about what the data mean, how they were collected, how computations were performed or how the data are stored.

   - Known limitations of the data are documented and readily available to the user.

   - The data system and/or study objectives are clearly stated in writing.

   - Changes over time in objectives and in all aspects of data collection and the processes and procedures for computerizing and for using the data are clearly documented in writing.

   - User manuals are available which contain the information necessary to use the data system.

   - When used, imputation methods are clearly and completely described, and imputed values are flagged in some way so as to give the user the flexibility to employ alternative methods for handling missing data.

6. The data are available in a timely fashion.

   - There is minimum delay between the data collection and their availability for analysis.

7. Costs are minimized.

   - Other considerations being equal, operational methods that lower costs are desirable.

8. Respondent rights are respected and maintained.

   - Providing information is not overburdensome to respondents.

   - Confidentiality is maintained.

2.2 Interplay between the Dimensions

Various comments come to mind concerning the above data system quality dimensions. In particular, building certain of the characteristics into a data system may preclude the inclusion of other characteristics. Many of the items discussed below have to do with tradeoffs between speed, convenience and accuracy. (For ease of reference, category numbers are used below that correspond to those above.)

1'. Appropriate data.

- More than one data system or study design may be needed to address varying issues and objectives. For example, with a specific data system it may not be possible to address a given issue because in the context of that data system it may not be possible to collect some of the relevant variables or study certain of the subpopulations of interest. As one illustration, most of the States' reporting systems for motor vehicle crashes do not contain the type of data needed to make detailed assessments of types and patterns of injury in automobile crashes. However, by linking these data to suitable hospital data, it may be possible to make an appropriate assessment.

2'. Data represent reality well.

- Effort and expense will be needed to assure that the data represent reality with reasonable accuracy.

- A formal mechanism is required to ensure that changes in the data collection and coding procedures will provide useful information. This should include input from field operations staff, data processing staff, statistical staff and users, among others.

- Preventing the occurrence of missing/incomplete data is the most reliable way to avoid bias from this source. It is difficult to know the magnitude of the bias that occurs when data are more than minimally incomplete.

- A complex sample design may be used to increase the precision of estimates in subpopulations of special interest or to reduce total data collection costs while maintaining overall precision. This will complicate some of the analyses, however. The documentation should carefully describe the sample design so that the user will employ appropriate methods of analysis.

- Corruption of data in the dissemination of a data release, or failure to properly identify the version of the data (and subsetting criteria) used in an analysis, may create confusion and waste of time. Also, care needs to be taken when data are released in more than one format to prevent idiosyncracies between formats from introducing inconsistencies in what should be equivalent data sets.

- In comments about the analysis of energy data, Loebl (1990) asserts that taking data out of context is the single most important source of error.

- Loebl (1990) also states that systematic errors are much more significant than random errors in large-scale data collection systems, at least in the arena of energy data. But clearly, there are systems in which random errors dominate. Both kinds of errors need to be considered.

3'. Computerized data.

- Again, expense needs to be incurred for the benefits of ease and speed of data retrieval and analysis.

4'. Data easy to use.

- The nature of the study issues or the study design may complicate the data structure, making it harder to use. For example, the data may have a hierarchical structure, which complicates analysis but enriches the analyses possible. For instance, in order to study the effect of vehicle characteristics on occupant outcomes in motor vehicle crashes, data are needed at both the vehicle and person levels.

- Note that there is often a tradeoff between simple-to-use software and software capable of flexible and/or complex analysis. An analysis appropriate to the design of the survey/experiment/study is the overriding consideration. Ignoring the study design in doing analysis is the intellectual equivalent of failing to use seat belts when operating a motor vehicle; in the short run, it appears harmless, but in the long run, it creates extra trouble.

5'. Documentation is needed.

- Without documentation, data are harder to use and likely at times to be misused.

- Yet again, expense needs to be incurred to prepare documentation.

- Over time, changes to a data system are necessary. As appropriate, variable mappings or other aids should be developed to facilitate trend analysis.

- Savage (1976) gave several comments about documentation and had harsh words about what he perceived as a pervasive lack of adequate documentation.

6'. Timely data.

- Some delay is unavoidable because time is taken by the processes required to collect information and build sound databases.

- Sometimes relevant variables are not available to address new issues that may arise, so there will be a lag time before the data system includes appropriate new variables.

7'. Low-cost data.

- There is likely to be a tradeoff between cost and accuracy. Too excessive a desire for low cost may yield inaccurate or misleading data.
- Related to the above, a reduction in sample size will lower cost (design remaining unchanged), but the precision of estimates will be reduced.
- It may be desirable to obtain more accurate data for certain variables of interest, by spending more on those variables.
- The more likely that specific data will be useful in helping to take action and the more important the issue, then the greater the willingness should be to increase spending in order to obtain the desired quality characteristics for those data.
- When a complicated experimental design or complex scientific sampling is used to keep cost down (or to increase precision in special subpopulations), ease of use decreases.
- When surrogate variables are used to keep measurement costs low, increased uncertainty concerning the validity of conclusions will occur, possibly hindering action.
- If funding becomes inadequate, restructuring of priorities and the magnitude of the data system efforts will be needed to maintain the integrity and utility of those data that continue to be collected.
- A case can be made that cost should not be considered as a dimension of data quality. However, for convenience cost has been included in this list of data characteristics. It is well to keep in mind Deming's (1960, p. 31) statement that "cost has no meaning without a measure of quality."

8'. Confidentiality and Burden

- Actual legal requirements may mandate confidentiality.
- Some variables that are otherwise appropriate to an issue of concern may not be collected to protect confidentiality.
- Timeliness, ease of computerization, ease of use and cost minimization may be hindered by the need to protect confidentiality.
- Excessive burden on respondents may lead to low accuracy in the data.
- Respondent perception that confidentiality may not be maintained may lead to low accuracy in the data.

3. DISCUSSION

Federal statistical agencies have a variety of responsibilities in the collection, storage, analysis and dissemination of data and statistics. An agency's databases are used to provide analyses and information to agency staff, other Federal agencies, Congress, state and local government, universities, private and nonprofit organizations, the media and individual citizens. Thus, an agency has an interest in ensuring that its data systems are respected as credible and useful.

Similarly, nongovernmental organizations have an interest in ensuring that their databases are of good quality. For example, Redman (1992) argues that commercial organizations may gain competitive advantage by increasing the quality of their data.

Data, statistics and information are useful and important to the extent that they reliably serve as guides to taking action. To be useful in this way, data systems must have a variety of qualities, or characteristics. This was discussed in some detail above. It behooves an organization to consider the qualities needed in data systems, and to then take cost-effective steps to ensure that its data systems have the appropriate characteristics.

Since the ultimate goal is to take prudent action on various issues of relevance, thought needs to be given to the entire process of obtaining and using data. Loosely speaking, this begins with the definition of purpose and of the data needs. It includes such items as training of the data collectors; methods for identification and sampling of basic observational units; and goes on to data collection, validation, error checking, storage, and analysis. Finally, the process ends with actions designed to make positive and cost-effective impact on the issues of concern.

Keeping in mind the interplay of statistical data system quality characteristics, the qualities needed for specific systems can be defined, and actions needed to improve and assure the quality of the data systems can be initiated. No system is ever perfect, and quality improvement is usually made a step at a time, by thought and effort. Teamwork is one of the tools in a quality improvement program. However, a detailed discussion of such a program is beyond the scope of this paper. There is a large literature on quality improvement, including books by Deming (1986, 1993).

4. SUMMARY

Accuracy of values recorded in a statistical data system is important. But as discussed above, there are other data quality dimensions to consider when planning, building, maintaining and using a statistical data system. Building in those characteristics that enable the data to guide prudent and rational action is critical. Ultimately, data systems and data sets should help people to take action toward solving problems and toward improving quality of service or manufactured product. Decision makers and data users should be aware of the dimensions of data quality and demand systems of high quality. Data system developers and managers should be responsive to these needs and take the steps required to produce and enhance such systems.

REFERENCES

Bailar, B.A. (1983), "Error Profiles: Uses and Abuses," in *Statistical Methods and the Improvement of Data Quality*, ed. T. Wright, Orlando, FL: Academic Press, pp. 117-130.

Dalenius, T. (1983), "Errors and Other Limitations of Surveys," in *Statistical Methods and the Improvement of Data Quality*, ed. T. Wright, Orlando, FL: Academic Press, pp. 1-24.

Deming, W.E. (1960), *Sample Design in Business Research*, New York: John Wiley.

______(1975), "On Probability as a Basis for Action," *The American Statistician*, 29, 146-152.

______(1986), *Out of The Crisis*, Cambridge, MA: MIT Center for Advanced Engineering Study.

______(1993), *The New Economics for Industry, Government, Education*, Cambridge, MA: MIT Center for Advanced Engineering Study.

Falter, K.H. (1981), "Data Quality Assurance," in *Statistics in the Pharmaceutical Industry*, eds. C.R. Buncher and J. Tsay, New York: Marcel Dekker, pp. 301-326.

Huh, Y.U., Keller, F.R., Redman, T.C., and Watkins, A.R. (1990), "Data Quality," *Information and Software Technology*, 32, 559-565.

Lessler, J.T. and Kulka, R.A. (1983), "Reducing the Cost of Studying Survey Measurement Error: Is a Laboratory Approach the Answer?," in *Statistical Methods and the Improvement of Data Quality*, ed. T. Wright, Orlando, FL: Academic Press, pp. 245-266.

Liepins, G.E., and Uppuluri, V.R.R. (eds.) (1990), *Data Quality Control: Theory and Pragmatics*, New York: Marcel Dekker.

Loebl, A.S. (1990), "Accuracy and Relevance and the Quality of Data," in *Data Quality Control: Theory and Pragmatics*, eds. G.E. Liepins and V.R.R. Uppuluri, New York: Marcel Dekker, pp. 105-143.

Naus, J.I. (1975), *Data Quality Control and Editing*, New York: Marcel Dekker.

NCSA (1993), "National Center for Statistics and Analysis Quality Plan," Document No. PF 88-01-CD3(h) in *Public File 88-01 (Motor Vehicle Safety Research Advisory Committee, Crash Data Analysis Subcommittee Meeting, October 6, 1993)*, National Highway Traffic Safety Administration, Technical Reference Division, Washington, DC 20590.

O'Day, J. (1993), "Accident Data Quality," National Cooperative Highway Research Program Synthesis of Highway Practice 192, Washington, DC: National Academy Press.

Redman, T.C. (1992), *Data Quality: Management and Technology*, New York: Bantam.

Savage, I.R. (1976), "Considerations of Data Quality," in *Setting Statistical Priorities*, ed. National Research Council, Washington, DC: National Academy of Sciences, pp. 101-104.

Shewhart, W.A. (1931), *Economic Control of Quality of Manufactured Product*, New York: Van Nostrand.

Spencer, B.D. (1985), "Optimal Data Quality," *Journal of the American Statistical Association*, 80, 564-573.