Obtaining Acceptable Quality Data from Carload Waybill and Other Samples

A. C. ROSANDER
Chief, Mathematics and Statistics Staff, Bureau of Transport Economics and Statistics, Interstate Commerce Commission

The purpose of this paper is to describe the various factors that determine the quality of data obtained from a probability sample study, to explain what is meant by "quality," to give examples illustrating quality of data with special reference to the Interstate Commerce Commission's continuous carload sample, and to describe steps that can be taken in the planning, implementation and presentation of a probability sample study to control the quality of the data obtained.

•STATISTICAL PROGRAMS and sample studies are planned not simply for the purpose of compiling numerical data which may be useful sometime, but as an integral part of an information system which aids management to make better decisions, exert closer control, appraise operations more accurately, and set policy more intelligently. This means that close attention is paid to the early stages of the study, to detailed analysis of the problem, definition of the population, isolation of basic characteristics, and enumeration of significant classes and subclasses for which data are required. In other words, close attention is paid to the purpose for which the information is needed, so that only the most significant information is collected.

There are many reasons for stressing analyzed purpose, but the most important is to collect the minimum amount of data required to meet a specified need, thereby saving time, money and personnel. The older notions that facts speak for themselves, that there is virtue in large masses of data, and that data collected for no specific purpose might sometime prove beneficial have been found to be wasteful guidelines. Simply collecting or compiling a mass of data is no longer enough.

Experience with applied probability sampling during the past 25 years has substantiated the following important ideas:

1. The quality of the information is as important as, if not more important than, the quantity of data.
2. Redundancy exists in numerical data (and, therefore, justifies sampling).
3. Numerical data have to be interpreted in terms of how they are collected or, to use an expression coined by R. A. Fisher, in terms of the logical structure. An efficient logical structure is one which can be expressed in terms of probability and mathematical statistics.
4. Data collected on a probability sample basis can be analyzed by mathematical statistics, thereby aiding both research and management to make more accurate and meaningful decisions about operations, planning, and policy.
5. A properly planned, designed, and managed probability sample can give quality data rapidly at a minimum cost.
6. Sources of nonsampling variation may equal or even exceed variations due to random sampling. A very important information problem is that of getting quality data at the source since the source is often the major cause of nonsampling variations, due to the respondent, the collector of the data, or to both. This means that a prob-

Paper sponsored by Committee on Intercity Highway Freight Transportation and presented at the 44th Annual Meeting.

ability sample properly designed and managed may actually prove better (have less total error) than a census or 100 percent tabulation that is not carefully planned and controlled.

7. Neither a purely inductive (statistical data) nor a purely deductive (model building) approach to information is most effective, but some combination of the two is required.

CHARACTERISTICS OF ACCEPTABLE QUALITY DATA

Some of the major characteristics of acceptable quality data derived from a probability sample are as follows:

1. The standard error of an estimate is known.
2. Since the variation due to sampling is known, assignable causes can usually be separated from random variations. Therefore, differences due to sampling can be distinguished from those arising from other sources. Therefore, statistical techniques, such as analysis of variance, can be applied to the data to clarify the meaning of differences.
3. Nonsampling variation is controlled so as to minimize or measure its effect, e.g., nonresponse or loss of sample elements due to cutoff date. This means that the sample study is properly managed and controlled, and that techniques such as a sample audit are used to measure nonsample variation.
4. If nonsampling error can be estimated, the magnitude of the total error can be computed.
5. The sample data answer specific questions of magnitude, level, control, comparison, and frequency of occurrence.
6. Basic concepts and terms are explained in enough detail that the user can distinguish them from slightly different concepts or terms. This is a problem of exposition.
7. Basic concepts and terms, including any questions on a questionnaire or items on a data sheet, are explained in operational terms so they can be distinguished from other concepts and so that answers to questions and data entries are additive. This comparability is facilitated by pre-tests, pilot studies, and written and oral instructions. Where many persons are involved in collecting the data, it is imperative that they have a common understanding of the information desired; hence, written as well as oral instructions are needed, with follow-up on questions and problems, statistical quality control techniques, and other controls designed to insure that the information is additive and not a function of some characteristic of those collecting the data.
8. Acceptable quality data also imply an adequate technical sample designed in terms of purpose and population, and adequate management of the sample so that the plan will be properly implemented. At every stage, the integrity of the data must be maintained so that the use of the powerful methods of statistical interpretation will be justified.
9. The data are relevant to the questions to be answered or the problems to be solved.

EXAMPLES ILLUSTRATING QUALITY IN DATA

The following examples illustrate what is meant by quality in data:

1. A careful probability area sample showed that about 25 percent of the businesses had been omitted in the French census of business for 1947. Therefore, the census data was of no value and was never published.
2. Revenue per ton-mile for railroads as obtained from annual reports is less than revenue per ton-mile obtained from the carload waybill sample because the former is derived from actual miles and the latter is derived from short-line miles. This means that in the expression $v = u/(xy)$, where u is total revenue, x is tons, and y is miles, the value of y is greater in the annual reports than in the carload sample. Hence, the ratio v is less. There is no problem here if the term miles is clearly defined in both instances.

3. On some piggyback-type waybills, the term net weight may mean the same as gross weight on other waybills. Furthermore, the concept of weight may include actual commodity weight in some instances and an arbitrary minimum weight in others. We "make" these weights additive by calling them billed weight. Obviously, this does not help the person who is interested in the total actual tonnage of commodities hauled. A correction factor could be obtained and applied to billed weight to get actual commodity weight in the same way as a circuity factor is applied to short-line miles to get actual miles.

4. In a proceeding before the Interstate Commerce Commission (ICC), a cost formula of the form $C = (Mcy)/x$ was presented where values for M and c came from accounting records, and x and y from a sample. Sampling errors were confined to y; x was ignored as were the possible errors in M and c. In this case, the standard error of C was needed, but this computation was not made. Considerable time was spent by the parties in the case discussing the sampling error in y when this magnitude was really not the error required.

5. The applicability of sample data collected for one day or one week to longer periods such as an entire year is questionable. Usually it is assumed, rather than proven, that data from a short-time population are applicable to a much longer-time population. The solution lies in using the time population inherent in, or applicable to, the problem.

6. In one of the carload waybill publications, the revenue per car-mile for all movements was shown as $0.60 for each of the years 1959 and 1960. The standard error of these estimates is about 0.01, and using one more digit gave $0.601 for the 1959 value and $0.598 for the 1960 value. A very significant difference of $0.003 existed between the two years but too much rounding off eliminated it.

7. In a random time study of personnel activity during one week, 25 random minutes are selected and one group of workers is observed each minute. The total number of persons for which a record is made is 225. The sample size is 25, not 225, and the standard error should be based on 25 values, not 225; otherwise, the standard error is one-third as large as it ought to be.

8. The waybill sample shows 25 carloads of a certain commodity movement for one year and 36 for the same commodity the next year. The probability is very high that this difference is due to sampling, not to a change in the commodity movement. In sample data, differences must be tested to determine whether or not they could be explained in terms of random sampling or its equivalent.

CARLOAD WAYBILL SAMPLE

The ICC's sample of waybills of carloads terminating on Class I railroads in the United States illustrates the problems which arise in connection with the quality of the data on the waybills, and the steps taken to insure acceptable quality:

1. Control over sample receipts by railroads is necessary to secure timely and complete shipment of the sample bills. A follow-up system is applied monthly and at the end of the calendar year to keep missing sample bills at a very low level.

2. A quick review of sample waybills is made on receipt to screen out those obviously in error or defective (such as those wholly or partially illegible).

3. Form letters are used to return waybills found in error to railroads for correction. These errors include missing stations, wrong station numbers, questionable commodity codes, inconsistency of type of car and commodity carried, and uncertain type of rate.

4. Consistency checks are programmed on the computer to read out waybills containing inconsistent information such as weight of car in relation to type of car.

5. About 90 percent of the waybills are miled mechanically, thus eliminating a major source of error. About 2,500 waybills are still manually miled monthly.

6. Statistical quality control techniques are now being developed or applied to commodity coding, to other coding operations, and to manual miling to maintain control over the error rate on waybills before they go to the punch room. This program is not fully operative but tests to date indicate that it is needed and would be effective.

Additional checks are made after the tabulations are run. Cell-by-cell comparisons are made with the tabulation for the previous year. In a few instances this review has detected some tabulation errors. Finally, the publications are reviewed to check for missing or illegible pages and omission of commodity identification. In the publications the limitations of the data are described and the limits of the population are drawn.

MAJOR FACTORS AFFECTING QUALITY OF DATA

Source

One of the major factors affecting quality is the source of the data. The respondent may not answer accurately or completely; he may misunderstand what is wanted or he may deliberately slant, withhold or distort the information. The collector of the data may not ask clearly worded questions, the right questions, or nonleading questions. The questionnaire and the data sheet may have deficiencies which affect the information. Very careful and complete planning is necessary to obtain quality of data at the source. For example, pre-testing data sheets, instructions, questionnaires, and even sample plans help to weed out unwanted and unwarranted deviations.

Analysis of Problem

Many times the crucial factor in determining the quality of the data is the analysis of the problem itself. This includes phrasing of specific questions, defining the population, identifying basic characteristics to be measured, and listing classes and subclasses for which data are required. When specific questions are formulated, effort is concentrated on collecting only those data giving unambiguous answers to these questions. Implicit assumptions should be made explicit, and subject to test by means of the data. Otherwise, the entire study may rest on an assumption or conjecture which may seriously impair the validity of the entire project.

Some years ago, a sample study of American families was made eliminating the foreign-born because it was assumed that they were different from native-born families. Before the study was over, however, estimates for all U. S. families were necessary and the foreign-born had to be included. Therefore, the original position was reversed and it was assumed the foreign-born were like the native-born at the same income class! These inconsistencies and unnecessary truncations of the population can be avoided by a careful analysis of the problem and the population at the outset.

Definition of Terms

Numerical data, whether derived from sample or census, are subject to semantic problems. These problems arise at some critical points: (a) in analyzing the problem; (b) in phrasing instructions, data sheets, etc., for collecting the data so that the respondents know what is wanted (part of this job is to explain clearly the various units of measurement); and (c) in explaining what these terms mean in various reports and publications. Much of the trouble lies in the fact that the terms are not explained carefully enough that the respondent, the reader, or the user knows what they mean. The carload waybill publications contain sections defining all major terms used in the tables, including ton-mile, carload, car-mile, revenue, and billed weight.

Use of Probability Sampling

A very effective way of obtaining data of acceptable quality is to take full advantage of probability sampling and the associated statistical analysis. This assumes, of course, that the probability sample is carefully designed and effectively managed when put into operation.

Probability sampling has the following merits:

1. It provides a way of measuring and controlling sampling variability (the standard error of an estimate);
2. It provides a built-in method of estimation;
3. It forces a better control over nonsampling errors;

4. It forces a more detailed analysis of the problem;
5. It provides data for its own progressive improvement;
6. It eliminates the need to assume that a sample is representative, a judgment often hard to defend; and
7. Mathematical statistics can be used to interpret the data.

The last point is of special importance because it allows us to use the sciences of probability and statistics to interpret the data, to distinguish real differences and relationships from apparent ones. The Shewhartian chart used for statistical quality control is an excellent example of how theory and practice of mathematical statistics are combined into a simple, yet very effective, device for decision making and control.

Random Time Sampling

A very important type of probability sampling found in work measurement, cost accounting, and other cost work is random time sampling. This method is used where operations must be observed as they are taking place. A suitable time frame is established and units of time, either instants or durations, are selected at random. At these random instants or durations, observations are made of workers or machines, of street or road traffic, or of persons entering supermarkets. In this way, accurate estimates of work performance, traffic density, and similar characteristics can be obtained. The randomly selected time unit may be a minute, an hour, a day, or some other unit, depending on the problem.

Design of Sample

The technical design of the sample is another factor in determining the quality of the data, but it is often the easiest problem to solve. Established standards of practice will be followed in designing the sample. Many different designs are possible, all of which may be valid, but the problem is to design that sample which is most efficient and most feasible, yet meets the specifications of variation imposed on the final results or the risks imposed on the final decisions. It is the statistician's responsibility to make clear what kinds of information the sample can provide and what kinds of data it is not designed to furnish.

Control Over Sampling Variability

This can be exerted in a number of different ways: (a) by the size of the sample, (b) by the method of estimation, (c) by stratification, and (d) by type of sampling unit. Usually, this control is not difficult to exert. Although control by size of sample is most common, the other methods can be very effective.

Pilot Studies and Pre-Testing

A sample survey can be greatly improved by making use of pilot studies including pre-testing of data sheets, forms and questionnaires. These exploratory tests or studies enable one to:

1. Test a questionnaire for ambiguity, misunderstanding, inconsistencies;
2. Estimate standard deviations and other quantities needed for determination of size of sample and efficiency of stratification;
3. Test personnel reaction, whether favorable or unfavorable;
4. Determine areas, questions, and operations that need to be emphasized when preparing instructions;
5. Test data recording forms;
6. Test control forms;
7. Examine possible frames for use in selecting the sample; and
8. Appraise the type of personnel available to implement the sample.

Example of Pilot Study

A good example of a pilot study was the five-day random time sample used by the Internal Revenue Service preparatory to introducing a continuous random time sample to determine salary costs of various projects for purposes such as of budget planning and work scheduling. The five-day test yielded information which was used to improve instructions and sample design and to simplify operations. The pilot study data in this instance were studied and procedures changed so that about four months elapsed before the final random time sample was put into effect. Much of this time was spent redesigning the sample, revising instructions, preparing activity code books, and otherwise improving procedures.

Management of Sample

The implementation of a probability sample study requires careful management of the entire project. This is done to keep nonsampling errors under control and to see that the technical sample design is carried out as planned. Careful management calls for detailed planning, design of various data sheets and control forms, preparation of instructions and any other necessary materials, and the effective implementation of these plans and controls.

Control Over Nonsampling Variability

Quality of data is preserved chiefly by controlling nonsampling variability, and here the major problem lies in getting acceptable quality data at the source. The problem does not arise primarily because some data involve opinions or judgments and others do not; it can arise even if the information is objective and subject to measurement. The problem consists of getting accurate and unambiguous data from the source, and this holds true whether we deal with income tax returns where most of the basic data are money figures, with carload waybills filled out by a railroad station agent, with freight bills filled out by a motor carrier (truck line), or with a cost-of-living study where all data are money figures or quantities furnished by a member of the household.

Statistical Analysis

Statistical analysis is needed to give logical quantitative meaning to the data; otherwise, misleading inferences may be drawn. We may conclude that a difference is due to economic factors, when it could very easily be due to sampling variations. This statistical analysis needs to be made before other specialists attempt to interpret the data in terms of their own subject matter fields. Some statistical analysis is now included in the carload waybill publications, and this will probably be in forthcoming issues.

CONCLUSION

To summarize, the quality of data depends on careful execution of the following procedures:

1. Define problem and terms;
2. Analyze problem into details;
3. Phrase questions carefully;
4. Use pre-testing and pilot study, and study the source;
5. Train personnel;
6. Prepare written instructions and other materials;
7. Design adequate probability sample and use random time sampling as needed;
8. Manage probability sample properly;
9. Apply statistical analysis implied in collection procedure;
10. Compute standard errors of estimates and functions;
11. Control nonsampling error;
12. Estimate total error;

13. Explain basic terms so the reader knows what they mean and distinguish them from similar or identical terms used elsewhere; and
14. Use sample audit and other control methods.