Total Quality Management,
Vol. 15, No. 1, 89–103, January 2004

Data Quality in Statistical Process Control

PATHIK MANDAL
SQC & OR Unit, Indian Statistical Institute, Baroda, Gujarat, India

(Received September 2002; revised April 2003; accepted July 2003)

ABSTRACT *Apart from the organizations conducting large-scale countrywide surveys, most of the large business houses today maintain large databases for controlling and improving their business processes. Naturally, data quality is a critical issue for these organizations. Consequently, data quality has emerged as a very important field of research. However, most of the research conducted is either generic in nature or deals with the problems faced in survey data. In this article, an attempt has been made to develop a framework for addressing the data quality problems that may be faced in implementation of Statistical Process Control. Specifically, a new set of data quality classes is proposed and various data quality problems are classified accordingly.*

KEY WORDS: Data quality scale, data collection process, causes of poor data quality, data collection planning

Introduction

We are in the data age. Data are everywhere. Modern business organizations, in particular, are now collecting an unprecedented amount of data for controlling their business processes. In fact, it is likely that some of these organizations are suffering from data overload. However this does not mean that these companies are also experiencing information overload. Why do we find this gap between data and useful information? In our opinion, it is not so much due to non-use and/or misuse of statistical techniques as is due to the fact that most data are of poor quality.

Most data are indeed of poor quality, although the concerned organization or the department may not recognize the same. A recent survey (Price-waterhousecoopers, 2001) has revealed that 75% of the 600 companies surveyed reported significant data quality problems. Firth (1996) reports that a fibre-optics manufacturer lost $500,000 when a wrongly labelled shipment caused the wrong cable to be laid along the bottom of a lake and a brokerage firm lost $500 million when a dealer entered an incorrect exchange rate. Dubois (2002) reports that, in the customer database of an insurance company, 62% of the names and 80% of the addresses were mere duplications. Elimination of such huge redundant

data not only reduced the processing time and cost of contacting customers substantially but also provided greater confidence the users have in their own data.

Informally, the data users have always been concerned about quality of the data being used. However, it is only in the recent past that the issue of data quality is being discussed explicitly. This is particularly true in the field of information technology. Others who have shown significant concern for data quality are the organizations conducting large-scale surveys and financial institutions maintaining large databases. The concern for data quality outside these groups is minimal. The references cited in this article attest to the prevalence of such a scenario. In particular, it is indeed disheartening to note a lack of significant interest among the researchers and practitioners of quality management since data play a central role in this field. Nevertheless, it may provide some comfort to a quality practitioner to note that researchers on data quality have drawn heavily from the literature on quality management.

The motivation for this work came while redesigning our short-term training courses, such as Statistical Process Control (SPC) and Six Sigma. It was felt that at least one session of one and a half hours should be devoted to the topic of 'data quality'. However, we could not find sufficient material for this purpose in the literature since most of the discussions are either generic in nature or related to problems faced in survey data.

In this article, our primary aim is to indicate various types of deficiencies that are likely to be associated with data generated for the purpose of process control and improvement. Throughout this article, by data we shall mean 'zeroth order data' (Davidson, 1996), i.e. data that have been keyed on to a database and not those which come out as output of statistical analysis.

Related Work

The developments in data quality management, both conceptually and methodologically, are closely related to those available in the general literature on quality management. This is not surprising since data can also be viewed as a product, i.e. an output of a process.

Definition of Data Quality

There are at least as many definitions of 'data quality' as there are of the term 'quality'. This is because most of these definitions are simple adaptations of the popular definitions of 'quality'. For example, according to Strong *et al.* (1997), high quality data are 'data that is fit for use by data consumers'. Clearly this echoes Juran's definition of quality. Similarly, following ISO 8402, Abate *et al.* (1998) opine that 'data is of required quality if it satisfies the requirements stated in a particular specification and specification reflects the implied needs of the user'. However, Orr (1998) presents a somewhat different but narrow view of data quality. Here, data quality is defined as 'the measure of the agreement between the data views presented by an information system and that same data in the real world'.

Table 1. Data quality categories and dimensions

Category	Dimension
Intrinsic	Accuracy, Objectivity, Relevancy, Reputation
Contextual	Value added, Relevancy, Timeliness, Completeness, Appropriate amount of data
Representational	Interpretability, Ease of understanding, Representational consistency, Concise representation
Access	Accessibility, Access security

Data Quality Attributes, Dimensions and Categories

It has long been recognized that data, like other products, have many attributes (King & Epstein, 1983; Agmon & Ahituv, 1987). The first step in measurement of data quality is to identify these attributes as applicable in a given situation. The ability to download, accuracy, compactness, quantity, timeliness, and traceability are a few examples of data quality attributes. However, in any given situation, the number of quality attributes may be too many and also the attributes may be highly correlated. To illustrate the possible correlation among attributes, consider the case of a market research organization, which conducts a countrywide retail store audit on an ongoing basis. It receives data from the field and these are processed in a central location. There is a due date for sending the audit report to the customers. Now, if data from the field arrive late, it is clear that all the attributes such as accuracy, timeliness, and completeness will be affected simultaneously.

Thus, instead of using the attributes for data quality measurement, it may be helpful to make use of a small set of comprehensive quality dimensions. Wang & Strong (1996) started with 179 attributes and used factor analysis to collapse the attributes into 15 dimensions. They also clubbed the dimensions into four quality categories (Table 1).

Data Quality Assessment

In the literature, data quality assessment refers to both assessment of a given database and that of an organization's data quality management system. In a database assessment framework, the quality dimensions play a central role. Pipino *et al.* (2002) suggests that each dimension be assessed both objectively (using a suitable metric) and subjectively. They also suggest that the measurements on all the dimensions may be combined suitably to form a 'data quality index', but one must be careful in using such an index. In another approach, the US Environment Protection Agency (2000) uses both the database and the data collection process involved for assessing whether the quality of the database is satisfactory or not.

International Monetary Fund (2001) has developed a framework for assessment of data quality from the organizational perspective. Structurally, this framework is very similar to those used for evaluating organizational performance

excellence (e.g. Malcolm Baldrige National Quality Award). But unlike the Baldrige Award model, the proposed data quality assessment framework is not recommended for rating different systems since all the elements involved may not be applicable in all cases and the degree of subjectivity involved may be high.

Process Control Tasks

In our discussion on data quality we shall assume that the data are to be used for the following purposes:

- Disposal of product/lot
- Estimation of lot quality
- Process approval and adjustment
- Process capability analysis
- Process monitoring
- Problem solving

Data Quality Scale

It is apparent from the above discussion that the issue of data quality having such a wide canvas cannot be discussed in a short-term course on SPC or Six Sigma. The discussion ought to be focused on issues related to process control. To this end, we first propose a data quality scale. The data quality classes in this scale have been formed keeping in mind the data quality problems usually faced in process control.

As a first step in developing the scale, following Taguchi & Wu (1979), we take a narrow view of data quality. In this view, product features are distinguished from product quality characteristics. Product features are those characteristics that are either assigned to the product (e.g. price) or vary from customer to customer depending on his/her needs (e.g. colour of a shirt, strength of liquor). Such characteristics were termed 'non-quality characteristics' (Mandal, 1997). The implication of this narrow view is that the two dimensions of data quality, i.e. accessibility and security (see Table 1) get eliminated from our scope of data quality.

Secondly, for the sake of simplicity we shall also leave out the aspects of perceived quality, measured by the dimensions such as believability and reputation (Table 1). In any case, product features are also likely to play a significant role in determining perceived quality. Accordingly, we feel that it will be more useful to measure perceived quality separately and study the gap, if any, between the real and perceived quality.

Thirdly, we focus our attention on the basic function of data, which is to provide useful information. Thus, the quality attributes are classified in terms of their effect on the nature and amount of information. This gives us eight classes of data, namely wrong, noisy, irrelevant, inadequate, hard, redundant, right and rich data. The definitions of these classes are given in Table 2. It may be noted from the definition of 'redundant data' that we are interested in the quality of a database not in isolation but as one of many in a process control system. Table 3

Table 2. Definitions of data quality categories

Category	Definition
Wrong data	Data values or the estimates obtained using the data values differ systematically from their true values.
Noisy data	The random error associated with the data values is more than the permissible maximum
Irrelevant data	Data that can not be used for the purpose at hand
Inadequate data	Data that is potentially useful but do not meet all the requirements for its intended use.
Hard data	Data that is difficult to comprehend/analyse/interpret.
Redundant data	Database (or a part of it) having multiple sources (modes) of availability (generation).
Right data	Data that is free from the deficiencies mentioned above but for which a better alternative exists.
Rich data	Data that provides better estimates and/or has wider applicability than right data.

shows the correspondence between our eight quality classes and 11 of the 15 dimensions of Wang & Strong. It is noticed that there is no match for our redundant class. This is because Wang & Strong eliminated the dimension containing the attribute of data redundancy. We have retained this class because of its particular importance in process control. Redundant data not only add to cost and affect believability in the case of discrepancy between two sources of measurement, they are also a potential source of misapplication of data. Very often, the data that are favourable to the user may be used in the face of conflicting evidence gleaned from other sources.

Finally, using subjective judgement, these eight classes are suitably placed in the information dimension to obtain the data quality scale as shown in Table 4. There can, however, be situations where a different ordering of the first six classes in the proposed scale may be more appropriate. In this connection, it is interesting to note that in a survey (Gendron & D'Onofrio, 2001) involving people in the health care industry, all 15 quality dimensions of Wang & Strong were rated to be almost equally important, with the highest rating given to the dimension 'accuracy'.

Data Collection Process

The generic steps involved in a data collection process are well known (Figure 1). The first step, obviously, is to identify the target population. Here, we shall define our target population as the totality of the characteristics (variables) of interest of a set of individuals (which can be persons, months, machines, parts etc.) and the environment surrounding the individuals. The sampling step is characterized both by the sampling design and the physical method of collecting the samples.

It may also be noted that the data product under consideration here is a

Table 3. Correspondence between the proposed data quality classes and 11 data quality dimensions of Wang & Strong

Wang & Strong		
Dimension	Definition	This work
Accuracy	The extent to which data is correct and reliable	Wrong data, Noisy data
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial	Wrong data
Relevancy	The extent to which data is applicable and helpful for the task at hand	Irrelevant data
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand	Irrelevant data
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand	Inadequate data, Hard data
Appropriate amount of data	The extent to which the volume of data is appropriate for the task at hand	Inadequate data
Value added	The extent to which data is beneficial and provides advantages from its use	Right data, Rich data
Interpretability	The extent to which data is in appropriate languages, symbols and units, and the definitions are clear	Hard data
Ease of understanding	The extent to which data is easily comprehended	Hard data
Representational consistency	The extent to which data is presented in the same format	Hard data
Concise representation	The extent to which data is compactly represented	Hard data

database and not just the data values. Further, since the scope of this article is limited to zeroth order data, the process described here (Figure 1) does not extend up to the analysis phase.

Causes of Poor Data Quality

Table 5 contains a list of 50 important causes related to the eight data quality classes defined above. This list is meant to be a guideline. It is neither exhaustive nor is the classification universal. The details of a few of these causes are discussed below.

Cooked Data

Data that are deliberately falsified to gain some advantage are called cooked data. Such data are found in almost every organization. Fear, lack of commitment and greed are the main causes for falsification of data. In the context of survey data, Loebl (1990) writes that 'while bias or misrepresentation are inevitably

Table 4. Data quality scale

Category	Example	Impact	Rank*
Wrong data	Flinching at the specification limits	Misleading information	1
Noisy data	Poor repeatability	Potentially misleading information	2
Irrelevant data	Old data	Useless information	3
Inadequate data	Poor least count of measurement	Useful but partial information	4
Hard data	Missing values	Information difficult to extract	5
Redundant data	Multiple copies of the database	Information overload and adds to cost	6
Right data	Measurement using go-no go gauges	Useful information that barely meets requirements	7
Rich data	Measurement of the basic function of a process	Useful information that meets requirements well	8

*Higher the better.

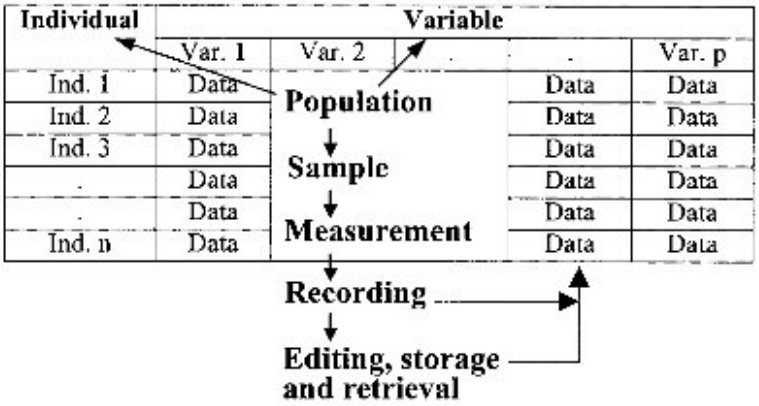


Figure 1. Data collection process

possible, they are likely to cause massive distortion, only when respondents consciously report misleading data for the purpose of gaining some kind of advantage'.

Controlled Noise during Experimentation

It may be desirable to control noise during experimentation if the sole aim is to establish the effect of a factor. However, in experimentation for the purpose of finding the best operating condition, controlling noise that is expected during operation may lead to the wrong conclusion. The best condition found from experimental data may not hold good in the operating condition if the controlled

Table 5. Important causes affecting data quality

Data quality	Causes
Wrong	(1) Nonrandom sample (2) Measurement bias (3) Cooked data (4) Duplicate data referring to the same individual in the same data base (5) Recording mistake (6) Wrong sample identity
Noisy	(7) Unstratified sample (8) Poor repeatability (9) Poor reproducibility
Irrelevant	(10) Irrelevant variable (11) Different background condition (12) Controlled noise during experimentation (13) Old data
Inadequate	(14) Aggregated individual (15) Aggregated variable (16) Yield as a measure for process improvement (17) Important variables left out (18) Improper data structure for the type of analysis to be made (19) Small sample (20) Unstratified sample (21) Fixed percent sample (22) One factor-at-a-time experiment (23) Poor least count of measurement (24) Long measurement delay (25) Coarse rounding (26) A part of the data lost during storage (27) Failure to locate all the files
Hard	(28) Correlation among dependent (independent) variables (29) Absence of unit (30) Confusing variable name (31) Very large sample (32) Censored sample (33) Truncated sample (34) Missing data (35) Variable repeatability (36) Subjectively coded data (37) Continuous charts (38) Images (39) Texts (40) Unknown measurement error (41) Descriptive unformatted data (42) Too many decimal places (43) Poor legibility
Redundant	(44) Multiple copies (45) Multiple acquisition
Right	(46) Sampling from heap (47) Full factorial experiment involving many factors (48) Go-No go measurements
Rich	(49) Measurement of the ideal function of a product/process (50) Bulk sampling while the material is on the move

noise factors interact with the control factors, or even the best may not be good enough if the noise factors have great impact on the response.

Aggregated Individual/Variable

These two causes are derived from the 'atomicity principle' of Davidson (1996), which says 'You can not analyze below the data level that you observe. (You can not analyze atoms if all you measure are molecules.)' For example, it is obviously not possible to get component level information from product level data. Similarly, we cannot find the most important cause of rejection if all we have are the data on 'total daily rejection', which is an aggregate of the variables defining the various causes of rejection.

Every industry routinely collects a lot of data for product and process monitoring. These databases are also expected to play an important role in improving the efficiency of the problem solving process. Thus, the question naturally arises regarding the optimal design of these databases whereby the data should not only serve the immediate purpose of process control but should also help in future in problem solving. One of the design questions here is to decide on the atomicity of the data. This however is an issue, on which, to the best of our knowledge, even guidelines are not available in the literature.

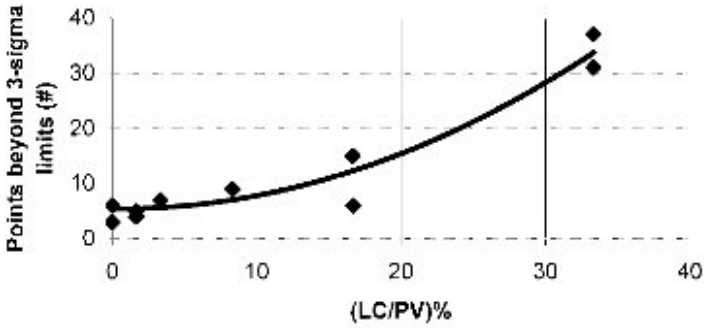


Figure 2. Effect of least count of measurement on performance of x-bar chart. The number of points falling outside the control limits (known sigma case) is plotted against least count (LC) as a percentage of process variation (PV).

Note that the situation here is different from the one where data are collected assuming some day someone will use it. What happens then is that no one actually uses it; the data become old and consequently loses its relevance to the existing process.

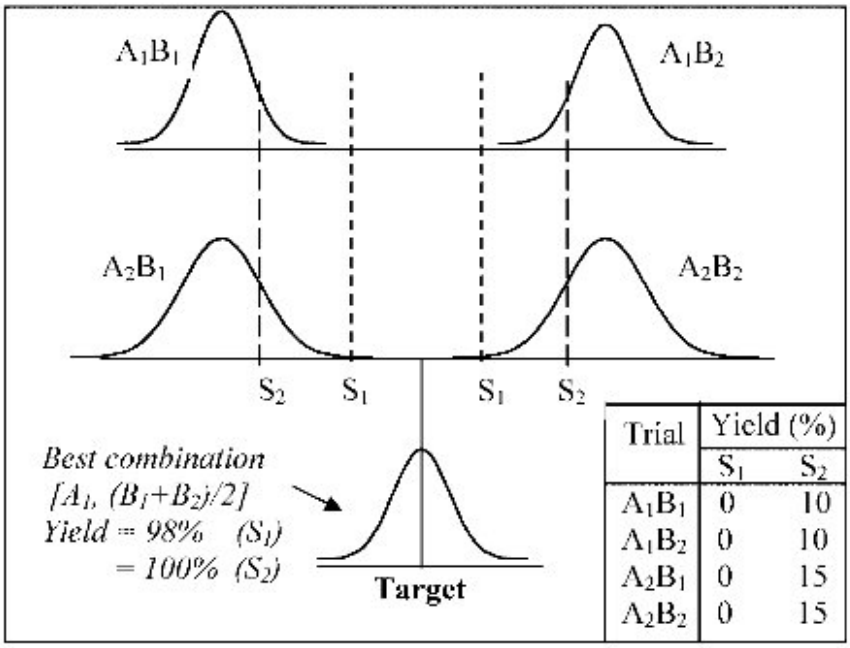
Poor Least Count of Measurement

Poor least count of measurement affects both the estimation of process variation and process monitoring adversely (Mandal, 1998). Here we shall record only the effect of least count on process monitoring using x-bar chart.

In a simulation experiment, 2000 random samples of size seven were generated from a normal distribution having mean of 20 and standard deviation of one. So the Process Variation (PV) of the hypothetical process is six. The simulated observations were rounded in six different ways to yield measurements having Least Count (LC) of 0.00001, 0.1, 0.2, 0.5, 1.0 and 2.0 respectively. Thus we had six sets of observations from an in-control process corresponding to LC/PV of 1/6000, 1/60, 1/30, 1/12, 1/6 and 1/3. The rounded observations were analysed using x-bar chart, the results of which are summarized in Figure 2. It is apparent from this figure that when the least count is less than or equal to 1/12th of the process variation the average number of out of control signals observed is not very different from that expected (about 5 in 2000) under ideal conditions. However, as the least count increases beyond 1/12th of the process variation, the number of out of control signals increases drastically. It is thus concluded that for effective process monitoring the least count of measurement, as recommended (Chrysler *et al.*, 1995), should be about 10% of process variation.

Recording Yield as a Measure of Process Performance

Consider a 2^2 experiment consisting of four trials denoted by A_1B_1, A_1B_2, A_2B_1 and A_2B_2, where A_iB_j represents the combination corresponding to the ith level of factor A and jth level of factor B. The objective of the experiment is to find the level combination of the two factors A and B that will maximize yield



Trial	Yield (%)	
	S_1	S_2
A_1B_1	0	10
A_1B_2	0	10
A_2B_1	0	15
A_2B_2	0	15

Figure 3. Results of a 2×2 experiment (involving factors A and B) corresponding to two specifications S1 and S2. The hypothetical results illustrate that process yield, which is a function of both process mean and variation, is not a good response variable for an experiment.

(percentage of output within specification limits). However, recording yield as a response for the experiment (as is frequently done in practice) may lead to an erroneous conclusion. This is explained below.

Let the result of the experiment be as shown in Figure 3, where S_1 and S_2 are two different sets of specifications. It is seen that the level A_1 giving lower output variation is better than A_2 and the factor B has strong impact on the mean. Assuming the effect of B on the mean is linear within the range of experimentation, the best combination is clearly $[A_1, (B_1+B_2)/2]$ and the corresponding expected yields with respect to S_1 and S_2 are 100% and 98% respectively. However, if we analyse the yield data with respect to S_1 our conclusion will be that none of the two factors have any impact on yield. With respect to S_2, we may even erroneously conclude that level A_2 is better than A_1. Moreover, since the resultant yield is very low for all the trials, we may unnecessarily plan further experimentation excluding the present zone.

Correlation Among Response (Predictor) Variables

In general, a process control database consists of a set of predictor variables and a set of response variables. High correlation among the response variables makes the data hard because it becomes difficult to have a good statistical procedure for monitoring the process, particularly the process variation (Montgomery, 1996). On the other hand, it is well known that in the presence of correlation

among the predictor variables the results of regression analysis need to be interpreted carefully (Box, 1966), which classifies the data as hard. Finally, if the correlation between the sets of predictor and response variables is low then the data may be classified as inadequate.

Confusing Variable Name

It is a common practice to measure the diameter of a round part at several locations by rotating it and recording the difference between the maximum and the minimum as the ovality. This, however, is not a sound practice since we have found that on many occasions the variation among such observations is purely random in nature.

To take another example, consider a lamination process in which molten plastic is pasted on paper. One of the defects, for which substantial amount of paper was getting rejected, was due to separation of layers of the paper through entrapment of an unknown gas. This defect was being referred to as 'air bubble'. Such a name, in the absence of definite knowledge that the entrapped gas is air and not something else, may be misleading. Also note that the name mystifies the issue since it seemed almost impossible for air to enter between two layers of the paper.

Failure to Measure the Ideal Function of a Product/Process

Of late, Taguchi is advocating measurement of the ideal function of a product or process and making the function robust or insensitive to the noise factors as a strategy for product/process improvement. Many successful applications of this strategy have also been published (Taguchi *et al.*, 2000). Here we shall explain with the help of a case example why data that are related to the ideal function have been classified as rich.

Various types of surface and subsurface defects, like laps and scratches, are found in hot rolled wire rods. It was desired to conduct an experiment for reducing the extent of surface and subsurface defects, which are detected by visual inspection and a hot/cold upset test. The visual inspection gives count data and the results of upset testing are somewhat subjective. Although both of these measures could be used as responses for the experiment, it would have been necessary to inspect and test many coils and thereby obtain right data. However, measuring the diameter of the rod, which is related to the basic function of rolling, was thought to be a better alternative. However, the management was not convinced about the suitability of measuring diameter for reducing surface defects, particularly because diameter was under control. Accordingly, apart from measuring the diameter of the experimental coils, these were also subjected to a hot upset test. Figure 4 shows that variance of the diameter and hot upset rating are indeed correlated and both indicate that best results are obtained with a soaking temperature of 1180°C.

Linking Data Quality to Data Collection Process

The cause(s) of a data quality problem can be traced back to some deficiency in the data collection process (Figure 5). This linkage is illustrated in Table 6 for

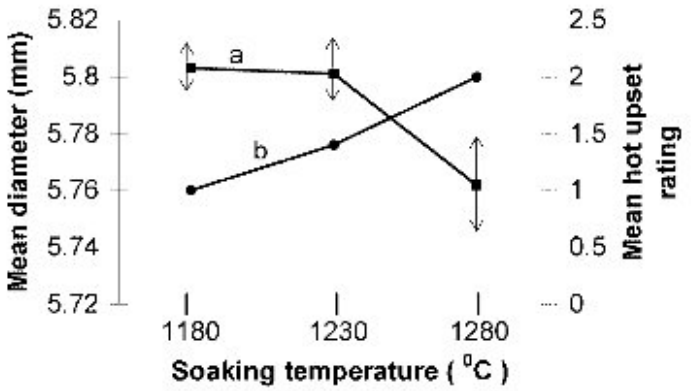


Figure 4. Effect of soaking temperature on diameter and hot upset rating of wire rods rolled under identical rolling conditions. (a) Diameter. (b) Hot upset rating. The lower the hot upset rating, the better the quality of a wire rod. The vertical arrows on the diameter curve indicate the 99% confidence intervals for the means.

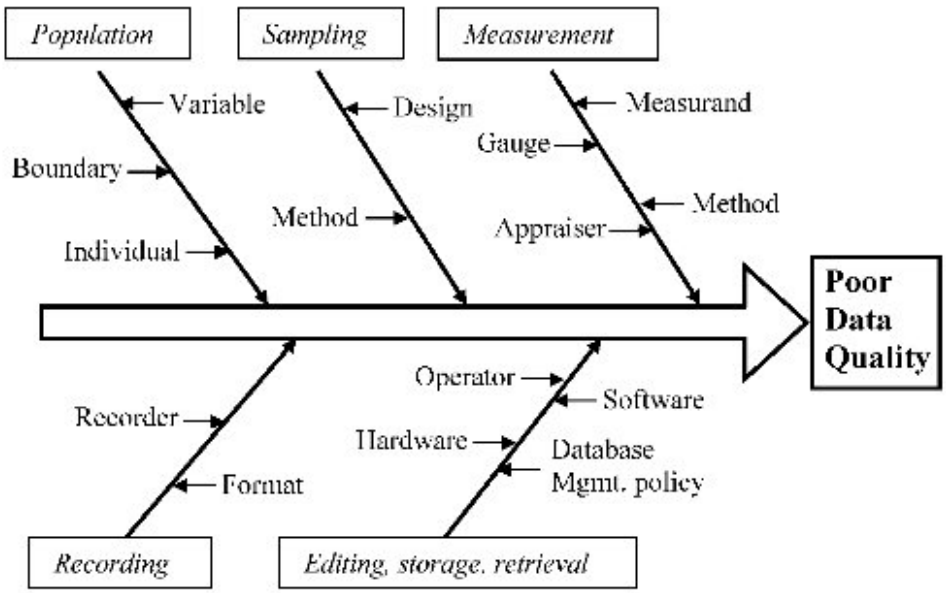


Figure 5. Cause and effect diagram showing the main sub-causes of poor data quality.

the 50 causes listed in Table 5. It may be noted that the redundant class is absent in this table since the causes for redundant data (as defined in Table 2) are related to management of databases and not collection of raw data.

Data Collection Planning

Harry (1997) has brought out the importance of data collection planning beautifully through his five step planning process. This planning process is

Table 6. Linking data quality to data collection process

		Data quality class						
Data collection process element		Wrong	Noisy	Irrelevant	Inadequate	Hard	Right	Rich
Target population	Individual				14			
	Variable			10	15,16,17	28,30,36		49
	Boundary			11,12				
Sampling	Procedure	1,6	7		18,20,22	28,32, 33,34	46,47	50
	Size				19,21	31		
Measurement	Gauge	2	8		23	37,38, 40	48	
	Appraiser	2	9			40		
	Others	2			24	34,35		
Recording	Format					29,39, 41,42,43		
	Recorder	3,5			25	43		
Editing, Storage, Retrieval		4		13	26,27	34		

Note: The numbers within the cells refer to the cause number of Table 5.

explained in Figure 6. The main thrust of the process is that execution is carried out in the reverse direction of planning, whereby the power of various statistical tools is better realized to obtain the desired solution. It is indeed satisfying to note that we can avoid many data quality problems by implementing the planning process. For example, the first step is a safeguard against irrelevant data. The second and third steps prevent occurrences of inadequate data and the fourth step facilitates generation of right data.

However, it must be noted that here we are concerned not only with the data requirement for a given situation but also with future use of a database. Thus, the first planning question is not just 'what do you want to know?' but also 'what may one want to know in future?' This makes the planning process difficult. Further, even when the purpose is known, it should be clear from our previous discussion of data quality that the simple five-step process as above might not provide rich data. Therefore, we recommend that at the end of the five steps, before we execute the process we should ask, 'Will the data be wrong?' 'Will the data be noisy?' and continue up to 'Will the data be rich?'

Concluding Remarks

In the introduction we mentioned that the main objective of this work is to discuss the issue of 'data quality' from an angle that will be suitable for courses on SPC, Six Sigma and the like. There is certainly a need to do better than merely issuing the famous warnings such as 'whenever you see data, doubt it' and 'whenever you see a measuring instrument, doubt it'. Hopefully, this article will be of some help to the trainers in this respect. However, we realize that

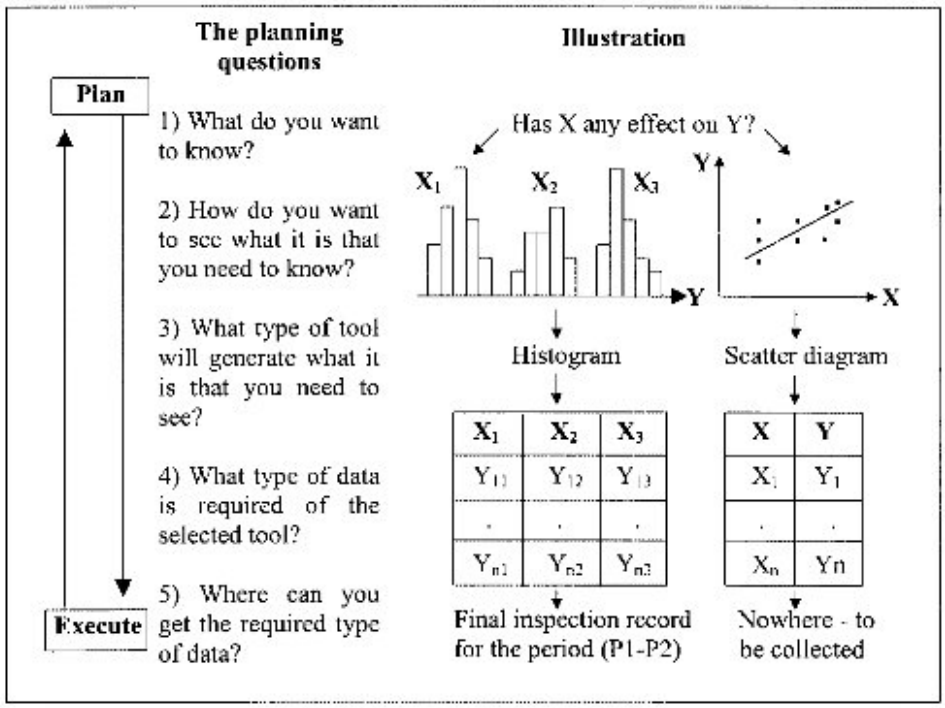


Figure 6. Data collection planning process.

although the topic may be introduced at the beginning of a course, it is preferable to have a detailed discussion near the end so that the participants appreciate the importance of data quality better.

Acknowledgement

The author would like to thank an anonymous referee for several suggestions that significantly improved the presentation of this paper.

References

Abate, M. L., Digert, K. V. & Allen, W. (1998) A hierarchical approach to improving data quality, *Data Quality*, 4(1), Available at: www.dataquality.com/998abate.htm

Agmon, N. & Ahituv, N. (1987) Assessing data reliability in an information system, *Journal of Management Information Systems*, 4, pp. 34–44.

Box, G. E. P. (1966) Use and abuse of regression, *Technometrics*, 8, pp. 625–629.

Chrysler Corporation, Ford Motor Company, & General Motors Corporation (1995) *Measurement Systems Analysis: Reference Manual*, 2nd edn, p. 20.

Davidson, F. (1998) *Principles of Statistical Data Handling* (California: Sage Publications).

Dubois, L. (2002) *Archiving Enterprise Data Quality*. Available at http://www.tdan.com/i006fe07.htm

Firth. C. P. (1996) Data quality in practice: experience from the frontline, *The 1996 Conference on Information Quality*, 25–26 October, Massachusetts Institute of Technology (see also http://sunflower: singnet.com.sg/~cfirth/dataquality/).

Gendron, M. S. & D'Onofrio, M. J. (2001) Data quality in the health care industry, *Data Quality*, 7(1). Available at: www.dataquality.com/90160.htm

Harry, M. J. (1997) *The Vision of Six Sigma: A Roadmap for Breakthrough* (AZ, USA: Tri Star Publishing).

International Monetary Fund (2001) *The IMF Generic Data Quality Assessment Framework*, Available at http://dsbb.imf.org/vgn/images/pdfs/dqrs-Genframework.pdf

King, W. & Epstein, B. (1983) Assessing information value: an empirical study, *Decision Sciences*, 14, pp. 34–35.

Loebl, A. S. (1990) An organizational and historical perspective of a decade of data collection, in: G. E. Liepins & V. R. R. Uppuluri (Eds) *Data Quality Control: Theory and Pragmatics* (New York: Marcel Dekker).

Mandal, P. (1997) *Whose Quality Is It – Producer's or Customers'? Juran's or Crosby's or . . .?*, Quality Quest, Technical Report No. 1 (Baroda: Indian Statistical Institute).

Mandal, P. (1998) Measurement systems analysis – some ambiguities, clarifications and proposals, In: *Proceedings of 7th National Convention of NIQR*, 9–10 January, Bangalore, India.

Montgomery, D. C. (1996) *Introduction to Statistical Quality Control*, 3rd edn (New York: Wiley), pp. 369–374.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. (2002) Data quality assessment, *Communications of the ACM*, 45(4), pp. 211–218.

Pricewaterhousecoopers (2001) *Global Data Management Survey*. Available at http://www.pwcglobal.com/Extweb/ncsurvres.nsf/

Orr, K. (1998) Data quality and systems theory, *Communications of the ACM*, 41(2), pp. 66–71.

Strong, D., Lee, Y. & Wang, R. (1997) Data quality in context, *Communications of the ACM*, May, pp. 103–110.

Taguchi, G. & Wu, Y. I. (1979) *An Introduction to Off-line Quality Control* (Nagaya 450, Japan: Central Japan Quality Control Association).

Taguchi, G., Chowdhury, S. & Taguchi, S. (2000) *Robust Engineering* (New York: McGraw-Hill).

US Environment Protection Agency (2000) *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9* (Washington, DC: Office of Environmental Information, see also http://www.epa.gov/quality1/dqa.html).

Wang, R. & Strong, D. (1996) Beyond accuracy: what data quality means to data consumers, *Journal of Management Systems*, 12, pp. 5–34.