CHAPTER

14 Functions of Measurement: Collection, Calculation, Comparison

"It is by the art of statistics that law in the social sphere can be ascertained and codified, and certain aspects of the character of God thereby revealed. The study of statistics is thus a religious service."
– Florence Nightingale[1] (1820 – 1910)

Purpose

This chapter describes the high-level process of taking the in-line measurements in terms of a set of functions common to many of the Data Quality Assessment Framework measurement types. Measurement types (generic patterns for taking specific measurements) begin with the collection of raw measurement data. They process that data through calculations that make the measurements comprehensible and through comparisons that make the results meaningful. These calculations and comparisons are intended to detect changes in data patterns that may indicate quality problems. The chapter presents relatively dense, technically oriented material. It should be read in conjunction with Chapters 15 and 16.

DQAF calculations and comparisons are based on statistical techniques. In order to understand how the DQAF measurement types work and how to apply them to particular data and to interpret results, it is important to know what these techniques are and the concepts behind them. There are software programs that can execute the calculations themselves. The chapter presents high-level definitions of statistical concepts that are necessary for the general context of measurement and to the DQAF's automation of data quality thresholds.

Functions in Measurement: Collect, Calculate, Compare

The DQAF describes the measurement of data quality in generic terms at the level of the measurement type. A measurement type is a category within a dimension of data quality that allows for a repeatable pattern of measurement to be executed against any data that fits the criteria required by the type, regardless of specific data content. Measurement types occupy a middle ground between abstract dimensions of data quality and measurements of specific data. Validity is a dimension of data quality. The percentage of records containing invalid diagnosis codes is a metric related to specific

[1] Attributed by F. N. David.

data. An example of a measurement type is the capability of measuring the validity of any designated data field based on a defined domain of values and calculating the percentage of records where values are invalid. (These relationships are illustrated in Figure 4.1 in Chapter 4.)

Automating DQAF measurement requires building functions to collect raw measurement data, execute appropriate calculations against that data, and compare against thresholds or past history to make the data meaningful so that data's quality can be assessed. The ways of collecting raw measurement data can be understood largely in terms of what features of the data need to be counted (file sizes, process duration, record counts, group-bys, etc.). Making measurements meaningful can be understood through the kinds of calculations (percentages, averages, etc.) that put results from large datasets into comprehensible terms and set up a means of making useful comparisons. The comparisons utilize statistical methods to understand results in relation to an established threshold or to determine how similar any instance of a measurement is to previous instances of the same measurement (see Figure 14.1).

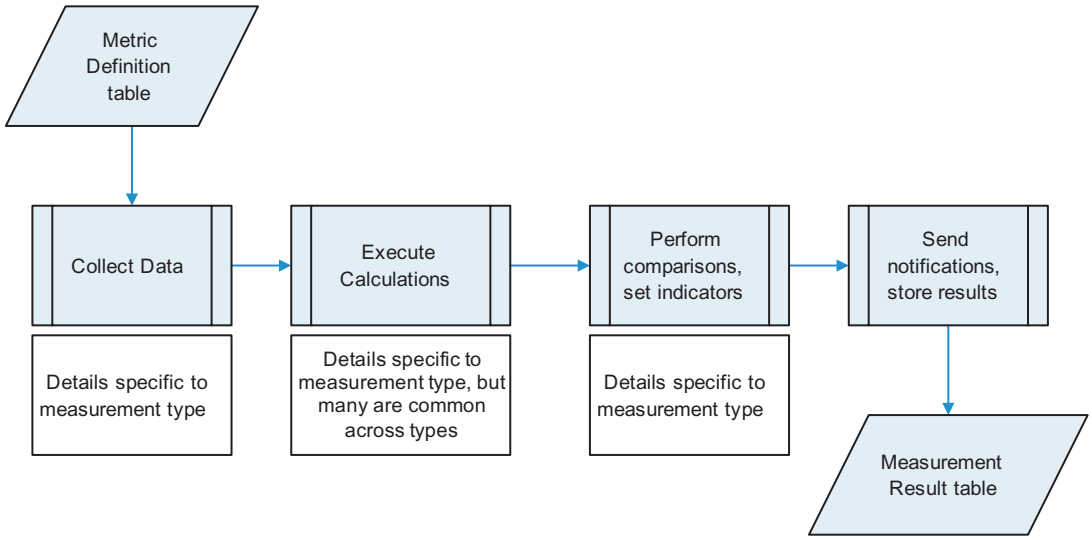

FIGURE 14.1 Functions in Measurement: Collect, Calculate, Compare

Definitions of specific metrics are stored in metric definition tables. These drive the collection of measurement data. Once this data is collected, calculations are executed against it. While each measurement type requires a specific set of calculations, there are many commonalities between them. For example, many of the consistency measurement types require the calculation of a percentage distribution of values. The results of the initial calculations provide input to comparisons. Comparisons are made either against quality thresholds defined by subject matter experts or based on the results of past measurements of the same dataset. The comparison process assigns indicators against these thresholds. If the measurements produce unexpected results, notifications will be sent and response protocols initiated. Results are stored in the measurement result tables, where they can be used for trend analysis and reporting.

Collecting Raw Measurement Data

Even in a complex dataset, there is a limited number of ways to collect basic input for quality measurement. These include:

- Capture the start and stop times of a process.
- Count total records in a dataset.
- Count records grouped by values in a field.
- Count records grouped by values across fields.
- Count and group records based on filters or qualifying criteria.

Data are abstract representations of real-world objects, concepts, or events, often referred to as entities. An entity may have numerous references (records) in a dataset. A distinct count of entities represented in a dataset determines how many individual entities the data contains references to, regardless of how many references there are. For example, a library may have 10 books (records) by Mark Twain, but Mark Twain is only one entity (author). Or it may have 10 copies of *Tom Sawyer*, but *Tom Sawyer* is only one entity (one novel, printed in numerous editions). Basic input related to distinct entities includes:

- Count the distinct number of a represented entity within a dataset.
- Count the distinct number of a represented entity within a dataset based on qualifying criteria

Date fields can help confirm expectations related to data content completeness and consistency and to assess the reasonableness of other data. However, the specific values of dates are not directly comparable in transactional data. Meaningful comparisons related to date data often require data be aggregated at the weekly, monthly, quarterly, or annual level. Basic aggregations based on date fields can make comparisons between datasets more meaningful.

- Count total records per aggregated date.
- Count records associated with two date fields.

Amount fields contain data in numeric form representing the number of objects associated with a record (such as number of products ordered) or representing currency amounts (such as net sales amount). For such fields, even the collection of "raw" data requires basic calculations that can be used as input for quality measures.

- Calculate the total value in an amount field within a dataset.
- Calculate the total value in an amount field based on filters or qualifying criteria.
- Calculate the value in an amount field grouped by values in another field or fields.
- Calculate the value in an amount field grouped by values in another field or fields based on filters or qualifying criteria.

Calculating Measurement Data

Raw counts, calculations, and even aggregations are of limited use by themselves. They can be made meaningful through calculations that enable people to see the relation of data within the set. Some

basic calculations that make raw counts more meaningful include:

- The percentage of records associated with each distinct value in a field (percentage distribution of values).
- The ratio of total records to represented entity (the average number of records per represented entity).
- The percentage of total records associated with each distinct represented entity.
- The percentage of total amount associated with each distinct value in a field.
- The average amount associated with each distinct value in a field.
- The ratio of two related amount fields.
- The percentage of total records per aggregated date.
- The summed amount per aggregated date.
- The percentage of total amount per aggregated date.

Percentages and averages enable people to understand raw measurement data better, especially if raw measurements involve very large numbers or large sets of numbers. In some data warehouses, tables may contain billions of rows, and any given update may contain millions of rows. Some code sets, such as ICD diagnosis and procedure codes, include thousands of distinct values. Percentages and averages situate the data within a comprehensible context and enable assessment of quality. These calculations are summarized in Table 14.1.

Table 14.1 Summary of Calculation Types for Processing Measurement Data

Description	Equation	Example
Percentage of records associated with each distinct value in a field	(Record count of distinct value/Total record count)*100	Percentage of records associated with each Place of Service Code
Ratio of total records to represented entity (average number of records per represented entity)	Total record count/Distinct number of represented entities	Average number of claims per member
Percentage of total records associated with a represented entity	(Record count of distinct entity/Total record count)*100	Percentage of claims per provider
Percentage of total amount associated with distinct values in a field	(Amount per distinct value/ Total amount)*100	Percentage of net paid dollars by Place of Service Code
Average amount associated with distinct values in a field	Amount per distinct value/ Record count for distinct value	Average net paid dollars for each Place of Service Code
Ratio of two related amount fields	Total amount field #1/Total amount field # 2	Total billed amount/Total net paid amount
Records per aggregated date	Records/Aggregated date	Claims per month-year
Percentage of total records per aggregated date	(Records per aggregated date/Total records)*100	Percentage of claims per month-year
Amount per aggregated date	Amount/Aggregated date	Total Net Paid per month-year
Percentage of total amount per aggregated date	(Amount per aggregated date/ Total amount)*100	Percentage Total Net Paid per month-year

Comparing Measurements to Past History

Understanding very large numbers or large sets of numbers is even more complex if measurements are taken over time. A primary goal of the DQAF is to describe in-line data quality measurement that enables analysis of trends in data quality. DQAF measurement types include not only basic calculations to make raw measurements meaningful within the context of one instance of measurement, but also additional comparisons to past measurements. These comparisons include:

- Comparison to the historical mean percentage of row counts.
- Comparison to the historical mean total amount.
- Comparison to the historical percentage of total amount.
- Comparison to the historical average amount.
- Comparison to the historical median duration.

For a stable process that is producing data that meets data consumers' expectations, comparisons to historical data can confirm that facets of data content are consistent. Historical data measurement data can provide the basis for automation of an initial level for data quality thresholds. To understand how such a process can work, we will discuss some basic concepts in statistics.

Statistics

Statistics is "the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample" (*NOAD*). Statistics use mathematical procedures to describe data—that is, to create data about data so that large datasets can be described in more comprehensible terms and so that comparisons can be made between datasets.

The general practice of measuring involves assigning quantities (numeric values) to qualities of the thing being measured in order to understand those qualities (Kircher, 1962, p. 72). Measuring data quality is the process of assigning quantities to qualities of data that we want to understand. Statistics provides critical tools for data quality, and statistical methods allow the creation of comprehensible measurements of data quality.

Many books on statistics begin by acknowledging that most people find statistics difficult (Rumsey, 2011; Salkind, 2011). Most studies also acknowledge that we live in a world described through statistics—the Dow Jones average, the gross domestic product, the unemployment rate, the Scholastic Aptitude Test score, the batting average, and so on.

Measuring data quality does not require mastery of statistics. But the use of statistical tools to measure data quality depends on applying the appropriate tool to the situation. To take basic measurements and interpret them, it is important to know the basic concepts. As with all data quality measurement, effective use of statistical tools depends on understanding expectations for the data.

Statistics includes a large set of methods for understanding data. I will focus on a small set of descriptive statistics that has a direct bearing on basic measures of quality: measures of central tendency (mean, median, mode) and measures of variability around a central tendency (range, variance,

standard deviation).[2] Measures of central tendency (also called measures of location) enable you to understand one value that can be understood as "typical" for the set. They describe "the tendency of quantitative data to cluster around some random value. The position of the central variable is usually determined by one of the measures of location, such as mean, median, or mode. The closeness with which the values cluster around the central value is measured by one of the measures of dispersion, such as the mean deviation or standard deviation" (Dodge, 2006, p. 60). While measures of central tendency quantify how data cluster around a value, measures of variability (also called measures of dispersion) describe how "spread out" the data in a set are from such a value.

Measures of Central Tendency

Common measures of central tendency include the mean, median, and mode. At its simplest, the *mean* refers to what most people understand as the mathematical average.[3] The mean is calculated by adding a set of numbers and dividing by the number of numbers in the set. The *median* of a dataset is the value that occupies the middle position of a dataset. When the data points are ranked in ascending order, an equal number of values are above and below the median. The median value does not have to exist as a value in the dataset. (If a dataset contains an even number of values, the average of the two middle numbers is the median.) The *mode* is the value that occurs most frequently in a dataset.

Each of these measures has advantages and disadvantages as a measure of central tendency. The mean is useful because many people understand it. But it can be influenced significantly by extreme values (lows, highs, outliers). It may also be misleading for nonsymmetrical data (data that has extreme values at one end of its set). This problem can be mitigated by calculating a *trimmed mean*, which is based on discarding extreme values in a distribution. The median is less influenced by extremes and can therefore be a better way to understand the central tendency of nonsymmetrical data (Boslaugh & Watters, 2008).

Measures of Variability

Measures of variability provide an understanding of how spread out data is. These measures include the range, variance, and standard deviation. The *range* is the difference between the largest and the smallest number in the dataset. Variance and standard deviation need to be understood in relation to each other and to a measure of central tendency, usually the mean. *Variance* is a measure of "the fluctuation of observations around the mean" (Mitra, 1993). Technically, *variance*, or *mean squared*, is the average of the squared deviations from the mean. A deviation is the difference from a standard value. The *standard deviation* is the square root of the variance.

While *range* describes the extreme values of the dataset, *variance* uses information associated with each observation. A larger value for the variance indicates a greater dispersion of data points (Mitra,

[2]The discussion on statistics draws from Boslaugh and Watters (2008), Mitra (1993), Ott and Schilling (1990), and Dodge (2006). I am grateful for feedback from Dick Janzig and Kent Rissman on the use of specific terms.

[3]Technically, the term *average* refers to the set of measures of central tendency (Salkind, 2011; Dodge, 2006). Practically, *average* and *mean* are synonyms. I will use the term *mean* in this discussion, except when I describe average amount as part of the individual measurements that can be taken on amount fields. I choose to use *average amount* to distinguish from *mean*, as I hope will be clear in the detailed definitions included later in the chapter.

1990). Statisticians look at the standard deviation as a measure of variance because it can be related to the distribution of data. Most of us are familiar with standard deviation because of its role in the bell curve. Roughly 68% of data fall within one standard deviation, 95% is within two standard deviations, and 99.7% is within three standard deviations of the mean on a bell curve (see Figure 14.2).

Understanding both the central tendency and the variation in a dataset is necessary to interpret data quality measurement results. Two datasets can have the same mean but have different ranges and different deviations from the mean, as is illustrated in Table 14.2. New values added to each set need to be assessed within the context of that set. Figures 14.3 and 14.4 illustrate the impact of the measures of dispersion on our understanding of the numbers.

Knowledge of the basic concepts of these measures of central tendency and variation helps in identifying outliers. Unfortunately, there is not an agreed-to definition of *outlier*. Conceptually, *outliers* are data points with values recognizably different from others in the dataset being analyzed. They may imply a different pattern, appear to come from a different population, or represent an error (Boslaugh & Watters, 2008).

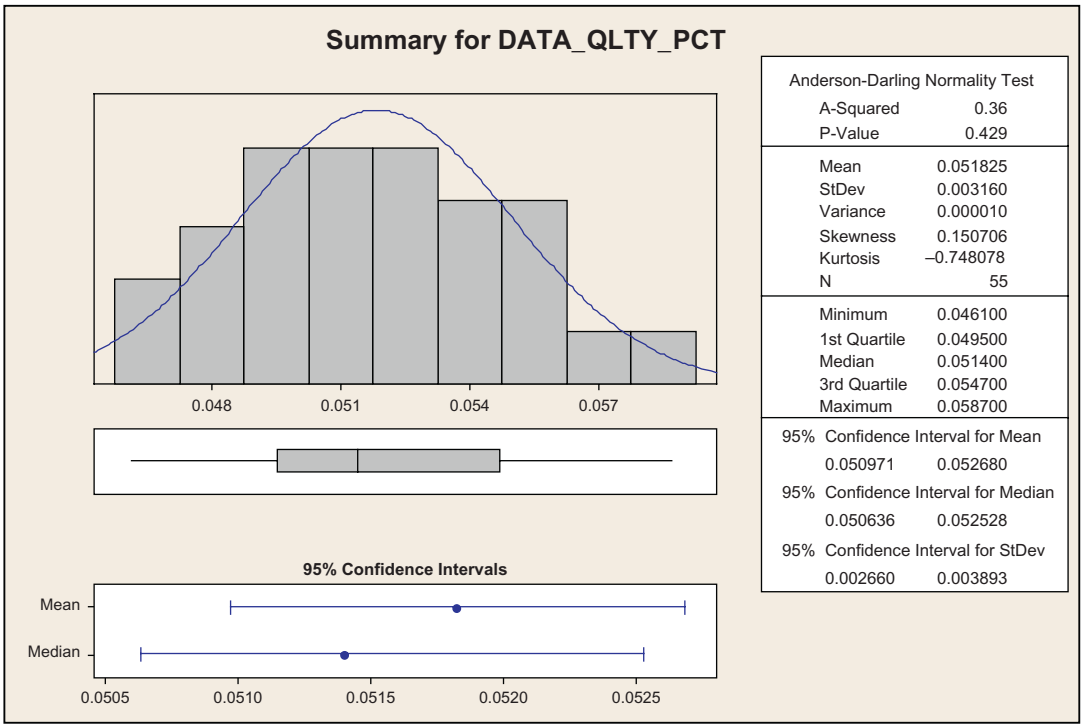

FIGURE 14.2 Summary Statistics for a Data Quality Metric

The figure, produced by the software package Minitab, includes summary statistics for a data quality measurement. The chart depicts a bell curve from the results. It is not completely symmetrical. I have defined only a few basics (the mean, standard deviation, and variance). Minitab generates a range of detail which you may want to take advantage of in analyzing your results.

Table 14.2 Value Sets with the Same Mean but Different Standard Deviations

Value Set 1	Value Set 2
12	20
12	23
14	23
14	25
15	25
15	25
16	27
44	31
50	32
74	35
Total = 266	Total = 266
Mean = 26.6	Mean = 26.6
Range = 62 (74–12)	Range = 15 (35–20)
Standard Deviation = 21.7	Standard Deviation = 4.8
Variance = 469	Variance = 21.8

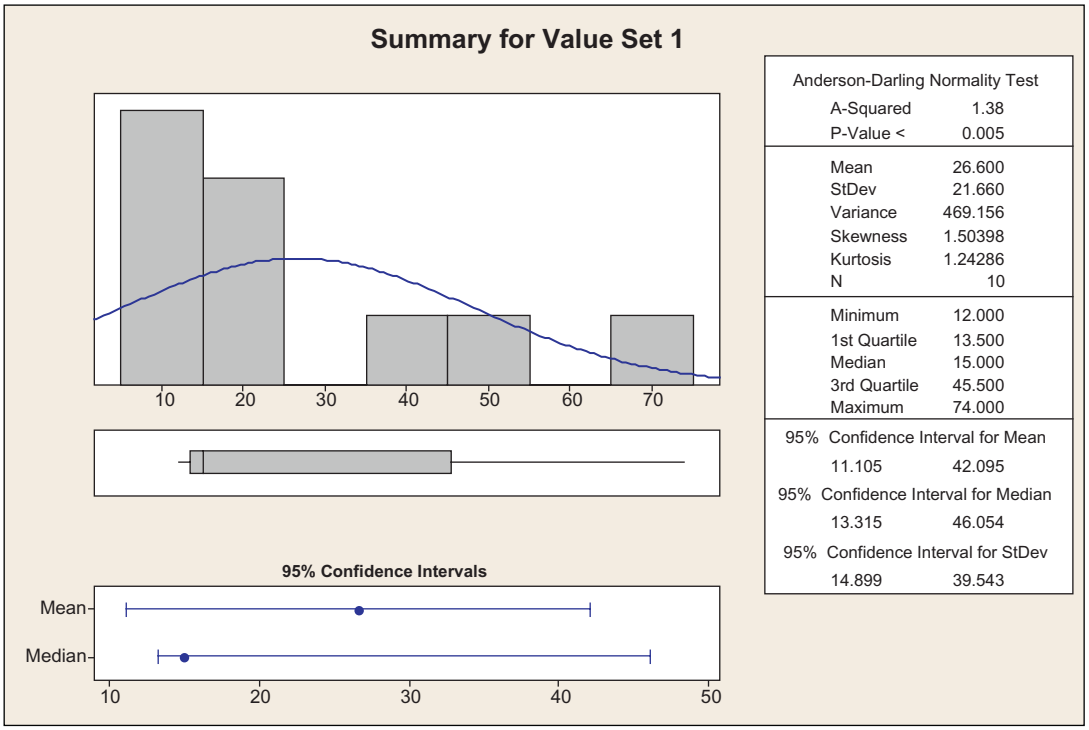

FIGURE 14.3 Summary for Value Set 1

The mean of the values in set 1 is 26.6. But the set has a range of 62 and a standard deviation of 21.7. So, while its measures of central tendency make it look similar to value set 2, its measures of dispersion are quite different.

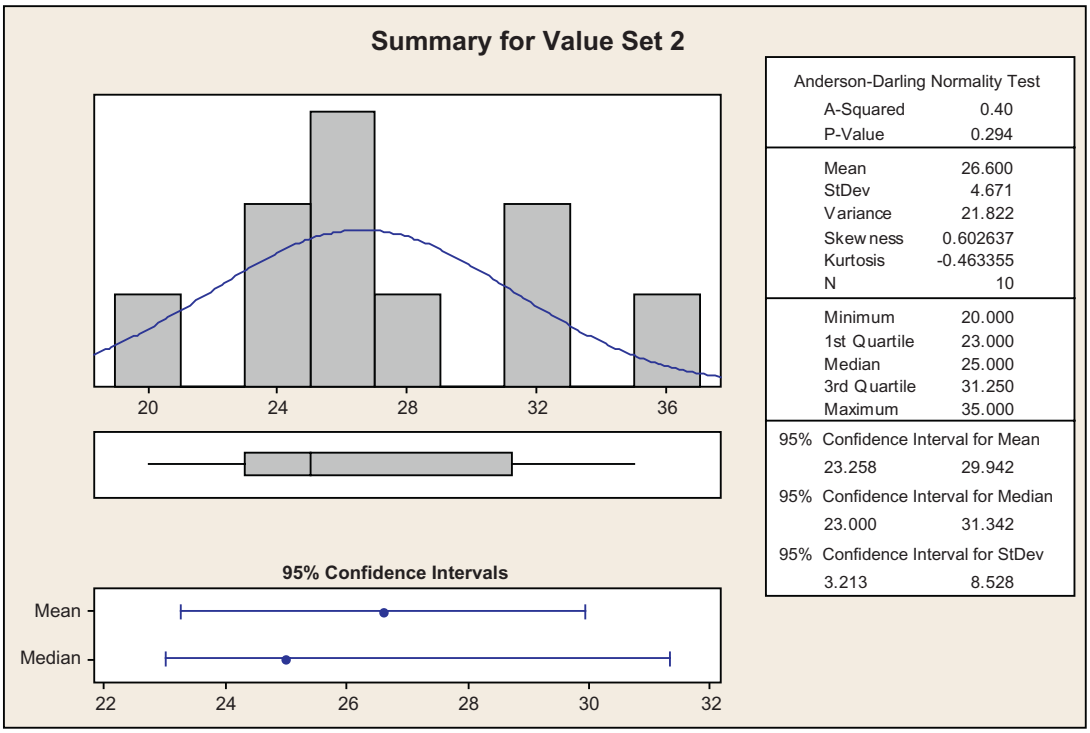

FIGURE 14.4 Summary for Value Set 2

The mean of the values in set 2 is the same as in set 1, 26.6. But set 2 has a much narrower range (15 as opposed to 62) and a smaller standard deviation (4.8 as opposed to 21.7) and variance (21.8 as opposed to 469). While measures of central tendency are similar between the two sets, measures of dispersion differ greatly.

The Control Chart: A Primary Tool for Statistical Process Control

Statistical process control (SPC), a method for measuring the consistency and helping ensure the predictability of manufacturing processes, was pioneered by Walter Shewhart in the first half of the twentieth century. A primary tool for SPC is the control chart, a time series plot or run chart that includes a representation of the mean of the measurements and of the upper and lower control limits (three standard deviations from the mean of the measurements). (See Figure 14.5 Sample Control Chart.) One of the biggest benefits of the control chart is its visual nature. From the chart, one can see the range of measurements (the difference between the maximum and minimum Y values) as well as the consistency of the measurements (how much up and down there is between the measurements and across the set) to gain an understanding of the stability of the overall process.

Shewhart used control charts to measure levels of product defects. From his analysis he discerned that two kinds of variation contributed to differences in the degree to which products

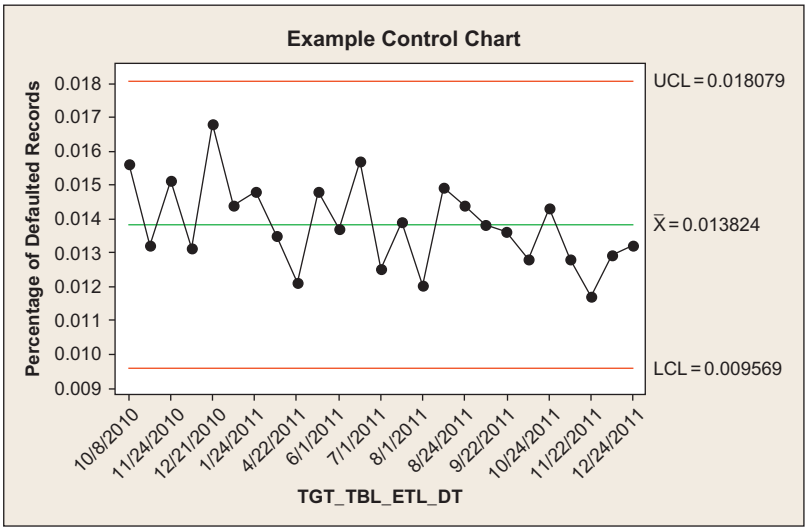

FIGURE 14.5 Sample Control Chart

The figure shows results from a process that is stable (in statistical control). All the measurement points are within the upper and lower control limits (UCL and LCL on the chart) around the mean.

conformed to specification: normal or common cause variation that is inherent in the process; and special cause or assignable variation that is the result of elements outside the process. Process measurement aims first to detect and remediate the root causes of special cause variation. Once a process is stable and predictable, efforts to improve it are directed at reducing the effects of common cause variation.

For a process that is in statistical control (special cause variation has been eliminated), 99.7% of all measurements fall within the upper and lower control limits (because 99.7% of all measurements should be within three standard deviations from the mean of the set of measurements). The question of whether a process is statistically under control is separate from the question of whether the end product of the process meets requirements. A process can be under control (exhibiting only common cause variation), but the quality of the end product still might not meet customer expectations.

The DQAF and Statistical Process Control

As discussed in Section Five, most approaches to data quality improvement start with a comparison between the production of data and the production of manufactured goods and recognize the value of treating data like a product. Much thinking about data quality is directly rooted in methods for process quality. Statistical process control methods have been successfully applied to the measurement of data quality both for initial analysis which identifies special causes and for ongoing measurement to

confirm whether a process remains in control.[4] The DQAF draws directly on this body of work, especially for its approach to automating measurement.

The third basic step in making quality measurements meaningful is to explicitly compare any new measurement to a standard for quality. The DQAF describes how to collect the raw data of measurement (record counts, amounts, etc.) and how to process this data to make it comprehensible and meaningful through the calculation of percentages, ratios, and averages. Most measurement types compare new measurements to the history of past measurements. These comparisons include:

- Comparison to the historical mean percentage of row counts.
- Comparison to the historical mean total amount.
- Comparison to the historical percentage of total amount.
- Comparison to the historical average amount.
- Comparison to the historical median duration.

Comparisons to historical data can be used to measure the consistency of data content for processes where content is expected to be consistent. For example, tests of reasonability based on the distribution of values can be automated to identify instances where changes in distribution fall outside of three standard deviations from the mean of past measurements. These comparisons can identify changes in data distribution that are statistically unusual (those that are outside of the 99.7% of all measurements). Not every measurement that falls outside of three standard deviations from the mean of past measurements represents a problem. Such measurements point to data that needs to be reviewed. Because thresholds based on three standard deviations from the mean of past measurements provide a test for potential problems, historical data can be used as the basis for automation of an initial level for data quality thresholds. Keep in mind that there is risk involved in using historical data as the standard. For this data to provide an effective standard, the data must come from a process that is both under control (stable, not influenced by special causes) and meeting expectations. If it does not meet both of these conditions, then it can be used to gauge consistency, but review of results requires significant skepticism (in the pure sense of that word, "doubt and questioning").

Concluding Thoughts

One of the goals of the DQAF is to describe in-line measures—those taken as part of data processing, especially processes that prepare data to be loaded to a database. In-line measures enable measurement of large amounts of complex data in an automated manner. To effectively monitor data quality, they need to be designed to detect unexpected conditions or unexpected changes to data. Guidance for what is unexpected comes from the fundamentals of statistical process control.

An understanding of basic statistical concepts is needed to see how the DQAF measurement types work and how to apply them to particular data. Fortunately, most tools for managing large data stores include software that can execute the calculations themselves. Measurement types that are based on distribution of values (these include most of the consistency measurement types) rely largely on

[4]My initial exposure to the concept of applying SPC to data quality measurement comes from Redman (2001). Loshin's description of SPC for data quality (2011, pp. 99–113) is the clearest, most succinct I have read.

comparisons to the historical mean. Such measures are based on an assumption about consistency of content. If this assumption is not reasonable (if there are good reasons to expect content to be inconsistent), then the measurement type will not produce useful results. If the assumption is reasonable, then historical measurements can surface unusual individual measurement results and increase the chances of discovering potential problems.

Knowledge of the options can help you make better choices. For example, because it is less influenced by extremes, the median may be more effective than the mean for measuring file sizes and process duration. If you have an understanding of your data and the processes that manage it, assumptions about the best options can be tested before measurements are implemented.

Ultimately, the goal of implementing data quality measurement within an organization is to ensure that the organization knows the quality of its data and can take actions to improve and maintain data quality. To achieve these goals requires that you understand measurements of specific data (the data critical to your organization) in relation to the processes that comprise the data chain within your organization. It is said that statistics is as much art as it is science. Because measurements must be interpreted, it important to understand how they are generated and what their limitations are.