THE ABCs of Statistics

By Praveen Gupta

Quality can be viewed objectively or subjectively. If subjective (as in the eyes of the beholder), then quality is just a fancy name for a perception, which is difficult to define. But taken objectively, quality must be measurable.

There are many definitions. The ASQC Standard A3-1987 defines quality as "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs." J.M. Juran's take is simpler: "Quality consists of those product features which meet the needs of customers and thereby provide further satisfaction."

These definitions essentially lead to the same conclusion: Quality is doing the right things right. Here, the right things are the ones that our customers expect from us, and doing them right implies we should do whatever necessary to satisfy customer expectations. And keep in mind who those customers are. For example, a supervisor at a manufacturing company is an operator's customer – a major one, in fact! In this case, quality means doing whatever has been agreed to with your supervisor and doing it correctly. What's more, doing quality work saves on time needed to do the job, reduces product cost, and delights the customer.

One way we maintain and, indeed, improve quality is through the use of statistical tools. statistical techniques are used to identify instances of excessive variation in a process so that corrective actions can be taken to remedy the causes of variation.

A process is any specific sequence of operations that is required to transform a needed set of inputs into a desired output. Control is a feedback mechanism to correct instances of excessive variation. Statistics has to do with collecting sample data, analyzing that data, and making inferences regarding the population.

To employ SPC in the manufacturing environment, there are things one needs to know. In this and next month's column, we will provide a primer of key terms.

Statistics. One dictionary defines statistics as, "The science of the systematic collection of data so as to present descriptive information about a larger population of which the data is (sic) construed as representative, or to infer the significance of underlying factors whose effects are reflected in the data."

You may wonder, "What do you do with the statistics?" Well, statistics is the collection, analysis, and interpretation of quantitative data, from which objective conclusions can be drawn. Take a simple example of an electronics manufacturing facility. During a five-day week, 145, 155, 175, 195, and 205 boards were assembled on each of the five days. If this were a typical week, one could state that, on average, 175 boards are produced per day, even though the actual number of boards manufactured daily varies.

The application of statistical methods is increasing in every industry. We hear about the monthly statistics regarding the unemployment rate, deficit, gross domestic product, inflation rate, etc., all of which are indicators of national performance.

Legislators make policy based on these indicators. Similarly, statistics is used in a company to describe the performance of its operations and assist management in making sound decisions to improve the company's performance.

Thus, statistics is the science of collecting, analyzing, and interpreting data from any operation, manual or automatic, to evaluate its performance (and from which necessary actions are determined to improve the operation).

Descriptive statistics. Statistics can be used to either describe data in statistical terms or to predict the behavior of a population based on sample data. Descriptive statistics deals with the graphical or pictorial presentation of data and the calculation of descriptive measures such as mean, mode, median, range, and standard deviation, among others. Descriptive statistics is used to describe only what one has observed or the historical data.

Inferential statistics. Inferential statistics deals with questions regarding general conclusions or the ability to predict behavior based on available data. Quality control or quality assurance functions emphasize the use of inferential statistics. Using inferential statistics, one can predict the probability of producing defects, then take corrective actions to prevent defects from occurring. For example, control charts are a widely used tool that is based on the principle of inferential statistics. Control charts identify scenarios where the probability of producing defects has become high and the process is out of control. In those cases, the situation must be corrected before the operation is resumed.

Population. Population, in statistical terms, is the group of items available for data collection. For example, if you want to know the average diameter of vias on a board, all the vias on the board make up the population.

Sample. Sample is a part of the population. For example, instead of measuring every via on the board, you may measure a small percentage of vias that represent the population. The data collected from these vias will become the sample data. Then, using inferential statistics, you can determine the average via diameter on the board. If you collect multiple samples, you may get a different average for each, due to sampling variation.

Why collect data from samples rather than the entire population? Since the population is much larger than the sample, it may take many man-hours and personnel to collect the data. To minimize cost, use data from samples that represent the population. Most statistical analyses are performed on sample data rather than the population. There are methods to determine sample size to accurately estimate the characteristics of the population. The most important attribute of sampling is randomness. Samples must be randomly selected.

Attribute data. Data can be either attribute or variable. The attribute data contain information like go/no-go, pass/fail, good/bad, high/low, or positive/negative. With this type of data collection system, defects are measured as percent defective or by count. Examples: pass-or-fail grades, a good or bad movie, accept or reject product.

Variable data. Variable data are quantitative information, measured on a continuous scale. Variable data go beyond attribute data by providing information regarding the degree of performance, that is, how well or poor or how high or low.

Example: If a radio does not work, variable data tell us its voltage output. Variable data are summarized using mean and standard deviation, which describe the distribution of a product's measurable characteristics. Variable data are preferred to attribute data because they contain more information.

Randomization. Applications of statistical techniques in quality control are based on the assumptions that events are independent or uncorrelated. Sometimes we can influence or bias the data due to stratification, our own actions, or other unknown factors. Randomization is a way to eliminate these factors. Randomization is the assignment of samples in a chance manner to various factors during the experiment and is performed to cancel the effects of some known causes and to validate the application of statistical techniques.

Variation. Differences exist from product to product, person to person, or machine to machine. These differences among products, or the process output over time, are called variation. Example: Several TVs may show the same program but have different picture quality. This may be due to the fact that the TVs are tuned differently, or perhaps they were produced by different manufacturers, or there are variations in a manufacturer's operations. There are two types of variation, random (chance) variation or assignable variation.

Random variation. Random variation occurs, as its name implies, due to random causes or chance. Random variation is inherent in a system. It is hard to detect and reduce. For example, if you measure your blood pressure five times on a coin-operated machine in a drugstore, you will get five different readings. These five readings may not be very different from each other. You take an average of the five readings to estimate your average blood pressure. The reason for the variation is inherent in the blood pressure measuring process. The variation can be due to the position of your arm, the machine itself, your mental state, and some other causes, or some combination thereof. Similarly, in manufacturing operations, when a product is built at several workstations by different people using raw material from different suppliers and different tools, random variation will occur from product to product due to such uncontrolled factors.

Assignable variation. Suppose you have been measuring your blood pressure every week in your family drugstore, with satisfactory results. Suddenly, the machine readings indicate your blood pressure is high. There are two potential causes for the new reading: Either your blood pressure has really changed, (temporarily or otherwise), or the machine is out of order. In either case, there is an assignable cause for the high reading. This variation in the blood pressure reading is called the assignable variation, which can be attributed to specific causes. In manufacturing operations, the assignable variation in the product performance occurs due to a change in machine setup, chemicals, operator, procedure, or other specific causes. Assignable variation is easy to detect and easier to reduce than random variation because its causes are known. Once all assignable causes are removed from a process, then the process is in statistical control.

Normal distribution. When you drop a fistful of sand on the ground, it forms a pile or heap that is high in the middle and low at the edges. Similarly, when a frequency distribution of a process or product characteristics is plotted, the distribution appears to be a bell-shaped curve, or the normal distribution. Normal distribution is characterized by two parameters, mean and standard deviation. Calculating probabilities using the normal distribution requires the estimate of the process mean, the standard deviation. In the industrial environment, the normal distribution is used to predict the probability of producing defective product.

Process mean. In manufacturing operations, mean is the value where a process is expected to operate or the target value. If the mean plating thickness is 0.0012, it suggests that most of the time the process will typically deposit a material layer about 0.0012 thick. Average values are plotted to monitor the process output using tools such as trend charts, control charts, or pre-control charts.

Standard deviation. Standard deviation is widely used to quantify the variability of a process. Standard deviation is the square root of the mean sum of the squares of the deviations from the mean. A process capability is defined as six times standard deviation. The standard deviation is a measure of inconsistency in a process.

Specification limits. Specification limits imply an acceptable range of operation for a product. Specification limits are set by customer requirements or by allowable performance criteria. Usually, the specification limits include the lower specification limits (LSL) and the upper specification limits (USL).

Target. Target is the expected value of the product or process performance. Example: The input voltage for appliances is 110 V (in the U.S.); thus, for appliance designers, the target for input voltage is 110 V. A product built to the specs of a well-established target will last much longer than one that is not. Our aim is to build product to target values.

Tolerance. Allowable variation from the target is called tolerance. Tolerances apply to characteristics of the process or the product. Tolerances are specified to verify functionality of the product. Specification limits are used to define allowable tolerance.

Realistic tolerance. Tolerances can be set or defined in many ways, such as based on historical data, experiences, or comparison with similar functions. Tolerances that are determined based on the real capability of the product or the processes are called realistic tolerances.

Statistical problem-solving. Statistical problem-solving is a systematic approach using statistical methods and tools. In industry, problems occur daily: Some are detected and corrected right away, but others are overlooked or missed and can become chronic or catastrophic. There are two objectives to statistical problem-solving. One is to solve an existing problem and the other is to prevent the problem from occurring. A well thought-out approach makes problem-solving a quick process and saves money. The four phases of statistical problem-solving are data collection, analysis, experiment, and corrective action. A few examples of various statistical tools include the histogram, Pareto chart, cause-and-effect diagram, multi-vary chart, control chart, design of experiments, and the response surface methodology.