Quality Management

L_6

27/01/2023

Fundamentals of Statistics

Before a description of the next SPC tool, it is necessary to have a background in statistical fundamentals. *Statistics* is defined as the science that deals with the collection, tabulation, analysis, interpretation, and presentation of quantitative data. Each division is dependent on the accuracy and completeness of the preceding one. Data may be collected by a technician measuring the tensile strength of a plastic part or by an operator using a check sheet.

Measures of Central Tendency

A measure of central tendency of a distribution is a numerical value that describes the central position of the data or how the data tend to build up in the center. There are three measures in common use in quality: (1) the average, (2) the median, and (3) the mode. The average is the sum of the observations divided by the number of observations. It is the most common measure of central tendency and is represented by the equation

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where $\overline{X}$ = average and is read as "X bar"
n = number of observed values
X_i = observed value
Σ = sum of

Median

Another measure of central tendency is the median, Md, which is defined as the value that divides a series of ordered observations so that the number of items above it is equal to the number below it.

Numbers 3, 4, 5, 6, 8, 8, and 10 has a median of 6

Numbers 3, 4, 5, 6, 8, and 8 has a median that is the average of 5 and 6, which is $(5 + 6)/2 = 5.5$

Mode

The mode, Mo, of a set of numbers is the value that occurs with the greatest frequency. It is possible for the mode to be nonexistent in a series of numbers or to have more than one value.

- Numbers 3, 3, 4, 5, 5, 5, and 7 has a mode of 5;
- The series of numbers 22, 23, 25, 30, 32, and 36 does not have a mode;
- The series of numbers 105, 105, 105, 107, 108, 109, 109, 109, 110, and 112 has two modes, 105 and 109

A series of numbers is referred to as unimodal if it has one mode, bimodal if it has two modes, and multimodal if there are more than two modes.

Average

The average is the most commonly-used measure of central tendency. It is used when the distribution is symmetrical or not appreciably skewed to the right or left; when additional statistics, such as measures of dispersion, control charts, and so on, are to be computed based on the average; and when a stable value is needed for inductive statistics.

Note

The median becomes an effective measure of the central tendency when the distribution is positively (to the right) or negatively (to the left) skewed. The median is used when an exact midpoint of a distribution is desired. When a distribution has extreme values, the average will be adversely affected, whereas the median will remain unchanged.

Average

Thus, in a series of numbers such as 12, 13, 14, 15, 16, the median and average are identical and are equal to 14. However, if the first value is changed to a 2, the median remains at 14, but the average becomes 12.

A control chart based on the median is user-friendly and excellent for monitoring quality. The mode is used when a quick and approximate measure of the central tendency is desired. Thus, the mode of a histogram is easily found by a visual examination. In addition, the mode is used to describe the most typical value of a distribution, such as the modal age of a particular group.

Measures of Dispersion

A second tool of statistics is composed of the measures of dispersion, which describe how the data are spread out or scattered on each side of the central value. Measures of dispersion and measures of central tendency are both needed to describe a collection of data.

To illustrate, the employees of the plating and the assembly departments of a factory have identical average weekly wages of $325.36; however, the plating department has a high of $330.72 and a low of $319.43, whereas the assembly department has a high of $380.79 and a low of $273.54. The data for the assembly department are spread out, or dispersed, farther from the average than are those of the plating department.

Measures of Dispersion

One of the measures of dispersion is the range, which for a series of numbers is the difference between the largest and smallest values of observations. Symbolically, it is represented by the

$$R = X_h - X_l$$

where R = range
X_h = highest observation in a series
X_l = lowest observation in a series

The other measure of the dispersion used in quality is the standard deviation. It is a numerical value in the units of the observed values that measures the spreading tendency of the data. A large standard deviation shows greater variability of the data than does a small standard deviation. In symbolic terms, it is represented by the equation

Measures of Dispersion

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

where s = sample standard deviation
X_i = observed value
$\overline{X}$ = average
n = number of observed values

Unless otherwise noted, s stands for s_x, the sample standard deviation of observed values. The same equation is used to find

$S_{\overline{X}}$ = sample standard deviation of averages
s_p = sample standard deviation of proportions
s_s = sample standard deviation of standard deviations, etc.

Measures of Dispersion

The standard deviation is a reference value that measures the dispersion in the data. It is best viewed as an index that is defined by the formula. The smaller the value of the standard deviation, the better the quality, because the distribution is more closely compacted around the central value. The standard deviation also helps to define populations.

Measures of Dispersion

1. In quality control the range is a very common measure of the dispersion.
2. It is used in one of the principal control charts.
3. The primary advantage of the range is in providing a knowledge of the total spread of the data.
4. It is also valuable when the amount of data is too small or too scattered to justify the calculation of a more precise measure of dispersion.
5. As the number of observations increases, the accuracy of the range decreases, because it becomes easier for extremely high or low readings to occur.
6. It is suggested that the use of the range be limited to a maximum of ten observations.
7. The standard deviation is used when a more precise measure is desired.

EXAMPLE PROBLEM

Determine the average, median, mode, range, and standard deviation for the height of seven people. Data are 1.83, 1.91, 1.78, 1.80, 1.83, 1.85, 1.87 meters.

$$\overline{X} = \Sigma X/n = (1.83 + 1.91 + \cdots + 1.87)/7 = 1.84$$

$$\text{Mo} = 1.83$$

$$\text{Md} = \{1.91,\ 1.87,\ 1.85,\ 1.83,\ 1.83,\ 1.80,\ 1.78\} = 1.83$$

$$R = X_h - X_l = 1.91 - 1.78 = 0.13$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(1.91 - 1.84)^2 + \cdots + (1.78 - 1.84)^2}{7-1}}$$
$$= 0.04$$

Population and Sample

In order to construct a frequency distribution of the weight of steel shafts, a small portion, or sample, is selected to represent all the steel shafts. The population is the whole collection of steel shafts. When averages, standard deviations, and other measures are computed from samples, they are referred to as statistics.

Because the composition of samples will fluctuate, the computed statistics will be larger or smaller than their true population values, or parameters. Parameters are considered to be fixed reference (standard) values or the best estimate of these values available at a particular time. The population may have a finite number of items, such as a day's production of steel shafts.

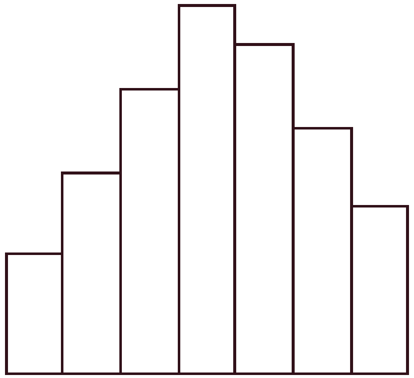
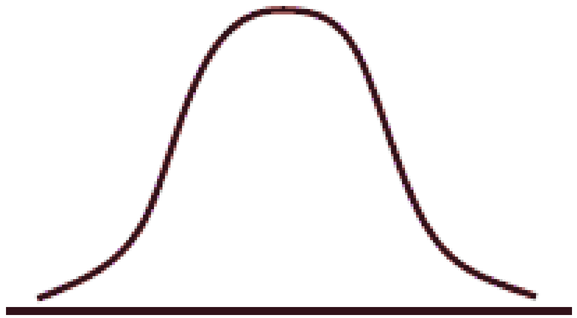
Population and Sample

Because it is rarely possible to measure all of the population, a sample is selected. Sampling is necessary when it may be impossible to measure the entire population; when the expense to observe all the data is prohibitive; when the required inspection destroys the product; or when a test of the entire population may be too dangerous, as would be the case with a new medical drug.

Population and Sample

When designating a population, the corresponding Greek letter is used. Thus, the sample average has the symbol $\bar{X}$ and the population mean the symbol μ (mu). Note that the word *average* changes to *mean* when used for the population. The symbol $\bar{X}_0$ is the standard or reference value. Mathematical concepts are based on μ, which is the true value— $\bar{X}_0$ represents a practical equivalent in order to use the concepts. The sample standard deviation has the symbol s, and the population standard deviation the symbol σ (sigma). The symbol s_0 is the standard or reference value and has the same relationship to σ that $\bar{X}_0$ has to μ. The true population value may never be known; therefore, the symbol $\hat{\mu}$ and $\hat{\sigma}$ are sometimes used to indicate "estimate of."

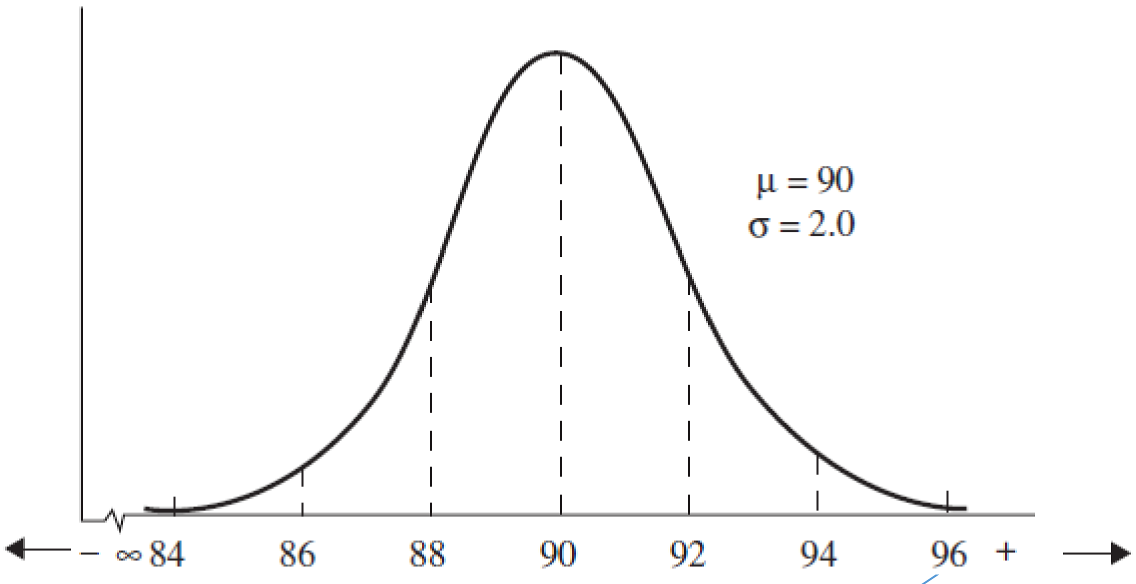
Comparison of Sample and Popultion

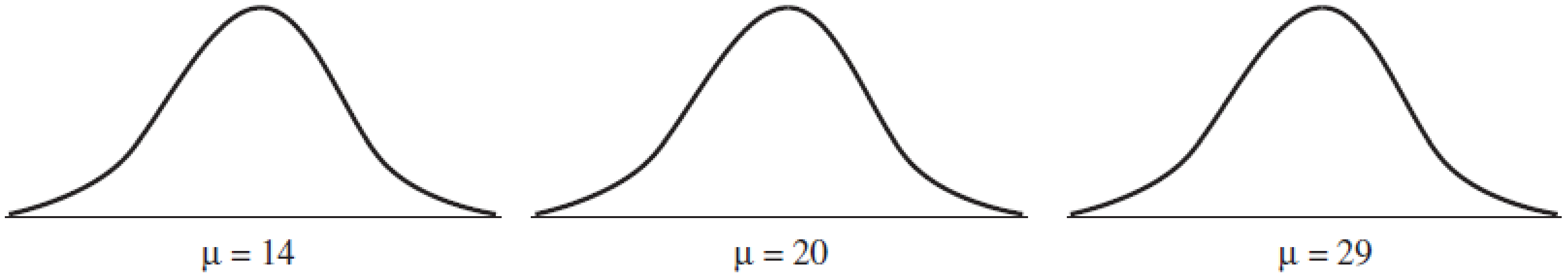
Sample	Population
Statistic X—average s—sample standard deviation	Parameter (X_0)—mean $\sigma(s_0)$—standard deviation
	

How successfully the sample represents the population is a function of the size of the sample, chance, the sampling method, and whether or not the conditions change.

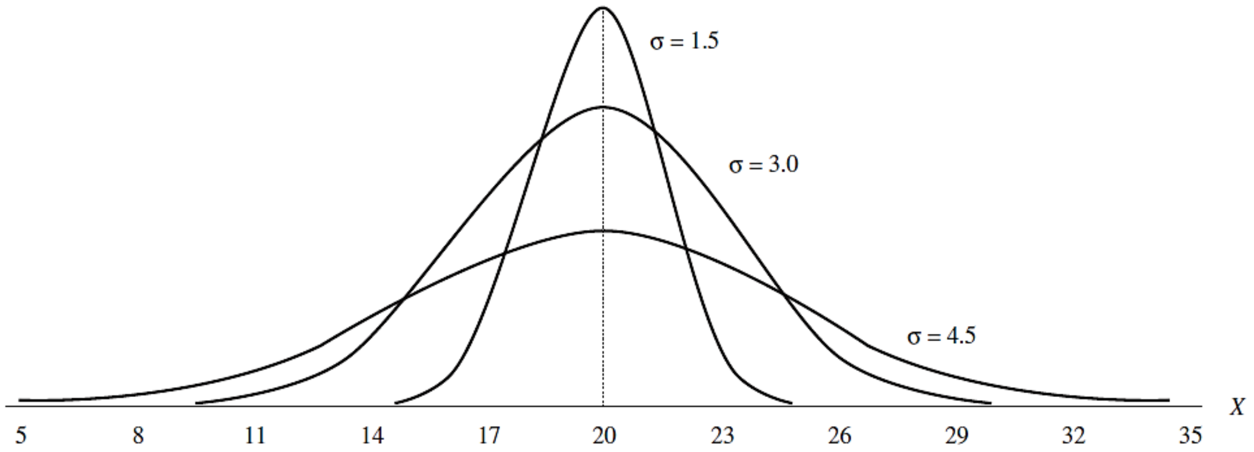
Normal Curve

Although there are as many different populations as there are conditions, they can be described by a few general types. One type of population that is quite common is called the normal curve, or Gaussian distribution. The normal curve is a symmetrical, unimodal, bell-shaped distribution with the mean, median, and mode having the same value.





Normal Curves with Different Means but identical Standard Deviations

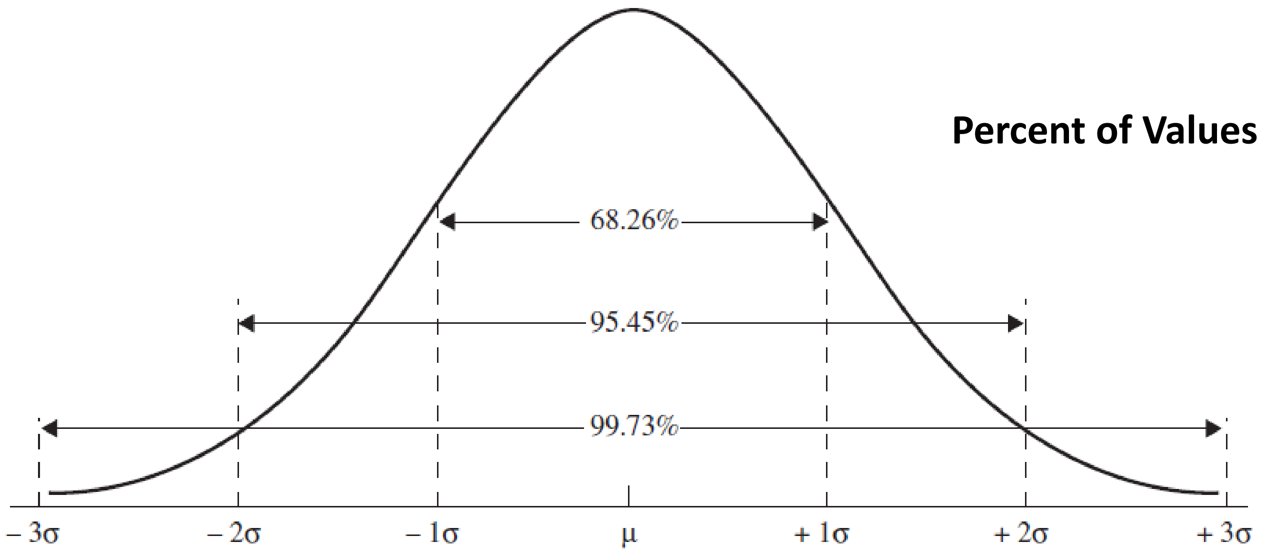


The normal distribution is fully defined by the population mean and population standard deviation. Also, as seen by Figures, these two parameters are independent. In other words, a change in one parameter has no effect on the other.

Normal Curves with Different Standard Deviations but identical Means

Normal Curve

A relationship exists between the standard deviation and the area under the normal curve, as shown in Figure. The figure shows that in a normal distribution, 68.26% of the items are included between the limits of $\mu + 1\sigma$ and $\mu - 1\sigma$, 95.46% of the items are included between the limits $\mu + 2\sigma$ and $\mu - 2\sigma$, and 99.73% of the items are included between $\mu + 3\sigma$ and $\mu - 3\sigma$.



Percent of Values Included Between Certain Values of the Standard Deviation

Introduction to Control Charts

Variation

One of the axioms, of production is that no two objects are ever made exactly alike. In fact, the variation concept is a law of nature because no two natural items in any category are the same. The variation may be quite large and easily noticeable, such as the height of human beings, or the variation may be very small, such as the weights of fiber-tipped pens or the shapes of snowflakes. When variations are very small, it may appear that items are identical; however, precision instruments will show differences.

Introduction to Control Charts

Variation

There are three categories of variations in piece part production:

1. Within-piece variation is illustrated by the surface roughness of a piece, wherein one portion of the surface is rougher than another portion or the width of one end of a keyway varies from the other end.

2. Piece-to-piece variation occurs among pieces produced at the same time. Thus, the light intensity of four consecutive light bulbs produced from a machine will be different.

3. Time-to-time variation is illustrated by the difference in product produced at different times of the day. Thus, product produced in the early morning is different from that produced later in the day, or as a cutting tool wears, the cutting characteristics change.

Introduction to Control Charts

The first source of variation is the equipment. This source includes tool wear, machine vibration, workholding-device positioning, and hydraulic and electrical fluctuations.

The second source of variation is the material. Because variation occurs in the finished product, it must also occur in the raw material (which was someone else's finished product).

A third source of variation is the environment. Temperature, light, radiation, particle size, pressure, and humidity all can contribute to variation in the product.

A fourth source is the operator. This source of variation includes the method by which the operator performs the operation.

- The preceding four sources account for the true variation. There is also a reported variation, which is due
- to the inspection activity. Faulty inspection equipment, the incorrect application of a quality standard, or too
- heavy a pressure on a micrometer can be the cause of the incorrect reporting of variation.

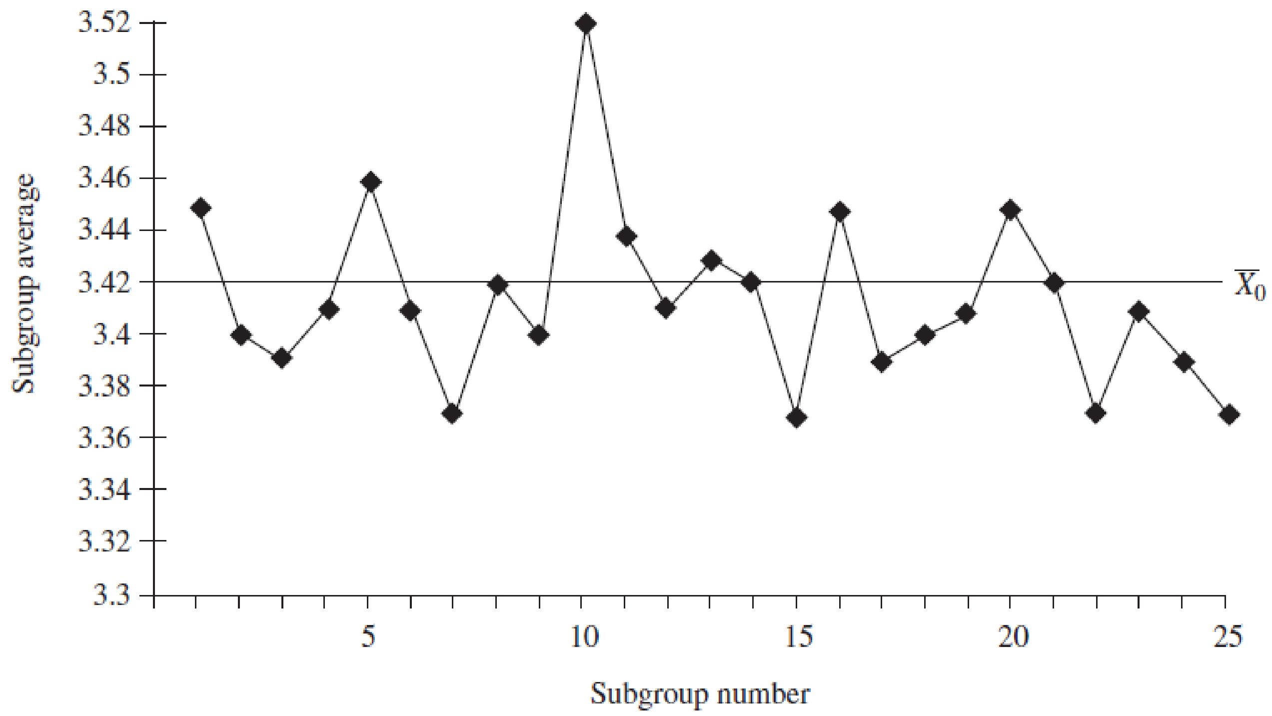
Introduction to Control Charts

Run Chart

A run chart, which is shown in Figure, is a very simple technique for analyzing the process in the development stage or, for that matter, when other charting techniques are not applicable. The important point is to draw a picture of the process and let it "talk" to you. A picture is worth a thousand words, provided someone is listening. Plotting the data points is a very effective way of finding out about the process. This activity should be done as the first step in data analysis. Without a run chart, other data analysis tools—such as the average, sample standard deviation, and histogram—can lead to erroneous conclusions.

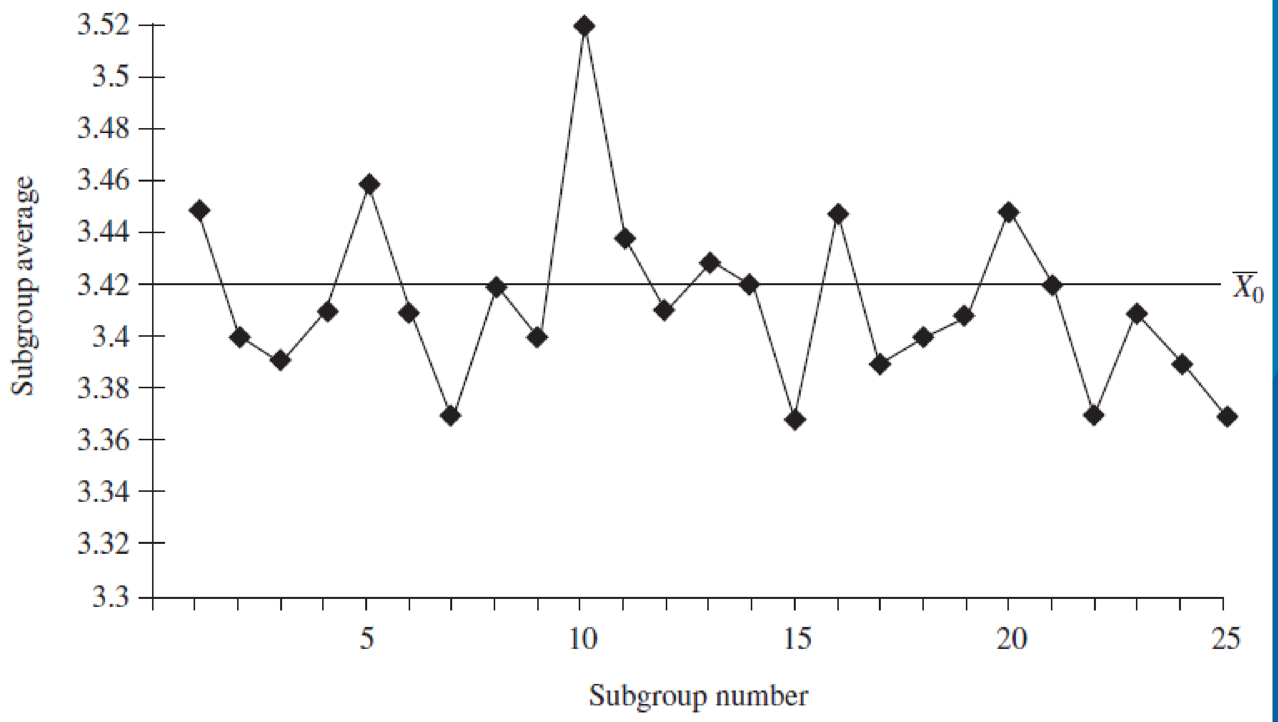
Introduction to Control Charts

The particular run chart shown in Figure is referred to as an *X* chart and is used to record the variation in the average value of samples. Other charts, such as the *R* chart (range) or *p* chart (proportion) would have also served for explanation purposes. Each small solid diamond represents the average value within a subgroup. Thus, subgroup number 5 consists of, say, four observations, 3.46, 3.49, 3.45, and 3.44, and their average is 3.46 kg. This value is the one posted on the chart for subgroup number 5.



Introduction to Control Charts

Averages are used on control charts rather than individual observations because average values will indicate a change in variation much faster. Also, with two or more observations in a sample, a measure of the dispersion can be obtained for a particular subgroup. Averages are used on control charts rather than individual



Introduction to Control Charts

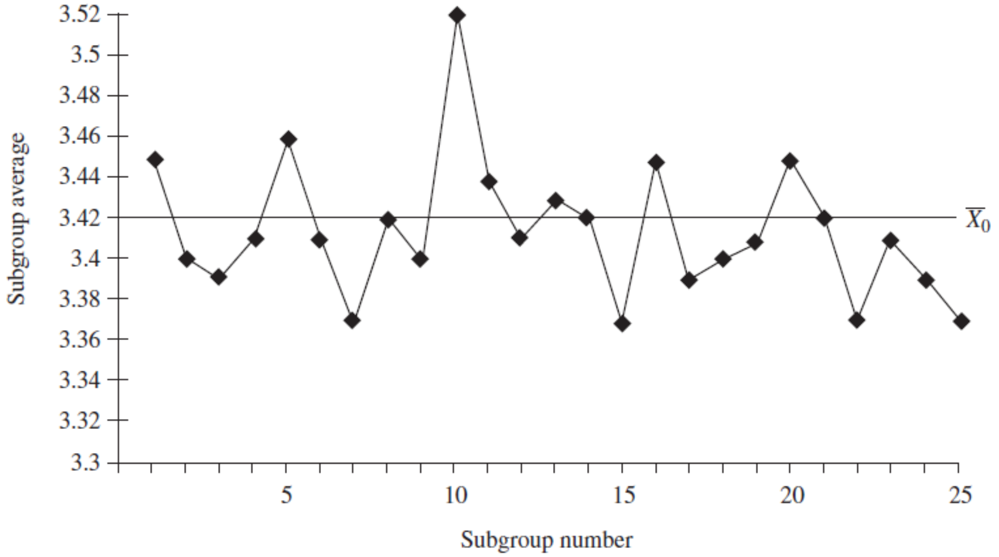
The solid line in the center of the chart can have three different interpretations, depending on the available data. First, it can be the average of the plotted points, which in the case of an $\overline{X}$ chart is the average of the averages or "X-double bar." Second, it can be a standard or reference value, $\overline{X}_0$, based on representative prior data, an economic value based on production costs or service needs, or an aimed-at value based on specifications. Third, it can be the population mean, μ, if that value is known.

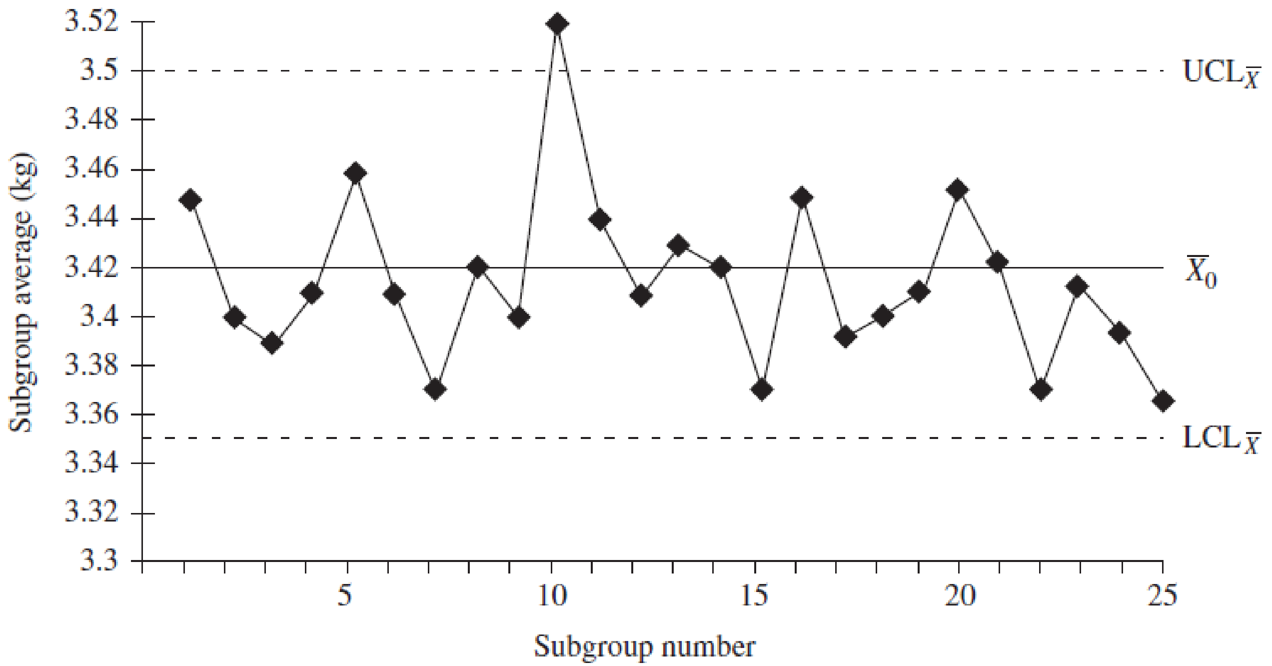
Control Chart Example

One danger of using a run chart is its tendency to show every variation in data as being important. In order to indicate when observed variations in quality are greater than could be left to chance, the control chart method of analysis and presentation of data is used.

The control chart method for variables is a means of visualizing the variations that occur in the central tendency and dispersion of a set of observations. It is a graphical record of the quality of a particular characteristic. It shows whether or not the process is in a stable state by adding statistically determined control limits to the run chart.

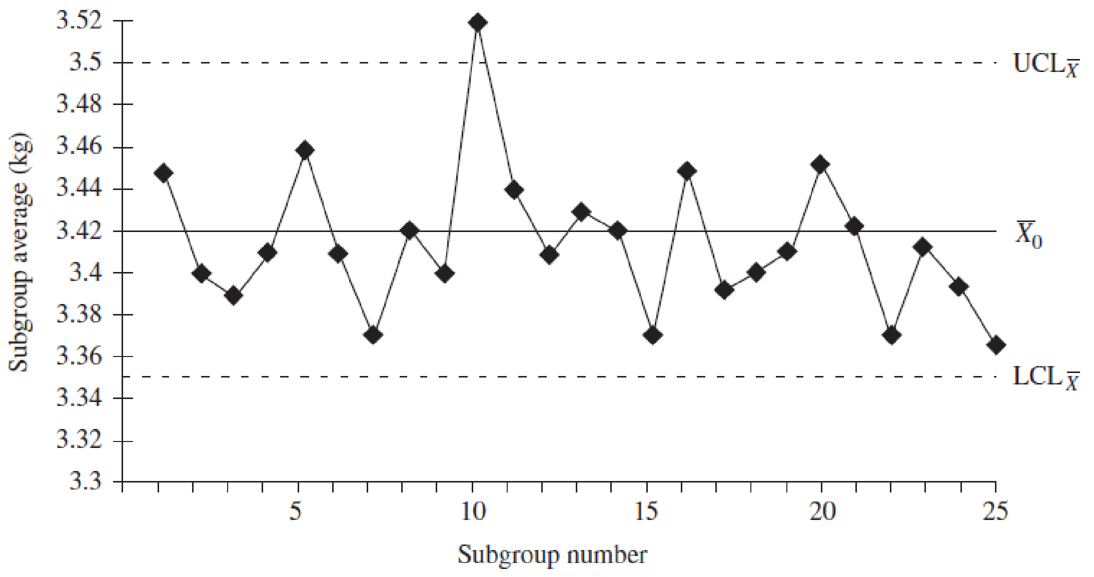
Control Chart Example





Control Chart Example

They are the two dashed outer lines and are called the upper and lower control limits. These limits are established to assist in judging the significance of the variation in the quality of the product. Control limits are frequently confused with specification limits, which are the permissible limits of a quality characteristic of each individual unit of a product. However, control limits are used to evaluate the variations in quality from subgroup to subgroup. Therefore, for the X-bar chart, the control limits are a function of the subgroup averages. A frequency distribution of the subgroup averages can be determined with its corresponding average and standard deviation.



Control Chart Example

The control limits are then established at ± 3σ from the central line. Recall, from the discussion of the normal curve, that the number of items between + 3σ and -3σ equals 99.73%. Therefore, it is expected that more than 997 times out of 1,000, the subgroup values will fall between the upper and lower limits.