PDHonline Course P208 (1 PDH)

# Statistical Methods for Process Improvement - Part 2: Using the Normal Distribution

*Instructor: Davis M. Woodruff, PE, CMC*

**2020**

**PDH Online | PDH Center**

5272 Meadow Estates Drive
Fairfax, VA 22030-6658
Phone: 703-988-0088
www.PDHonline.com

An Approved Continuing Education Provider

# Statistical Methods for Process Improvement Part 2: Using the Normal Distribution

***Davis M. Woodruff, PE, CMC***

*"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." John Tukey*

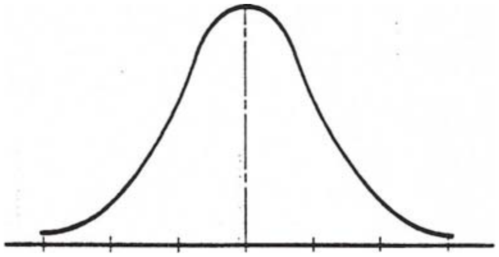
## Course Content

### Introduction

The primary distribution utilized in Statistical Process Control and in using statistical methods for process analysis is the normal distribution. This distribution occupies a prominent place in statistics for the following reasons:

- The distribution is applicable to many situations in which it is necessary to make inferences from samples taken from a population.
- A normal distribution comes close to fitting the actual observed frequency distribution of many phenomena (weights, heights, times, etc.) and it adequately describes outputs from many processes (dimensions, yields, etc.).

The distribution, a probability density function, resembles a bell-shaped curve as shown below.



**Normal Distribution Curve**

### The Equation of the Distribution

The general formula for the probability density function of the normal distribution is:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

Since this is a basic course that emphasizes using the distribution, we will not be developing curves using the equation. Rather, it is shown here for completeness and to show that the mean (μ) and the standard deviation (σ) are the two parameters of the distribution. The mean locates the distribution on a given scale and the standard deviation describes the "spread" of the distribution. For more information on the average and standard deviation and using descriptive statistics and pictures of data, see **Statistical Methods for Process Analysis, Part 1: Using Data for Process Improvement.**

### Characteristics of a Normal Distribution

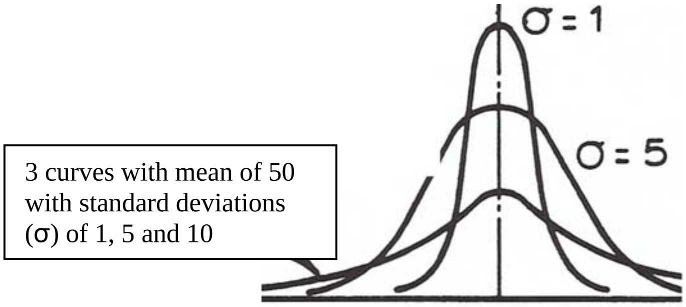
The following are the basic characteristics of a normal distribution:

- The curve has a single peak; it is uni-modal.
- The mean, median and mode are exactly the same and lie at the center of the distribution.
- The curve is symmetrical; each side is a mirror image of the other.
- The two tails of the curve extend indefinitely. In other words, the tails of the curve approach but never reach the horizontal axis.
- The area under the curve represents the probability of occurrence. The total area is 1.

## Significance of the Mean and the Standard Deviation

There is no single normal curve, but a family of curves which are accurately described by their mean and standard deviation. The mean locates the curve; the standard deviation measures the spread of the data. Any number of distributions could have a mean of, say 50, yet the curves be very different if the standard deviations are different.

Three different normal distribution curves are shown below. Each of these curves has a mean(μ) of 50, but varying standard deviations. Even though the curves are located at an identical spot on the horizontal axis, the spread or variability of the curves is quite different.

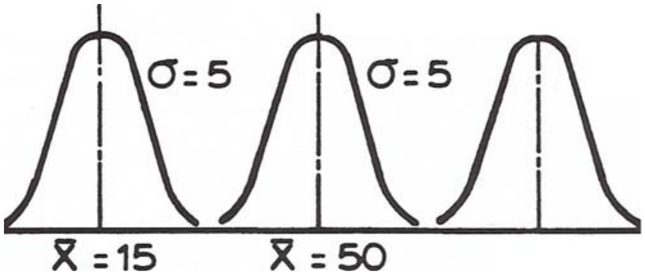


3 curves with mean of 50 with standard deviations (σ) of 1, 5 and 10

**Mean = 50**

**Family of Normal Curves with Different Standard Deviations and the Same Mean (50)**

A similar graph showing differing locations of the mean is shown in Figure V-3. Curve A has a mean of 15 and is centrally located at 15 on the horizontal axis. Curve B has a mean of 50 and is located on the horizontal axis at 50, and Curve C has a mean of 85. All of these curves have the same standard deviation or variability, and so, have the same shape. The mean merely indicates the relative location of the distribution.

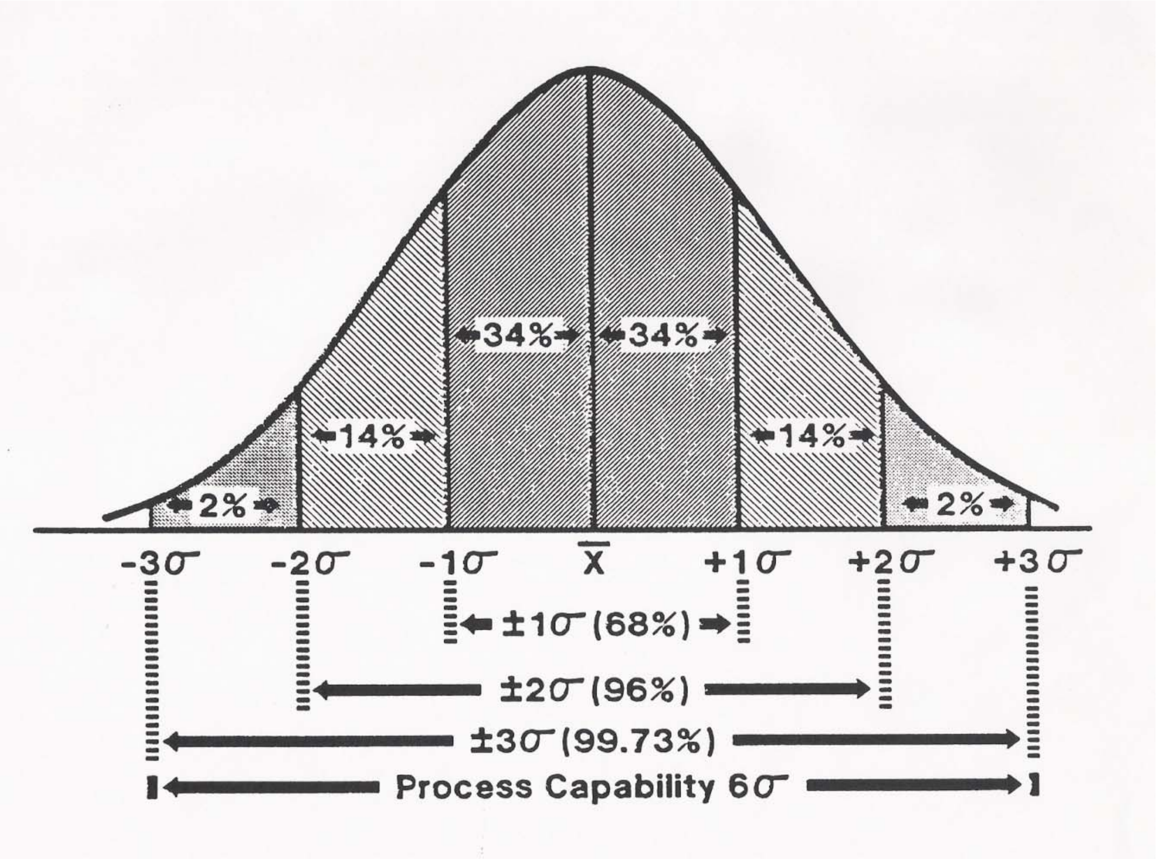


**Family of Normal Curves**
**with Different Means and Same Standard Deviation**

Regardless of the value of the mean and standard deviation of any normal distribution, the total area under the curve is always considered equal to 1.00. This allows us to look at the areas under the curve as probabilities. The area under the curve has been calculated for various standard deviations from the mean and are shown in the graph of the normal distribution. From this graph, it is readily apparent that the area under the curve between plus and minus one standard deviation is approximately 68%. Specifically, these probabilities are summarized in Table V-1.

**Areas Under the Normal Curve for One, Two, and**
**Three Standard Deviations**

| Limits | Area (or Probability) |
|---|---|
| $\bar{X} \pm 1\sigma$ | 0.6826 |
| $\bar{X} \pm 2\sigma$ | 0.9546 |
| $\bar{X} \pm 3\sigma$ | 0.9973 |
| $\bar{X} \pm 4\sigma$ | 0.9999 |

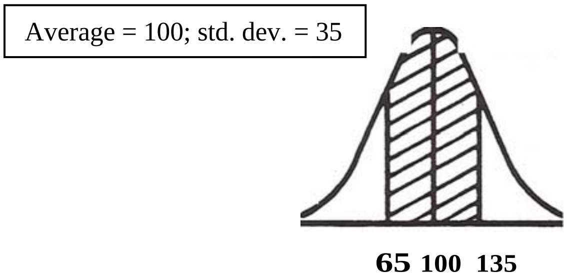


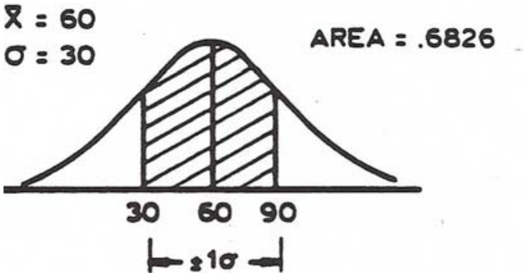
**Area Under the Normal Curve**

### Areas Under the Normal Curve

While it is convenient that the areas under the normal curve have been calculated for exactly l, 2, and 3 standard deviations from the mean, use of the normal distribution frequently involves intervals other than these. Statistical tables have been compiled for precisely these situations and using excel® it is easy to perform a variety of important, yet simple calculations using these distributions. They indicate portions of the area under the normal curve that are contained within any number of standard deviations from the mean.

It is not necessary to have a different table for every normal curve. Instead, we use a standard normal probability distribution table to find areas under any normal curve. This is possible because the areas under the curve are identical for given standard deviations. In other words, the area for ±lσ is 0.6826 regardless of the mean or standard deviation of the distribution. The drawings below show this area.


Average = 100; std. dev. = 35



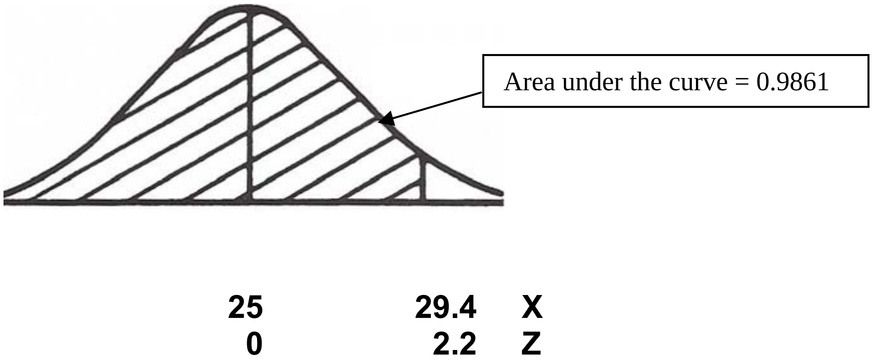
**Area Under Two Normal Curves**
**for ±1 Standard Deviations**

A Standard Normal Probability Distribution Table can be found in most statistics books. These tables contain two types of numbers: the areas under the curve and Z scores. For this reason, it is often referred to as a Z table. The Z scores are actually standard deviation units or the number of standard deviations a given value in the distribution is from the distribution mean. The equation for a Z score is:

$$z = \text{Point of interest on the curve - Average/Standard Deviation}$$

The $\overline{X}$ and σ are the mean and standard deviation of the distribution and X is the value or point of interest in the distribution with which we are concerned.

Really, the Z value is nothing more than a change of scale of measurements on the horizontal axis.


Area under the curve = 0.9861

**Area Under the Normal Curve for a Z Score of 2.20**

In this example, suppose a normal distribution has $\overline{X}$ = 25.0 and σ of 2. 0. If we were interested in the value of 29.4 of this distribution, it would represent a Z score of:

**Z = (29.4-25)/2.0 = 2.20**

This leads to the other numbers in the Z table, the areas under the curve. These numbers represent the area under the curve from minus infinity to the Z score. Looking up Z = 2.20 in the Z table, we find the area is 0.9861 which is shown in the curve above. Thus, we can conclude that 98.61% of the values will be 29.4 or less.

The probabilities of areas under the curve may be found by following these step-by-step instructions:

I. Determine the mean and the standard deviation of the process.

2. Draw a normal curve and indicate the mean and the area of interest.

3. Calculate the Z score

4. Find the area from the Z table corresponding to the Z score obtained in step 3.

5. Read the value from the Z table and relate it to the area of interest under the curve.

6. Convert the area to percent by multiplying by 100.

Or, using excel® follow these steps:

1. Go to f(x), then go to “statistical function”
2. Next, select Normal Distribution
3. Enter x, the point of interest, average, and standard deviations when prompted
4. Enter TRUE for cumulative probability

5. The answer shown on the screen is the area under the curve to **(–)**∞
6. If we are interested in the value higher than the point of interest on the curve, then use 1**–** area under the curve to **(–)**∞

The following example demonstrates the use of Z scores using excel® instead of converting it to the standard normal distribution and using Z score tables.

**Example: Evaluating the potential impact of a customer specification change**

The current upper specification for a product is 0.0023 inches and the average is 0.00175. The customer is requesting the upper limit be reduced to 0.0020 inches and the lower limit remains unchanged. The standard deviation is 0.001. What is the potential impact of this change?

Using excel® we can readily evaluate the potential impact assuming the average and standard deviation remain unchanged:

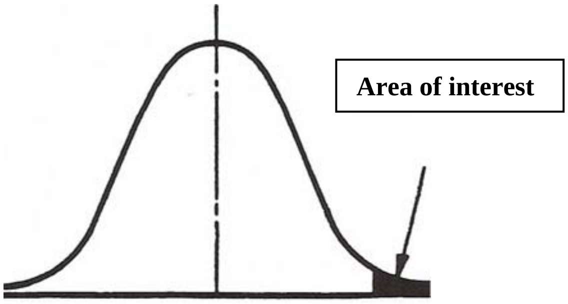
1. The area under the curve from 0.0023 to minus infinity is 0.7088, which means that 70.88% of the product will be less than the upper limit of 0.0023.

2. The area under the curve from 0.0020 to minus infinity is 0.5987, which means that 59.87% of the product will be less than 0.0020.

3. The impact will be 11.01% more rejected material if the upper limit is reduced and there are no other process changes to compensate for the specification change. The alternatives would be to lower the average by changing the process or reducing the variability.

## Example 2-- Wastewater Discharge Limits

The plant operating permit requires that a certain chemical in the discharge be **<** 9 ppm. In April, the values averaged 8.49 ppm with a standard deviation of 0.19. You are interested in knowing if the amount of chemical in the wastewater exceeded the upper limit of 9 ppm, and if so, what % of the time this occurred. Just to be sure we all understand the concept, let’s do this one by hand and then use excel®. First, draw a normal curve with a mean of 8.49 and σ of 0.19 to be sure we are answering the “right” question. This is shown below:



**Normal Curve, $\bar{X}$ = 8.49, σ = 0.19**

Next, calculate the Z-score for a value of 9:

$$Z = (9 – 8.49)/0.19$$

Remember, the area of interest is discharge ppm > 9, so the table value must be subtracted from I.00. The probability associated with a Z score of 2.68 is 0.9963, or in other words this is the area under the normal curve to the left of 9. The area of interest is: I.00 - 0.9663 = 0.0037. So the values of this particular chemical would exceed the upper limit of 9 about 0.37% of the time.

In excel®, this is simply a matter of plugging in the values and subtracting the cumulative probability from 1.000 just as done above.

These two examples illustrate the versatility and power of using the standard normal distribution. Remember that the major consideration is that the underlying distribution is a normal distribution!

In my consulting practice I have recommended and observed the use of this powerful technique in many different settings to answer many questions. In one case, the question involved how long it should take most employees in a large organization to complete an on-line training class. In this particular case, an average and standard deviation was calculated using 30 different employees selected randomly from the population to be tested. From this information and assuming the population was normally distributed, the client's HR department determined the expected time requirement for ~95% of the population and planned the training sessions accordingly. The result was a smoothly run training session. This is just another example of the many uses of the normal distribution in business today.

In another case a Plant Manager of a large corporation told me that at last he had a tool to use when evaluating specification changes being proposed by the marketing department. Obviously, that had been a significant issue for his operation.

The focus should be on fact based decision making and answering the "right" questions using data from the real world.

## Glossary of Statistical Terms

**Accuracy** - The closeness of agreement between and observed value and an accepted reference value.

**Alternative Hypothesis** - The hypothesis that is accepted if the null hypothesis is disproved.

**Attribute Data** - Qualitative data that typically shows only the number of articles conforming and the number of articles failing to conform to a specified criteria. Sometimes referred to as Countable Data.

**Average** - The sum of the numerical values in a sample divided by the number of observations.

**Bar Chart** - A chart that uses bars to represent data. This type of chart is usually used to show comparisons of data from different sources.

**Batch** - A definite quantity of some product or material produced under conditions that are considered uniform.

**Bias** - A systematic error which contributes to the difference between a population mean of measurements or test results and an accepted reference value.

**Bimodal Distribution** - A distribution with two modes that may indicate mixed data.

**Binomial Distribution** - A distribution resulting from measured data from independent evaluation, where each measurement results in either success or failure and where the true probability of success remains constant from sample to sample.

**Cells** - The bars on a histogram each representing a subgroup of data.

**Common Cause** - A factor or event that produces normal variation that is expected in a given process.

**Confidence** - The probability that an interval about a sample statistic actually includes the population parameter.

**Control Chart** - A chart that shows plotted values, a central line and one or two control limits and is used to monitor a process over time.

**Control Limits** - A line or lines on a control chart used as a basis for judging the significance of variation from subgroup to subgroup. Variation beyond a control limit shows that special causes may be affecting the process. Control limits are usually based on the 3 standard deviation limits around an average or central line.

**Countable Data** - The type of data obtained by counting -attribute data.

**Data** - Facts, usually expressed in numbers, used in making decisions.

**Data Collection** - The process of gathering information upon which decisions to improve the process can be based.

**Detection** - A form of product control, not process control, that is based on inspection that attempts to sort good and bad output. This is an ineffective and costly method.

**Distribution** - A group of data that is describable by a certain mathematical formula.

**Frequency Distribution** - A visual means of showing the variation that occurs in a given group of data. When enough data have been collected, a pattern can usually be observed.

**Histogram** - A bar chart that represents data in cells of equal width. The height of each cell is determined by the number of observations that occur in each cell.

**k** - The symbol that represents the number of subgroups of data. For example, the number of cells in a given histogram.

**Lower Control Limit (LCL)** - The line below the central line on a control chart.

**Mean** - The average value of a set of measurements, see Average.

**Measurable Data** - The type of data obtained by measurement. This is also referred to as variables data. An example would be diameter measured in millimeters.

**Median** - The middle value ( or average of the two middle values) of a set of observations when the values have been ranked according to size.

**Mode** - The most frequent value in a distribution. The mode is the peak of a distribution.

**n** - The symbol that represents the number of items in a group or sample.

**np** - The symbol that represents the central line on an np chart.

**Nonconformities** - Something that doesn't conform to a drawing or specification; an error or reason for rejection.

**Normal Distribution** - A symmetrical, bell-shaped frequency distribution for data. This is a distribution that is often seen in industry.

**Null Hypothesis** - The hypothesis tested in test of significance that there is no difference (null) between the population of the sample and the specified population (or between the populations associated with each sample).

**Out of Control** - The condition describing a process from which all special causes of variation have not been eliminated. This condition is evident on a control chart by the presence of points outside the control limits or by nonrandom patterns within the control limits.

**p** - The symbol on a p chart that represents the proportion of nonconforming units in a sample.

**Parameter** - A constant or coefficient that describes some characteristics of a population (e.g. standard deviation, average, regression coefficient).

**Pareto Charts** - A bar chart that arranges data in order of importance. The bar representing the item that occurs or costs the most is placed on the left-hand side of the horizontal axis. The remaining items are placed on the axis in descending (most to least) order.

**Population** - All members, or elements, of a group of items. For example, the population of parts produced by a machine includes all of the parts the machine has made. Typically in SPC we use samples that are representative of the population.

**Prevention** - A process control strategy that improves quality by directing analysis and action towards process management, consistent with the philosophy of continuous quality improvement.

**Process** - Any set of conditions or set of causes working together to produce an outcome. For example, how a product is made.

**Product** - What is produced; the outcome of a process.

**Quality** - Conformance to requirements.

**Random Sampling** - A data collection method used to ensure that each member of a population has an equal chance of being part of the sample. This method leads to a sample that is representative of the entire population.

**Range** - The difference between the highest and lowest values in a subgroup.

**Repeatability** - The variation in measurements obtained when one operator uses the same test for measuring the identical characteristics of the same samples.

**Reproducibility** - The variation in the average of measurements made by different operators using the same test when measuring identical characteristics of the same samples.

**Run Chart** - A line chart that plots data from a process to indicate how it is operating.

**Sample** - A small portion of a population.

**Sampling** - A data collection method in which only a portion of everything made is checked on the basis of the sample being representative of the entire population.

**Significance Level.()** - The risk we are willing to take of rejecting a null hypothesis that is actually true.

**Skewed Distribution** - A distribution that tapers off in one direction. It indicates that something other than normal, random factors are effecting the process. For example, TIR is usually a skewed distribution.

**Special Cause** - Intermittent source of variation that is unpredictable, or unstable; sometimes called an assignable cause. It is signaled by a point beyond the control limits or a run or other nonrandom pattern of points within the control limits. The goal of SPC is to control the special cause variation in a process.

**Specification** - The extent by which values in a distribution differ from one another; the amount of variation in the data.

**Standard Deviation (σ)** - The measure of dispersion that indicates how data spreads out from the mean. It gives information about the variation in a process.

**Statistic** - A quantity calculated from a sample of observations, most often to form an estimate of some population parameter.

**Statistical Control** - The condition describing a process from which all special causes of variation have been eliminated and only common causes remain, evidenced by the absence of points beyond the control limits and by the absence of nonrandom patterns or trends within the control limits.

**Statistical Methods** - The means of collecting, analyzing, interpreting and presenting data to improve the work process.

**Statistical Process Control (SPC)** - The use of statistical techniques to analyze data, to determine information, and to achieve predictability from a process.

**Statistics** - A branch of mathematics that involves collecting, analyzing, interpreting and presenting masses of numerical data.

**Subgroup** - A group of consecutively produced units or parts from a given process.

**Tally or Frequency Tally** - A display of the number of items of a certain measured value. A frequency tally is the beginning of data display and is similar to a histogram.

**Tolerance** - The allowable deviation from standard, i.e., the permitted range of variation about a nominal value. Tolerance is derived from the specification and is NOT to be confused with a control limit.

**Trend** - A pattern that changes consistently over time.

**Type I Error (α)** - The incorrect decision that a process is unacceptable when, in fact, perfect information would reveal that it is located within the "zone of acceptable processes".

**Type II Error(β)** - The incorrect decision that a process is acceptable when, in fact, perfect information would reveal that it is located within the "zone of rejectable processes".

**u** - The symbol used to represent the number of nonconformities per unit in a sample which may contain more than one unit.

**Upper Control Limit (UCL)** - The line above the central line on a control chart.

**Variables** - A part of a process that can be counted or measured, for example, speed, length, diameter, time, temperature and pressure.

**Variable Data** - Data that can be obtained by measuring. See Measurable Data.
**Variation** - The difference in product or process measurements. A change in the value of a measured characteristic. The two types of variation are within subgroup and between subgroup. The sources of variation can be grouped into two major classes: Common causes and Special Causes.

**X** - The symbol that represents an individual value upon which other subgroup statistics are based.

**$\bar{X}$ (x bar)** - The average of the values in a subgroup.

**$\bar{\bar{X}}$ (x double bar)** - The average of the averages of subgroups.

**Suggested Readings and/or References:**

Duncan, A.J., *Quality Control and Industrial Statistics*, 5th ed. Irwin.

Grant, E.L. and Leavenworth, R.S., *Statistical Quality Control*, 5th ed., McGraw Hill.

Blank, L.T., *Statistical Procedures for Engineering, Management and Science*, 1st ed, McGraw Hill.

Wadsworth, H.M, Stephens, K.S. and Godfrey, A.B., *Modern Methods for Quality Control and Improvement*, 1st ed., Wiley.

Juran, J.M., et al, *Quality Control Handbook*, McGraw Hill

Ott, E.R., *Process Quality Control: Troubleshooting and Interpretation of Data*, McGraw hill

Shewhart, W.A., *Economic Control of Quality of Manufactured Products*, Van Nostrand.

Levin, R.I., *Statistics for Management*, 3rd ed., McGraw Hill.

Feigenbaum, A.V., *Total Quality Control*, 3rd ed., McGraw Hill

*Statistical Quality Control Handbook*, Western Electric.

Charbonneau, Harvey C. and Webster, Gordon L., *Industrial Quality Control.*