

CHAPTER 4

BASIC QUALITY CONCEPTS

- 1.0 Continuous Distributions
- 2.0 Measures of Central Tendency
- 3.0 Measures of Spread or Dispersion
- 4.0 Histograms and Frequency Distributions
- 5.0 Shapes of Distributions
- 6.0 The Normal Curve
- 7.0 Discrete Distributions
- 8.0 Tolerances
- 9.0 Determination of Sample Size
- 10.0 Process Capability Analysis
- 11.0 Pareto Analysis

“In earlier times they had no statistics. They did it with lies and we do it with statistics.”

Stephen Leacock

STATISTICS

1.0 CONTINUOUS DISTRIBUTIONS

Continuous distributions are formed because everything in the world that can be measured varies to some degree. Measurements are like snowflakes and fingerprints, no two are exactly alike. The degree of variation will depend on the precision of the measuring instrument used. The more precise the instrument, the more variation will be detected. A distribution, when displayed graphically, shows the variation with respect to a central value.

Everything that can be measured forms some type of distribution that contains the following characteristics:

Measures of central tendency:

- Arithmetic mean or average
- Median
- Mode

Measures of spread or dispersion from the center:

- Range
- Variance
- Standard deviation

Shapes of distributions:

- Symmetrical - normal
- Symmetrical - not normal
- Skewed right or left
- More than one peak

2.0 MEASURES OF CENTRAL TENDENCY

Measures of central tendency are values that represent the center of the distribution.

2.1 Arithmetic Mean or Average

The arithmetic mean or average of sample data is denoted by \bar{x} . The mean or average of an entire population or universe is denoted by μ . The value of \bar{x} may always be used as an estimate of μ .

$$\bar{x} = \frac{\sum x_i}{n} = \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}$$

The symbol Σ stands for “sum of.”

Five parts are measured and the following data are obtained:

2.6", 2.2", 2.4", 2.3", 2.5"

$$\bar{x} = \frac{\sum x_i}{n} = \frac{(2.6 + 2.2 + 2.4 + 2.3 + 2.5)}{5} = 2.4$$

2.2 Median

The median is the middle value of the data points.

To find the median, the data must be rank ordered in either ascending or descending order.

2.2, 2.3, 2.4, 2.5, 2.6
The Median is 2.4

For an even number of data points, the median is the average of the two middle points.

2.3 Mode

The mode is the value that occurs most frequently.

The data 2.6, 2.2, 2.4, 2.3, 2.5 do not contain a mode because no value occurs more than any other.

The following data are taken from another product:

6, 8, 13, 13, 20
The Mode is 13

3.0 MEASURES OF SPREAD OR DISPERSION FROM THE CENTER

How much can data points vary from a center or central value and still be considered reasonable variation? The question can be answered by calculating what is considered to be the natural spread of the data values.

3.1 Range

The calculation of the range provides a simple method of obtaining the spread or dispersion of a set of data. The range is the difference between the highest and lowest number in the set and is denoted by the letter r . The range and average are points plotted on control charts (a subject covered in a subsequent chapter). For the data set 2.6, 2.2, 2.4, 2.3, and 2.5, the high value is 2.6 and the low value is 2.2.

$$\text{Range} = r = (2.6 - 2.2) = .4$$

3.2 Variance

The variance is the mean squared deviation from the average in a set of data. It is used to determine the standard deviation, which is an indicator of the spread or dispersion of a data set.

$$\text{Variance} = \sigma^2 = \text{Sigma Squared} = \frac{\sum(x_i - \bar{x})^2}{n}$$

3.3 Standard deviation

The standard deviation is the square root of the variance. It is also known as the root-mean-square deviation because it is the square root of the mean of the squared deviations.

$$\text{Standard Deviation} = \sigma = \text{Sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

The average and standard deviation together can provide a great deal of information about a process or product. These two statistics are very powerful values used to make inferences about the entire population based on sample data.

When an inference is made about a population from sample data, $(n - 1)$ is used instead of n in the denominator of the variance formula. The term $(n - 1)$ is defined as degrees of freedom. When $(n - 1)$ is used, the calculated value is called the unbiased estimator of the true variance and is usually denoted by s^2 . When the standard deviation is obtained from the unbiased estimator of the variance it is denoted by s or σ' .

If a sample is taken and the average and standard deviation are not used to make inferences about the entire population, then the sample is considered to be the population and the standard deviation is indicated by σ . The symbol μ is used to denote the population average and \bar{x} is used to denote the sample average. The value of \bar{x} may always be used as an estimate of μ .

3.4 Variance and Standard Deviation Formulas

The following terminology and formulas will be used for the variance and associated standard deviation:

- Variance and standard deviation using all data values of a finite population:

$$\text{Variance} = \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sigma$$

- Variance and standard deviation using a subset (sample) of an infinite (very large) population:

$$\text{Variance} = s^2 = (\sigma')^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = s \text{ or } \sigma'$$

This is called the unbiased estimator of the population variance σ^2 .

- Variance and standard deviation using a subset (sample) of a finite population:

$$\text{Variance} = s^2 = (\sigma')^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \cdot \frac{(N - 1)}{N}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = s \text{ or } \sigma'$$

This is also called the unbiased estimate of the population variance σ^2 .

- The standard deviation for a distribution of averages is called the standard error.

$$\text{Standard Error} = s_{\bar{x}} = \sqrt{\frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{N_s}} \text{ or } \frac{s}{\sqrt{n}}$$

N_s is the number of samples and n is the sample size.

Example 1

Compute the variance and standard deviation for the data: 2.6", 2.2", 2.4", 2.3", 2.5". Assume that the data is the entire population.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \quad \text{where } \mu = 2.4$$

$(2.6 - 2.4)^2 =$	$(.2)^2 =$.04
$(2.2 - 2.4)^2 =$	$(-.2)^2 =$.04
$(2.4 - 2.4)^2 =$	$(0)^2 =$	0
$(2.3 - 2.4)^2 =$	$(-.1)^2 =$.01
$(2.5 - 2.4)^2 =$	$(.1)^2 =$.01
	Total =	.10

$$\text{Therefore } \sigma^2 = .10/5 = .02$$

The standard deviation is the square root of the variance. For this example, the standard deviation is

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = \sqrt{.02} = .1414$$

Many scientific hand calculators have a function to compute the mean, variance and standard deviation. The calculator is the preferred method of obtaining the values. The example is to ensure that you know what your calculator is doing when performing the calculations.

Another formula known as the working formula may also be used to calculate the variance and standard deviation. When the calculation for the variance and standard deviation must be done manually, the working formula may be easier than the formula given above. The answer will be the same using either formula. The working formula for the variance is

$$\sigma^2 = \frac{\sum(x_i)^2}{N} - \mu^2$$

$(x_i)^2$	
$(2.6)^2 =$	6.76
$(2.2)^2 =$	4.84
$(2.4)^2 =$	5.76
$(2.3)^2 =$	5.29
$(2.5)^2 =$	6.25
Total =	28.90

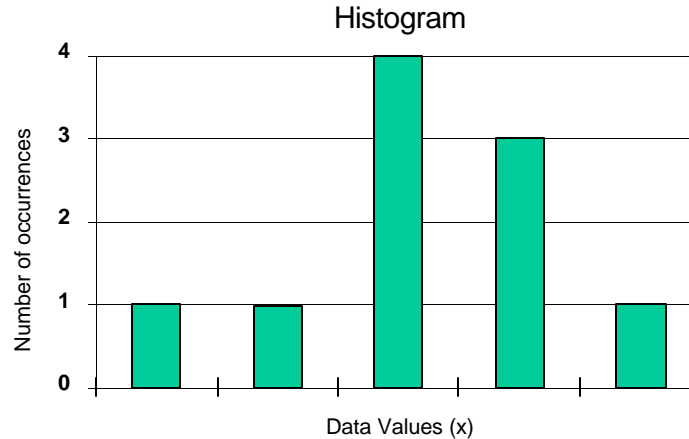
$$\text{Variance} = \sigma^2 = \frac{\sum(x_i)^2}{N} - \mu^2 = \frac{28.90}{5} - (2.4)^2 = 5.78 - 5.76 = .02$$

$$\text{Standard Deviation} = \sigma = \sqrt{.02} = .1414$$

4.0 HISTOGRAMS AND FREQUENCY DISTRIBUTIONS

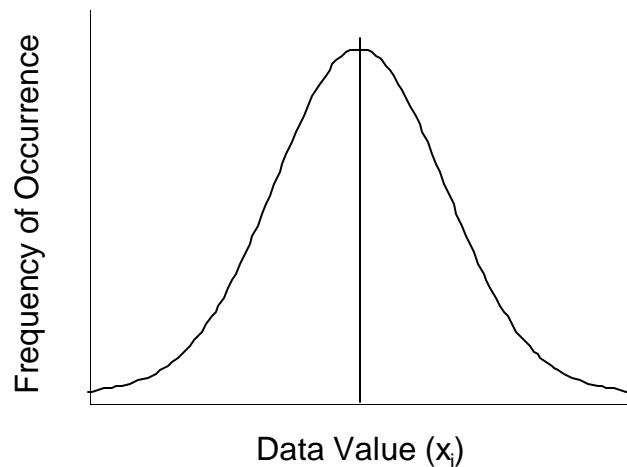
4.1 Histograms

A histogram is a simple frequency distribution. It is a plot of the actual data showing the data values versus the number of occurrences for each value. The plot will give a general indication of the shape of the distribution. It is a picture of a number of observations. The more data values that are plotted, the more informative it will be. As more observations are plotted, the histogram will approach the distribution of the population from which the data were obtained.



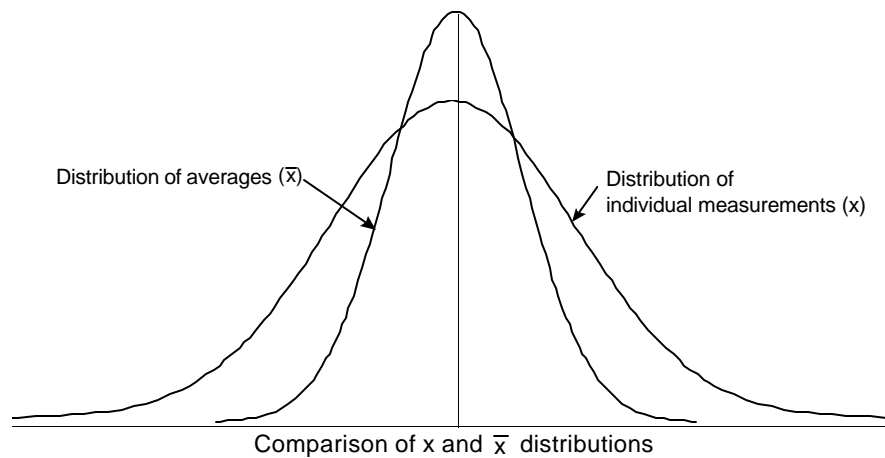
4.2 Frequency Distributions

A frequency distribution is a model that indicates how the entire population is distributed based on sample data. Since the entire population is rarely considered, sample data and frequency distributions are used to estimate the shape of the actual distribution. This estimate allows inferences to be made about the population from which the sample data were obtained. It is a representation of how data points are distributed. It shows whether the data are located in a central location, scattered randomly or located uniformly over the whole range. The graph of the frequency distribution will display the general variability and the symmetry of the data. The frequency distribution may be represented in the form of an equation and as a graph.



When using a frequency distribution, the interest is rarely in the particular set of data being investigated. In virtually all cases, the data are samples from a larger set or population. The population may be a specified number of items already produced or an infinite set of items that are continually made by some process. Sometimes, it is wrongfully assumed that data follow the pattern of a known distribution such as the normal. The data should be tested to determine if this is true. Goodness of Fit tests are used to compare sample data with known distributions. This topic will be covered in a subsequent chapter. The inferences made from a frequency distribution apply to the entire population.

Quality engineers and statisticians deal with distributions formed from individual measurements as well as distributions formed by sets of averages. Control charts, which are covered in a subsequent chapter, are applications of a distribution of averages. If the data are taken from the same population, there is a relationship between the distribution of individual measurements and the distribution of averages. The means will be equal ($\bar{x} = \bar{\bar{x}}$). If the standard deviation for individual measurements is s , then the standard error for the distribution of averages is s/\sqrt{n} . If a sample of 100 parts is divided into 20 subsets of 5 parts each, then n is 100 when calculating the variance and standard deviation of individual measurements and n is 5 when calculating the standard error using s/\sqrt{n} .



Some distributions have more than one point of concentration and are called multimodal. When multimodal distributions occur, it is likely that portions of the output were produced under different conditions. A distribution with a single point of concentration is called unimodal.

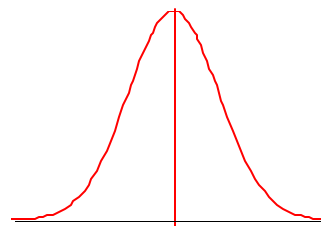
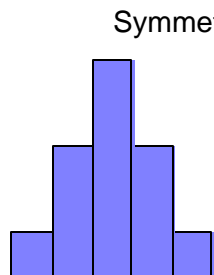
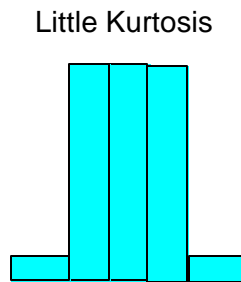
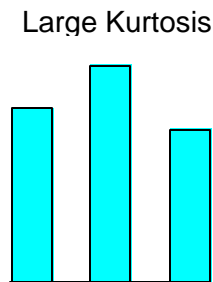
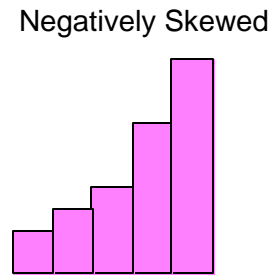
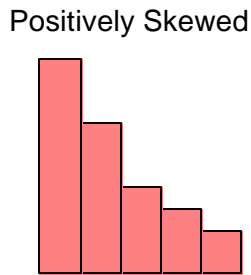
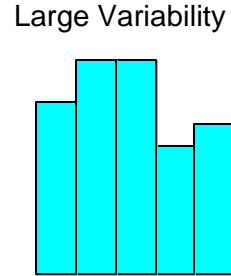
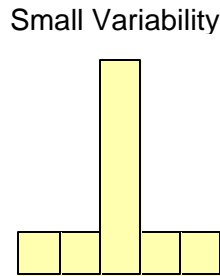
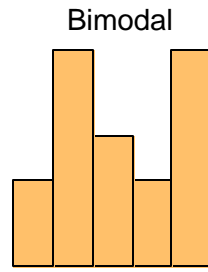
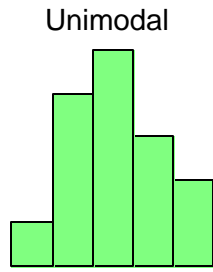
A distribution is symmetrical if the mean, median and mode are at the same location.

The symmetry of variation is indicated by skewness. If a distribution is asymmetrical it is considered to be skewed. The tail of a distribution indicates the type of skewness. If the tail goes to the right, the distribution is skewed to the right and is positively skewed. If the tail goes to the left, the distribution is skewed to the left and is negatively skewed. A symmetrical distribution has no skewness.

Kurtosis is defined as the state or quality of flatness or peakedness of a distribution. If a distribution has a relatively high concentration of data in the middle and out on the tails, but little in between, it has large kurtosis. If it is relatively flat in the middle and has thin tails, it has little kurtosis.

If the frequencies of occurrence of a frequency distribution are cumulated from the lower end to the higher end of a scale, a cumulative frequency distribution is formed.

5.0 SHAPES OF DISTRIBUTIONS



6.0 THE NORMAL CURVE

The normal curve is one of the most frequently occurring distributions in statistics. The pattern that most distributions form tend to approach the normal curve. It is sometimes referred to as the Gaussian curve named after Karl Friedrich Gauss (1777-1855) a German mathematician and astronomer. The normal curve is symmetrical about the average, but not all symmetrical curves are normal. For a distribution or curve to be normal, a certain proportion of the entire area must occur between specific values of the standard deviation. There are two ways that the normal curve may be represented: The actual normal curve and the standard normal curve.

6.1 Actual Normal

The curve represents the distribution of actual data. The actual data points (x_i) are represented on the abscissa (x-scale) and the number of occurrences are indicated on the ordinate (y-scale).

6.2 Standard Normal

The sample average and standard deviation are transformed to standard values with a mean of zero and a standard deviation of one. The area under the curve represents the probability of being between various values of the standard deviation. By transforming the actual measurements to standard values, one table is used for all measurement scales. A Standard Normal Curve table is included in appendix A and various iterations of the table can be found in most probability and statistics textbooks.

The abscissa on the actual normal curve is denoted by x and the abscissa on the standard normal curve is denoted by Z .

The relationship between x and Z :

$$Z = \frac{(x_i - \bar{x})}{s}$$

This is known as the transformation formula. It transforms the x value to its corresponding Z value. A distribution of averages may also be represented with the normal curve. The abscissa on the actual normal curve for a distribution of averages is denoted by \bar{x} . The center is denoted by $\bar{\bar{x}}$, the average of averages.

The relationship between \bar{x} and Z :

$$Z = \frac{(\bar{x}_i - \bar{\bar{x}})}{s/\sqrt{n}}$$

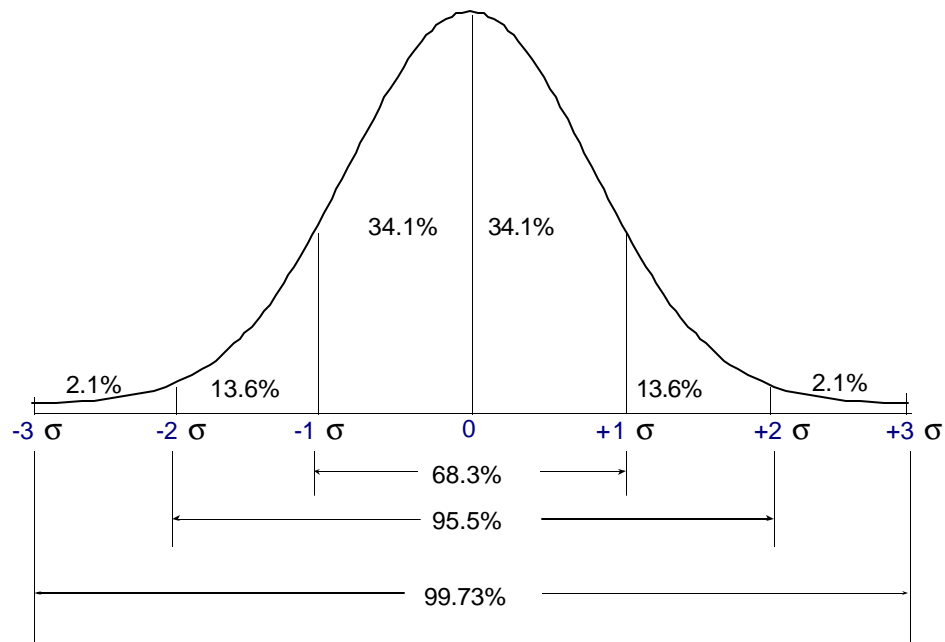
The statistic s/\sqrt{n} is the standard error or the standard deviation for a set of averages.

The statistic $\bar{\bar{x}}$ is an estimate of the parameter μ , the population average.

The standard normal curve areas are used to make certain forecasts and predictions about the population from which the data were taken. The standard normal curve areas are probability numbers. The area indicates the probability of being between two values

on the Z scale.

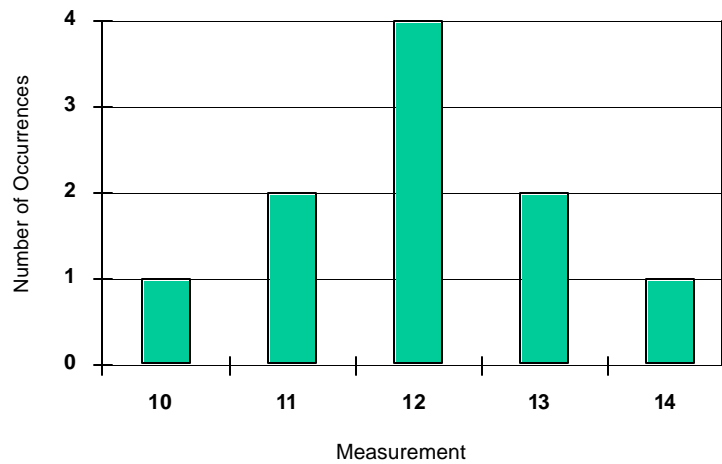
6.3 Areas Under the Standard Normal Curve

**Example 2**

The following data represent ten measurements (timing in seconds) from an electronic device. This is a sample taken from a production run.

10, 11, 11, 12, 12, 12, 12, 13, 13, 14

A histogram is drawn to get a general idea of the shape of the distribution.

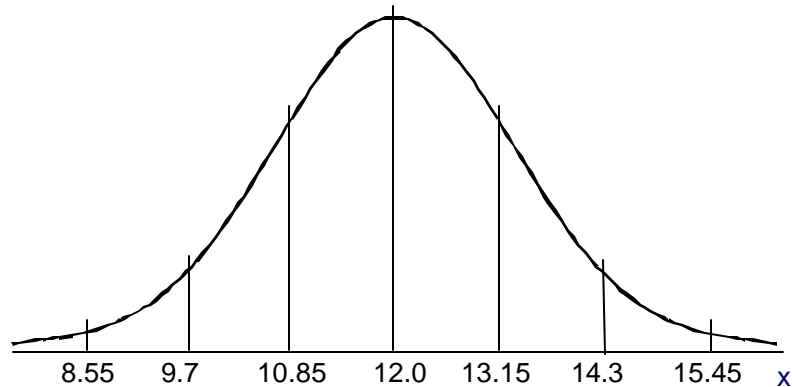


The mean and standard deviation are calculated: $\bar{x} = \frac{\sum(x_i)}{n} = \frac{120}{10} = 12$

The standard deviation from the unbiased estimator of the variance using the working formula: (Using the calculator is much easier.)

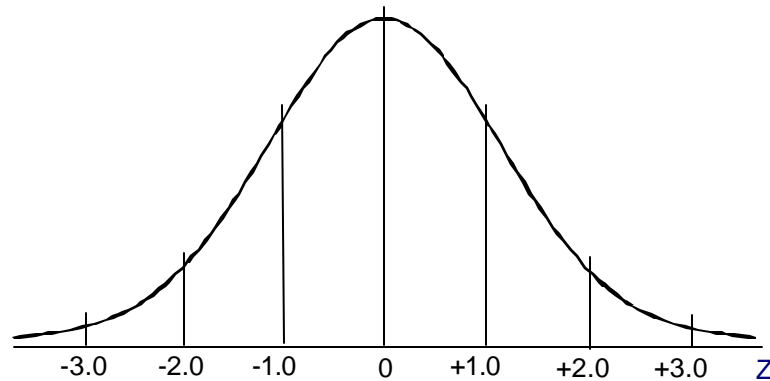
$$s = \sqrt{\left(\frac{\sum(x_i)^2}{n} - \bar{x}^2\right)\left(\frac{n}{n-1}\right)} = \sqrt{\left(\frac{1452}{10} - 144\right)\left(\frac{10}{9}\right)} = \sqrt{1.2\left(\frac{10}{9}\right)} = \sqrt{1.333} = 1.15$$

The normal curve areas are used to make predictions about the process.



To use the standard normal tables the x values must be converted to their equivalent Z values.

Using $Z = \frac{(x_i - \bar{x})}{s}$, the x value 10.85 converts to $Z = -1.0$, the x value 12 converts to $Z = 0$, the x value 13.15 converts to $Z = +1.0$, the x value 14.3 converts to the $Z = +2.0$, etc.



Area from $-\infty$ to $+\infty$	= 1.0
Area from $-\infty$ to 0	= .5
Area from 0 to $+\infty$	= .5

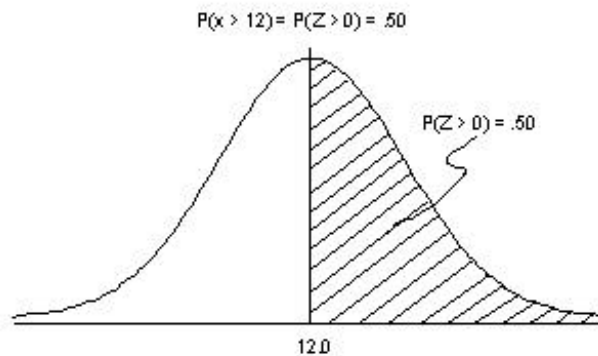
Example 3

Use the standard normal curve table to find the area between $Z = +1.0$ and $Z = +2.0$.

$$\begin{aligned} \text{Area from } 0 \text{ to } +2.0 &= .4772 \\ \text{Area from } 0 \text{ to } +1.0 &= .3413 \\ \text{Area between } +1.0 \text{ and } +2.0 &= .4772 - .3413 = .1359 \end{aligned}$$

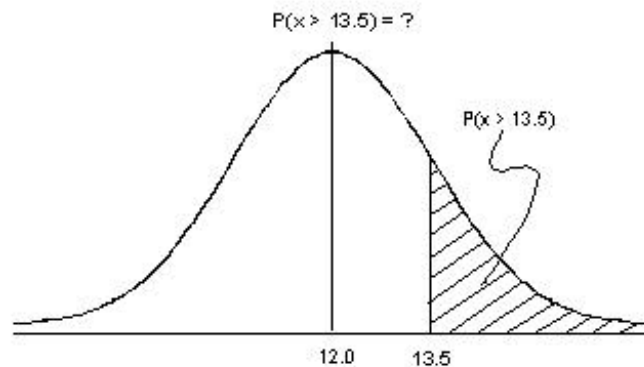
Example 4

For $\bar{x} = 12.0$ and $s = 1.15$, find the probability that a measurement will be greater than 12.0. This is written as $P(x > 12)$. $P(x > 12) = .50$ which is the same as the probability that $Z > 0$ since the mean value on the x scale corresponds to 0 on the Z scale.

**Example 5**

What is the probability that a part will have a measurement greater than 13.5?

The first step is to draw a diagram indicating the area that represents the probability of a measurement greater than 13.5. This is a very important step because the areas under the normal curve are difficult to visualize and a diagram makes it easy.

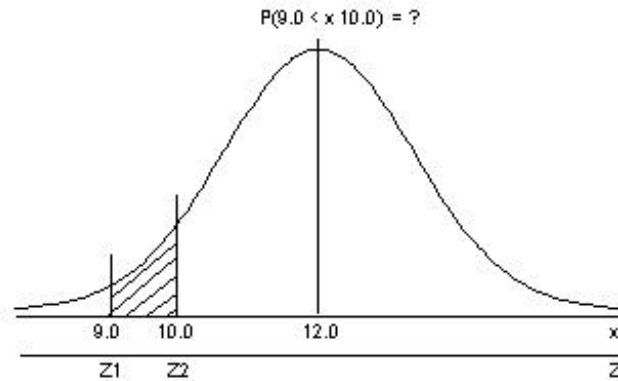


The next step is to convert the x value into a Z value. $Z = \frac{(x_i - \bar{x})}{s} = \frac{(13.5 - 12.0)}{1.15} = +1.30$
 This is the area from $Z = 0$ to $Z = +1.30$, therefore $P(x > 13.5) = P(Z > +1.30) =$

$$(.5000 - .4032) = .0968.$$

Example 6

What percentage of the population will have measurements between 9.0 and 10.0?



$$Z_1 = (9.0 - 12.0)/1.15 = -3.0/1.15 = -2.61$$

$$Z_2 = (10.0 - 12.0)/1.15 = -2.0/1.15 = -1.74$$

The standard normal curve table gives the following results:

Area from Z_1 to 0	= area from 9.0 to 12.0	= .4955
Area from Z_2 to 0	= area from 10.0 to 12.0	= .4591
Area from Z_1 to Z_2	= area from 9.0 to 10.0	= .4955 - .4591 = .0364

Therefore, 3.64% of the population will have measurements between 9.0 and 10.0.

7.0 DISCRETE DISTRIBUTIONS

There are many applications where the areas under the normal curve are used to approximate probabilities associated with discrete distributions. The mean and standard deviation are calculated using the formulas shown below. The procedures are the same as previously described for continuous distributions.

7.1 Hypergeometric Distribution

Mean and standard deviation for the hypergeometric distribution:

$$\text{In terms of } np: \quad \mu = np, \quad \sigma = \sqrt{npq \frac{(N-n)}{(N-1)}}$$

$$\text{In terms of } p: \quad \mu = p, \quad \sigma = \sqrt{\frac{pq}{n} \frac{(N-n)}{(N-1)}}$$

The parameter p is the fraction defective and $q = (1 - p)$ represents the fraction of good parts. To use the hypergeometric distribution formula the actual number of defective and goods parts in the lot must be known, not just the fraction defective.

7.2 Binomial Distribution

Mean and standard deviation for the binomial distribution:

$$\begin{array}{lll} \text{In terms of } np: & \mu = np, & \sigma = \sqrt{npq} \\ \text{In terms of } p: & \mu = p, & \sigma = \sqrt{\frac{pq}{n}} \end{array}$$

The parameter p is the fraction defective and $q = (1 - p)$ represents the fraction of good parts. The parameter p is also defined as the probability of a single success and must always be a value between zero and one.

7.3 Poisson Distribution

Mean and standard deviation for the Poisson distribution:

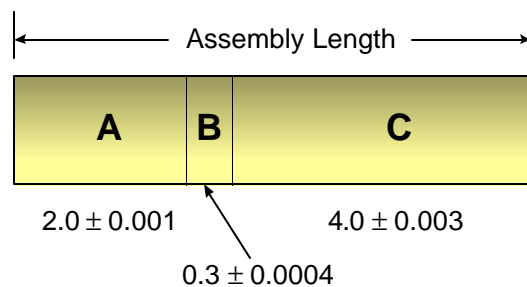
$$\begin{array}{lll} \text{In terms of } np: & \mu = np, & \sigma = \sqrt{np} \\ \text{In terms of } p: & \mu = p, & \sigma = \sqrt{\frac{p}{n}} \end{array}$$

The parameter p is either defects per unit or fraction defective. If p represents a fraction defective, it must be a value between zero and one. If p represents defects per unit, it is a value between zero and infinity. In terms of np , the mean is equal to the variance for the Poisson distribution.

8.0 TOLERANCES

Tolerances are usually specified in design drawings for interacting dimensions that mate or merge with other dimensions to obtain a final result.

A simple assembly is shown below:



8.1 Conventional Method of Computing Tolerances

Adding each individual tolerance in an assembly to form a final result is called the conventional method of computing tolerances.

Nominal value = nominal value_A + nominal value_B + nominal value_C
 Nominal value of the example assembly = 2.0 + 0.3 + 4.0 = 6.3

Addition of individual tolerances = T_A + T_B + T_C
 Tolerance of the example assembly = 0.001 + 0.0004 + 0.003 = 0.0044

The final value for the example assembly is 6.3 ± 0.0044.

Although this method is mathematically correct, the resulting tolerance may in some cases be quite large. Most mathematicians, statisticians, design engineers and quality engineers reject this method in favor of the statistical method shown below.

8.2 Statistical Method of Computing Tolerances

The nominal or center value is computed by adding the individual nominal values. This is the same computation for both the conventional and statistical methods.

Nominal value = nominal value_A + nominal value_B + nominal value_C
 Nominal value of the example assembly = 2.0 + 0.3 + 4.0 = 6.3

Statistical method for computing the tolerance = $\sqrt{T_A^2 + T_B^2 + T_C^2}$
 Tolerance of the example assembly = $\sqrt{(0.001)^2 + (0.0004)^2 + (0.003)^2} = 0.003187$

The final value is 6.3 ± 0.003187. Most of the assemblies will fall within this range.

9.0 DETERMINATION OF SAMPLE SIZE

9.1 Sample Size Determination for Variables Data

$$n = \left(\frac{Zs}{E} \right)^2$$

Z is the Z value corresponding to the level of confidence from the standard normal curve table. The symbol s is the standard deviation and E is the error factor. On the normal curve, E is the distance from the center (μ) to Z standard errors.

$$E = \left(\mu + Z \frac{s}{\sqrt{n}} \right) - \mu = Z \frac{s}{\sqrt{n}}$$

If the standard deviation is unknown, take thirty parts and calculate it using the standard deviation formula. Use this estimate for s in the above formula, and then recalculate s from the new sample size.

Example 7

What sample size is required so that there is a 90% chance that the sample mean will be within ± 0.2 inch of the true mean? The standard deviation is two.

From the standard normal curve table, Z is ± 1.645 for a 90% confidence level.
(E = ± 0.2)

$$n = \left(\frac{Zs}{E} \right)^2 = \left(\frac{(1.645)(2)}{0.2} \right)^2 = 271$$

9.2 Sample Size Determination for Discrete Data - Binomial

$$n = pq \left(\frac{Z}{E} \right)^2$$

The formula requires a value of p. When p is unknown, the worst case of p = .5 is used. This gives the largest value of pq (pq = .5 x .5 = .25).

Example 8

In conducting a public opinion poll, what sample size is required so that the poll takers are 95% confident that the poll is accurate to the nearest one percent?

From the standard normal curve table, Z is ± 1.96 for a 95% confidence level.
(E = ± 0.01)

$$p = (.5)(.5) \left(\frac{1.96}{.01} \right)^2 = 9604$$

9.3 Sample Size Determination for Discrete Data - Poisson

$$n = p \left(\frac{Z}{E} \right)^2$$

When used in the above formula, p represents defects per unit. If p is in terms of defective units, use the sample size formula for the binomial.

Example 9

In checking a characteristic on an assembly, what sample size is required so that there is a 99% confidence level that the average defects per unit recorded from the sample is within ± 0.1 of the true defects per unit in the population? Data from a random sample of one hundred parts yielded 0.5 defects per unit.

From the standard normal curve table, Z is ± 2.575 for a 99% confidence level.
(E = ± 0.1)

$$n = .5 \left(\frac{2.575}{.1} \right)^2 = 332$$

10.0 PROCESS CAPABILITY ANALYSIS

The term process capability refers to the normal behavior of product characteristic measurements when the process is in statistical control. It is the measured range of inherent variation of product characteristics turned out by the process. Process capability may be expressed by variables or attributes data. Process capability may also be defined as the range of values where 99.73% of the data values will fall. If a product characteristic yields an \bar{x} of 2.1" and an s of .01", the process capability is the range 2.07" to 2.13". A process capability study is a scientific procedure for determining the capability of a process to obtain the desired results.

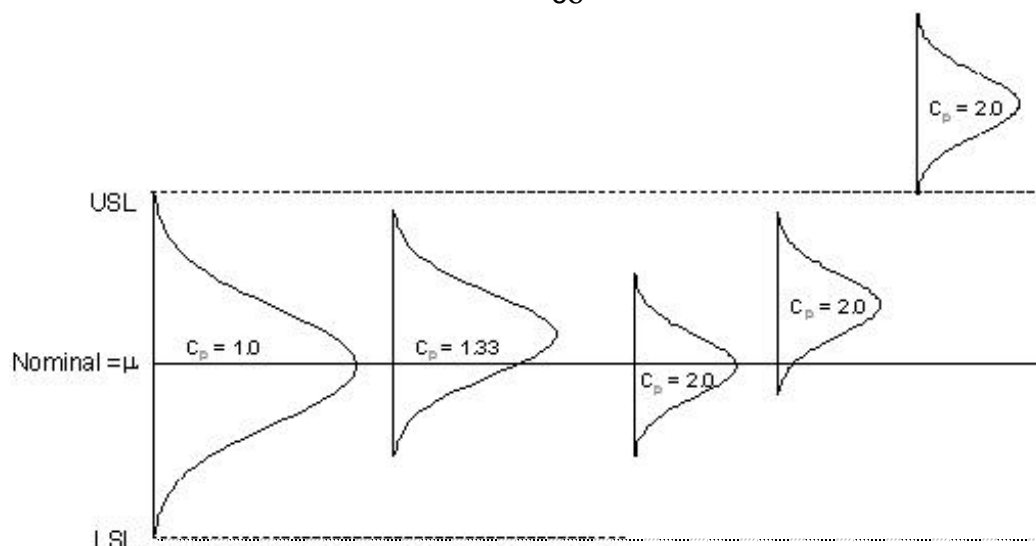
The standard deviation calculated from the sample data (s) is used as an estimate of the population standard deviation (σ).

$$\text{Process Capability} = 6 \text{ Sigma} = 6\sigma, \quad \text{where } \sigma = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

10.1 Process Capability Index = C_p

This is the ratio of the specification spread to the measured process variability or sample distribution (6σ). The sample distribution is an estimate of the population distribution because s^2 is the unbiased estimator of σ . The C_p does not indicate the location of the sample distribution relative to the specification. It is a comparison of the sample distribution width to the specification width. If the C_p is exactly 1.0, the 6σ spread is the same width as the distance between the specification limits. A C_p of 2.0 means that the 6σ spread is half of the specification range. A process with a C_p of 1 or greater may be within or totally outside of the specification limits. A C_p of less than 1 means that the sample distribution is wider than the specification range. (USL = upper specification limit and LSL = lower specification limit).

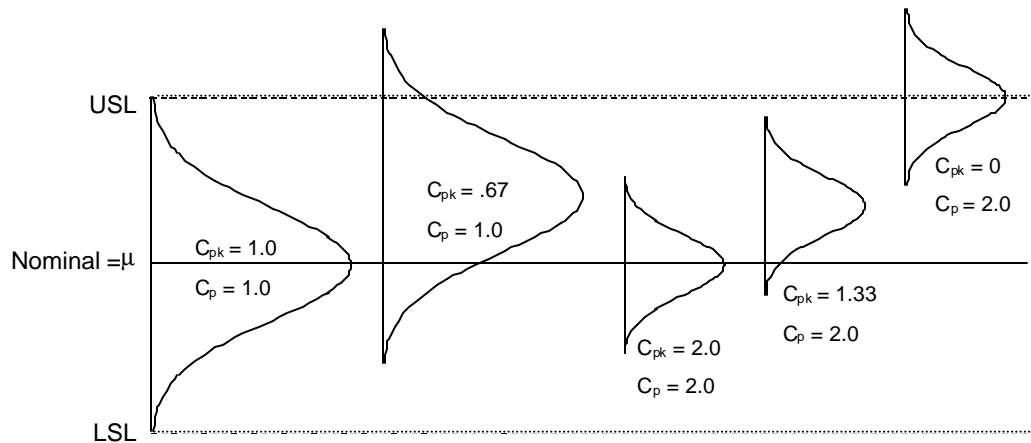
$$C_p = \frac{USL - LSL}{6\sigma}$$



10.2 Process Performance Index = C_{pk}

This index reflects the location of the sample distribution in relation to the specification midpoint. The maximum value of C_{pk} is equal to C_p and occurs when the sample distribution is centered on the specification midpoint or target. If the C_{pk} is 1.0 or less, there is no room for the process average to vary from the nominal dimension of the engineering specifications. A C_{pk} that is greater than one indicates that the 6σ spread is inside of the specification limits. A C_{pk} that is less than one indicates that some part of the distribution is outside of the specification limits. When the process average is located at one of the specification limits, C_{pk} is zero and 50% of the measurements will be outside of the limits. If the process average is outside of the specification limits, C_{pk} is a negative value. A C_{pk} of 1.3 to 2.0 is a respectable process performance index. To compute the C_{pk} , enter \bar{x} , LSL, USL and s into the formulas below. The lesser of the two values is the C_{pk} .

$$C_{pk} = \text{minimum} \left[\frac{\bar{x} - LSL}{3s} \text{ or } \frac{USL - \bar{x}}{3s} \right]$$

**Example 10**

The specifications for a certain product characteristic are $.005" \pm .0002"$. The control chart data ($n = 5$) indicate an \bar{x} of $.0051"$ and an average range of $.0001$. Calculate the C_p and C_{pk} for this characteristic. Is the process capability acceptable? What is the percent defective?

$$s = \frac{\bar{R}}{d_2} = \frac{.0001}{2.33} = .0000429$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{.0052 - .0048}{6(.0000429)} = \frac{.0004}{.0002574} = 1.55$$

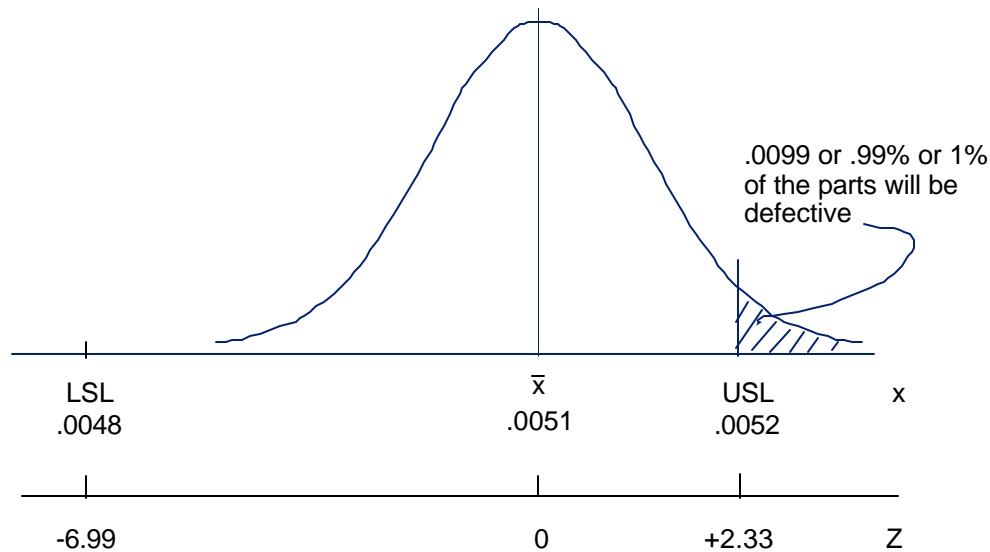
$$C_{pk}(1) = \frac{\bar{x} - LCL}{3\sigma} = \frac{.0051 - .0048}{3(.0000429)} = \frac{.0003}{.0001287} = 2.33$$

$$C_{pk}(2) = \frac{USL - \bar{x}}{3\sigma} = \frac{.0052 - .0051}{3(.0000429)} = \frac{.0001}{.0001287} = .77$$

$C_{pk}(2)$ is less than $C_{pk}(1)$, therefore $C_{pk} = C_{pk}(2) = .77$

Since the C_{pk} is less than one, a portion of the sample distribution will be outside of the specification limits. As shown below, the process will yield approximately one percent defective parts. One percent of the parts will be above the upper specification limit. This may or may not be an acceptable process capability. If the parts are expensive, the process capability may be unacceptable because of the high dollar value of one percent of the parts. If the parts are relatively cheap, the process capability may be acceptable.

$$Z = \frac{x_i - \bar{x}}{s} = \frac{.0052 - .0051}{.0000429} = 2.33$$



11.0 PARETO ANALYSIS

Vilfredo Pareto (1848 - 1923) was an Italian economist and sociologist whose theories influenced the development of Italian fascism. He was initially credited with the theory of maldistribution of wealth. This theory simply states that in any country a small percentage of the people own a large percentage of the money. The theory may really belong to M. O. Lorenz rather than Pareto. Since J. M. Juran identified the maldistribution of wealth and its similarities to defects in a manufacturing environment as the Pareto Principle in the first edition of his *Quality Control Handbook*, the term Pareto Principle been used.

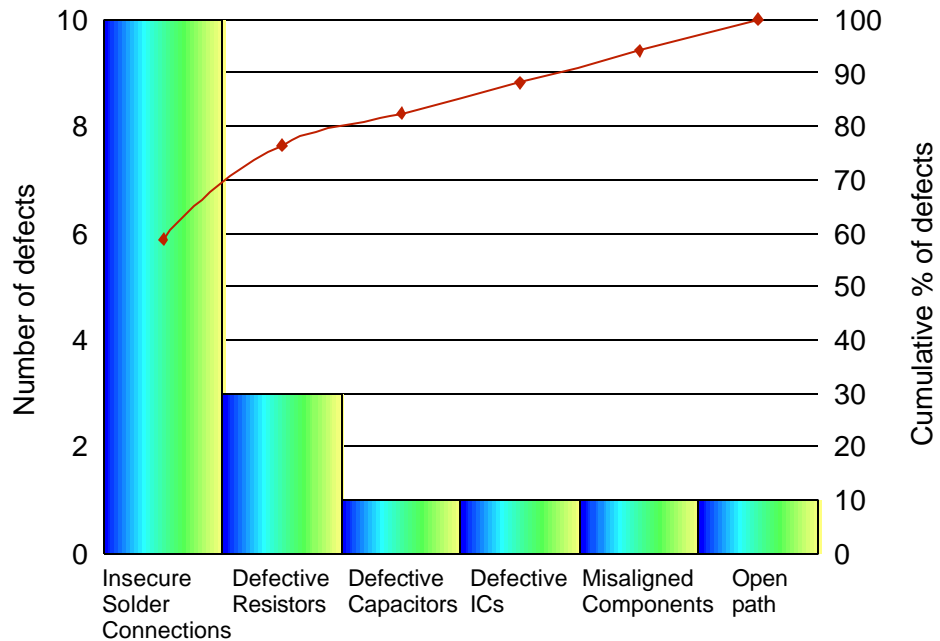
As in the maldistribution of wealth, it is also a fact that quality losses are maldistributed. A small percentage of the quality characteristics will account for a high percentage of the quality losses. The Pareto Principle is a simple yet powerful concept that provides a tool (Pareto diagram) for the analysis of data as well as information for action. Like all statistical tools, it does not provide the action itself.

A Pareto diagram indicates which problems should be worked on first in eliminating defects and improving the operation. The Pareto diagram is a way of portraying those problems that have

the greatest impact on the process or product, and once solved will yield the greatest return. A Pareto diagram is simply a bar chart arranged in order of importance.

Example 11

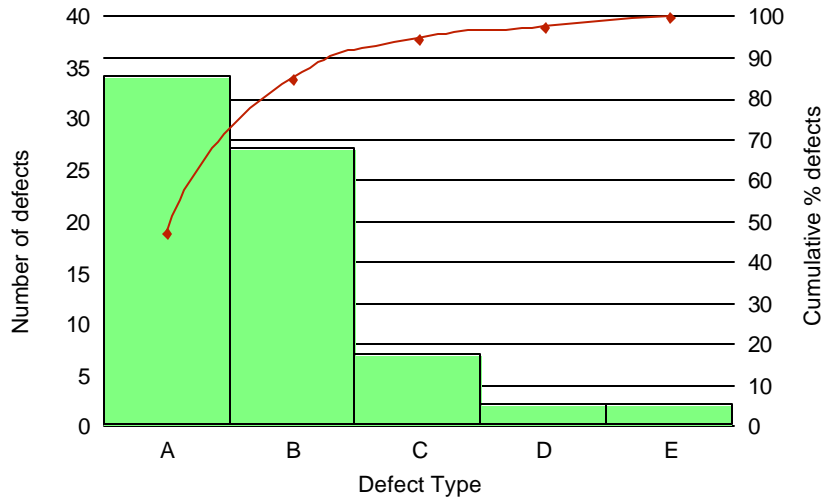
Defects recorded from a circuit board manufacturing operation



From this analysis, the first problem that may be pursued is the problem of insecure solder connections. This may not be obvious unless the frequencies of the various defects are plotted in some way. In most cases it is easier to see which defects are most important with a bar graph than by using a table of numbers. The diagram has two distinct parts: the “vital few” and the “trivial many.” Of course in an actual analysis a great many more defect types could occur.

Example 12 Simple analysis of defects

Defect Code	Number of Occurrences	Percent of Total
A	34	47.2
B	27	37.5
C	7	9.7
D	2	2.8
E	2	2.8
	72	100.0

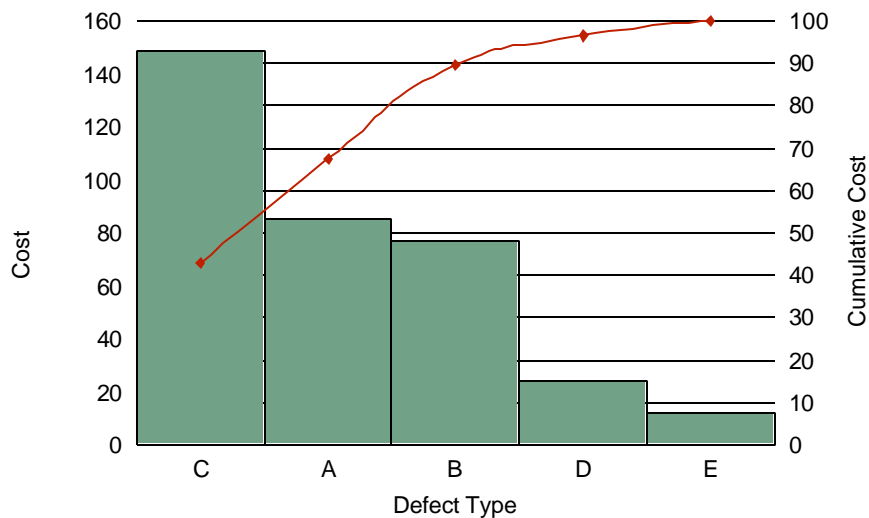


Defect A has the highest number of occurrences, but it may not have the greatest impact on the total operations. The key is to consider costs when making a Pareto analysis. Costs should always be taken into consideration. A separate study may have to be conducted to determine the costs of various defects.

Example 13 Pareto analysis considering costs

Defect Code	Number of Occurrences	Repair Costs*	Other Costs*	Total Costs	Percent of Total Costs
A	34	\$1.00	\$1.50	\$85.00	24.5
B	27	\$1.25	\$1.60	\$76.95	22.2
C	7	\$12.75	\$8.50	\$148.75	42.9
D	2	\$10.00	\$2.00	\$24.00	6.9
E	2	\$3.25	\$2.75	\$12.00	3.5
				\$346.7	100.0

*Incurred costs for each defect occurrence



From this diagram, it is evident that the root cause of defect C should be investigated first.

The elimination of this defect would reduce costs by 42.9%.

Pareto diagrams may be used to first identify major problems and then to display the impact of the improvement activity. The order of the bars will change if significant improvements to the process are made. The Pareto analysis itself will not actually solve the problem in question. A plan of attack must be devised after the problem is identified. The objective is to eliminate the root cause of the problem. Pareto charts and Pareto analyses are techniques to display data in a form that aids in the identification of the vital “few” and the “trivial many.”

When used alone, the Pareto analysis and associated diagram have several limitations. They should be used with good judgment and with knowledge of the process. If the samples are small, the diagram may not show much difference between the various classes of defects. It does not show variation over time for occurrences of a particular defect. A defect that occurred several times last month may not occur this month although no corrective action was taken. The Pareto diagram does not provide the trend of individual defects over time. In some rare cases, the diagram may show a new defect in the number one position each week although no corrective action was taken on the last number one defect. This is where knowledge of the process is important.

One way to make Pareto diagrams more effective is to use them together with trend charts for each specific defect class. The combination of Pareto diagrams and trend charts have many benefits. A particular defect class could be considered a significant problem if the Pareto diagram were used alone. A trend chart, however, may show that the high rate of occurrence of a particular defect last month was a one-time event. Trend charts show the effect of corrective actions.

Combining Pareto diagrams and trend charts provides a powerful analysis tool. More information is available than if they are used separately. This combination allows for the identification of critical problems and provides a method for determining the effectiveness of corrective actions.