PDHonline Course P207 (3 PDH)

Statistical Methods for Process Improvement - Part 1: Using Data for Process Improvement

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 I: Using Data for Process Improvement

An Introduction to Process Improvement and Statistics

Davis M. Woodruff, PE, CMC

Course Content

"There are three kinds of lies: lies, damned lies, and statistics."—Benjamin Disraeli.

Doing It Right the First Time

To prosper in today's economic climate, companies and their suppliers must be dedicated to never-ending improvement in quality and productivity. Leaders must constantly search for more efficient ways to produce products and services that meet customers' needs. Most organizations clearly state their dedication to achieve continual quality improvement in the company's Quality Policy.

Most everyone agrees that "Doing it right the first time" and the policy of prevention are sensible and even obvious philosophies which will, when adopted, improve the quality of all the products manufactured or services provided by any company.

Quantitative measures are needed to effectively monitor process and product or service performance. Statistical methods can provide the data and information to more effectively manage your business. The recent emphasis on Six Sigma and other similar "programs" is really nothing more than an organized approach to gathering, analyzing, managing and acting on data. The underlying principles are found in the basic concepts of statistics and Statistical Process Control or SPC and its associated problem solving techniques. These are simply the quantitative tools which will allow us to get a more objective handle on quality and productivity. These concepts have been taught under the banner of SPC, Statistical Methods and most recently the Six Sigma approach. All of these are based on simple statistical methods that can help manage process variability and thus improve quality.

This course is part 1 of a five part course, Statistical Methods for Process Improvement. The five parts are:

Part 1: Using Data for Process Improvement

Part 2: Using the Normal Distribution

Part 3: Understanding and Using Process Control Charts

Part 4: Using Data to Make Decisions

Part 5: Understanding Process Capability

The Definition

Statistics is not really math and it is not really science, but rather is "having to do with numbers." When we think of statistics many times it is about probability and game theory as opposed to practical applications in the world of work. Sometimes people just want to run away fast when they hear statistics. It is time to “get over it” and learn how to use these simple and powerful tools to help guide the journey of continual process improvement.

You may or may not have an idea of what SPC or Statistical Process Control is. In either case, let's see what it is not. It has already been said that it is not a quick fix for quality or productivity problems. Neither is it simply learning about control charting. Instead, it is a way of managing business operations to achieve continuing improvement in both process and people. Often SPC has been misunderstood to be simply using statistics to try to solve business problems or in the worst case, confusing people with statistical analysis methods that are theoretical instead of practical.

For this course we will define SPC as using statistical techniques to analyze data, then using the information to achieve predictability from a process. "It is a capital mistake to theorize before one has data,"—Sir Arthur Conan Doyle. In this course we will focus on the plan-do-check-act or PDCA approach to insure that we have a plan for gathering and using data that can make a difference.

SPC is a technique of monitoring and measuring performance using basic statistics such as averages, ranges or standard deviation along with the associated time sequenced charts and graphs. SPC methods can provide a picture that can be analyzed to detect an oncoming problem before the product of a process goes off specifications. Further, the data and charts can be analyzed to help identify the root cause of a problem--the first step toward problem elimination and prevention

The Beginnings

Statistical methods for process analysis or Statistical Process Control (SPC) is neither a new technique nor was it developed by the Japanese as many suppose nor was it developed by the recent Six Sigma “gurus.” To review the development of SPC requires that we discuss Dr. W. Edwards Deming, who could be considered the "father of SPC". The use of these techniques can be traced to the early 1900’s and the work of Dr. Walter A. Shewhart at the Bell Labs during and shortly after WWI. However, Dr. Deming has been credited with bringing these methods into industry.

Dr. Deming, a statistician by education, was associated with the Census Bureau, National Bureau of Standards and the Department of Agriculture early in his career. He became acquainted with Dr. Walter Shewhart of Bell Labs who introduced him to the concepts of process control.

During World War II, Dr. Deming taught statistics to thousands of individuals from hundreds of companies. The objective was to use this knowledge to bring better precision and more productivity to the nation's wartime manufacturers. A serious shortcoming of this effort was that it failed to involve top management but focused primarily on the engineering functions. Gradually, the techniques, though initially effective, disappeared due to lack of management support.

In 1945, Dr. Deming assisted Japan in certain agricultural areas. Later he was invited to teach statistical methods to industry. Through the efforts of Ichico Ishikawa, Japanese managers became not only familiar with, but committed to the concepts taught by Dr. Deming. Change was not immediate, but as you know, the Japanese have dramatically improved the quality of their products.

Dr. Deming has since been honored by the Japanese for his efforts. Their highest award for industrial excellence is known as the Deming Prize.

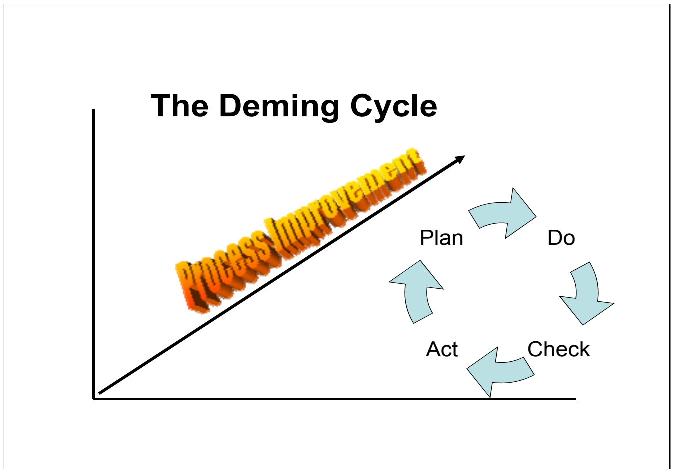
Today many managers still think that only minor changes in their organizations will solve productivity and quality problems. There are numerous productivity programs available, but many avoid the recognition of management's responsibility to change the systems they perpetuate.

Deming once asserted that 85 percent of all the problems of a company belong to the system which only management can change, and 15 percent of the problems can be solved by the workers. Before his death in 1994, he revised his estimates to 94% and 6%. If a company is to achieve significant gains in productivity, the management methods and styles of the organization must change.

An Organized Approach to Process Improvement

Process improvement or continual improvement must be addressed in a logical and organized manner. Statistical methods can help do this. Dr. Deming proposed that the

Plan-Do-Check-Act cycle is how organizations could implement and drive improvements in a logical manner. This is shown:



Expected Benefits

Successful implementation of using statistical methods (SPC) offers several benefits:

I. Increased customer satisfaction by producing a more trouble-free product.

2. Decreased scrap, rework and inspection costs by controlling the process.

3. Decreased operating costs by optimizing the frequency of process adjustments and changes.

4. Maximized productivity by identifying and eliminating the causes of out of control conditions.

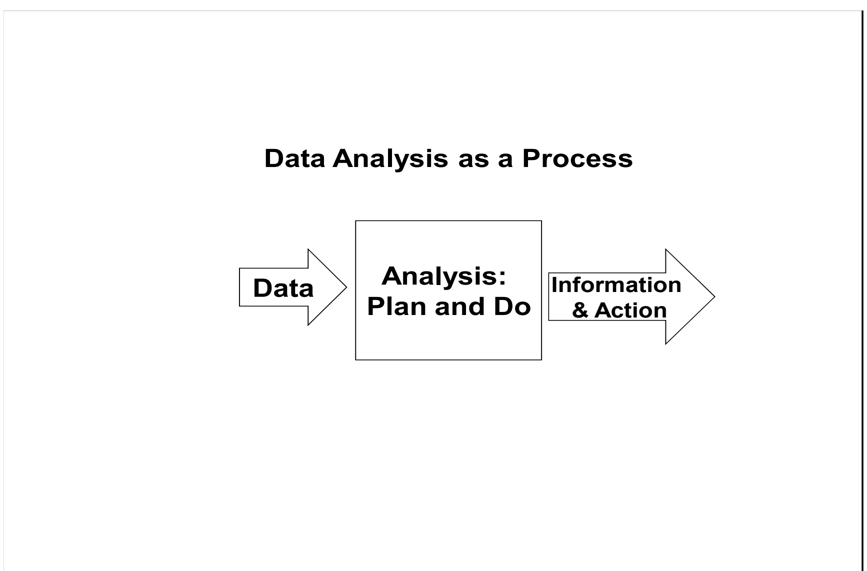
5. A predictable and consistent level of quality.

6. Elimination or reduction of receiving inspection by the customer.

The challenge is to use the statistical tools properly. The rewards are great. Only by everyone in the organization working together toward the same quality goal can your business ultimately be successful in the market place.

Statistical Methods as a Process

Many times organizations collect reams of data and do some sort of analysis or analyses without a planned approach. Data analysis is just another process to be managed. There are inputs, steps to take and outputs. However, this process requires an understanding of a number of underlying concepts which will be presented in this course. Let's look at the process first:



One can readily see that this follows the Plan-Do-Check-Act approach. Many times, the most important factor involves determining the question we are trying to answer by collecting and analyzing data. Remember, the output will only be as good as the input in this case. Bad data will yield bad or incorrect results or conclusions. **"No data is better than bad data,"** is one of the 31 lessons I use in my consulting practice. Also, data can be collected properly and the incorrect analysis technique can be used or underlying assumptions are not validated leading to incorrect conclusions.

This course is really about learning to manage this process correctly to lead to valid conclusions to improve process performance.

Variability is the Enemy of Process Improvement and Control

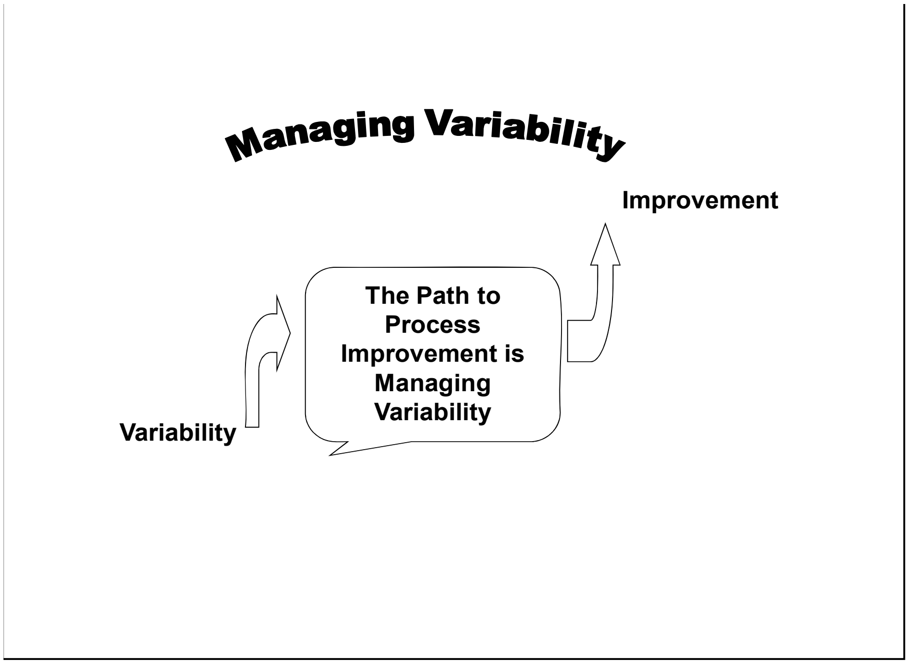
"Variability is simply the difference in things that ought to be alike." DMW

The Concept of Variability

We have all heard that no two snowflakes are alike; but neither are any two flowers or leaves or people. Even "identical" twins have different fingerprints (and probably different first names!). If your job involves producing a certain product, you know it is not possible to make each item exactly like the ones preceding it. The unit may be similar to the others already produced, but small differences can be detected.

This phenomenon of no two things being exactly the same is characterized by the term variability. Variability is the result of the many factors (sometimes un-measurable ones) which constantly influence a process. This is true whether the process is man-made, as in manufacturing, or natural, as in the formation of snowflakes or leaves.

In the case of snowflakes, temperature, wind speeds, humidity and air pollution can have a bearing on their formation. With man-made products, many factors can act simultaneously to affect a product parameter (dimension, purity, weight, etc.). Material composition, bearing wear, operator adjustments, speeds, pressures, tool wear or power fluctuations are just a few of the many factors which could influence the product. Also, feedstock fluctuations, changes in process conditions, and ambient conditions could have an effect on the product.



From observations of nature and manufacturing, we conclude that:

-No two objects are exactly alike.

-Variability is always present.

-Variability in processes is inevitable.

A few of the things which could cause variability in manufacturing are:

-Speeds

-Vibration

-Temperature

-Equipment adjustment

-Flow Rates

-Raw Materials

-Method Inconsistencies

-People

-Pressure (of assorted kinds!)

-Analytical problems

If the goal is to produce consistent, quality products and services, we must attempt to control the variation which results in the differences we observe. To control the variation requires that we trace it back to its source or cause.

Types of Variability

The causes of variability can be divided into those which are usually randomly occurring and those where a definite cause can be established. These are knows as Common Cause and Special Cause, respectively.

Common causes, sometimes called inherent causes, are those which randomly affect the system. These are usually many small and often un-measurable causes which add to the total variability.

Special causes, also called assignable causes, are usually of a greater magnitude. Normally, only a few special causes are acting on the system. These are usually characterized in the system as those things which come and go over time.

Any process can be influenced by both common causes and special causes. *One function of Statistical Process Control is to be able to distinguish between the two, and to provide a signal when special causes are present.*

Process Control can be achieved only when the special causes in the system can be identified and controlled. A summary of facts concerning common and special causes is shown below.

Common Causes	Special Causes
1. A large number are in effect at any one time.	1. Very few are in effect at any one time.
2. Each has an individual effect.	2. The effect is measurable.
3. Only a change in the system will reduce the part of the variability attributed to common causes.	3. They can be found and eliminated.
4. Only management has the ability to make changes affecting common cause variability.	4. The operator or supervisor is best able to discover and eliminate.
5. Common causes remain constant over time.	5. Special causes are sporadic and unpredictable.

Variability is the enemy of stable and predictable processes. Variability can be summed up as the "difference in things that ought to be alike." Remember, the path to process improvement is really the path of managing and reducing variability.

Understanding Statistical Process Control and Variability

Statistical Process Control can be summarized as follows:

In Control = Only common causes of variability present

Out of control = Special cause(s) of variability present

SPC allows us to separate the special causes from the common causes in a process. Once this is done, it is time to "swing into action" and identify, correct and eliminate the special causes. This enables actions to be taken before large amounts of nonconforming products are produced and is a part of the prevention cycle mentioned earlier.

When special causes are eliminated it is time to begin attacking the overall process variability and identifying steps to improve the system or common cause variation. This is what process capability and other such factors are all about.

Summary

- All processes have variability
- The control of quality is largely the control of variability
- Causes of variability are either common or special
- Special causes can be found and eliminated (no one said it would be easy!)
- To improve a process in statistical control requires changing the system to reduce or eliminate common causes

Data Display/Histograms

"A picture of the data is a good place to start the analysis process." DMW

Introduction

When collecting data, or a series of measurements, it is common practice to record successive readings in tabular form as shown below:

Tabulated Data

Subgroup	X_1	X_2	X_3	X_4	X_5
1	50	52	48	49	51
2	51	53	51	50	49
3	49	50	50	51	52
4	51	51	50	53	47
5	49	50	52	50	51
6	50	48	48	51	52
7	53	51	50	54	52
8	49	50	51	50	51
9	48	51	50	52	50
10	47	53	52	50	50

It is obvious from a table that the data is of limited use. Tables of data are difficult to analyze or interpret and fail to portray the distribution form which the samples were drawn. Tabulated data can be very useful if the order of samples is important, such as for a time plot analysis.

Frequency Tally

A frequency tally, or simple histogram, is another method of recording readings or measurements that allows easier interpretation of the data.

Frequency tallies simply record, or tally, the number of times a particular reading occurs. This tally allows one to see the shape of the distribution of the data. Further, the specifications or tolerances can be superimposed on the tally and you can readily see if samples fall outside these limits. The frequency tally actually provides a picture of the variation pattern.

Constructing a frequency tally is really very easy. The steps in developing a frequency tally are:

1. Determine the range of the data, or the expected range.
2. List the readings (or expected values) in a column.
3. Make a tally mark by the value each time it occurs. This is the frequency of occurrence for each value.
4. Total the frequency for each reading.

A frequency tally of the preceding table of data is shown below:

Frequency Tally

Readings	Frequency Tally	Frequency
47	//	2
48	////	4
49	////	5
50	//// //// ////	15
51	//// //// //	12
52	//// //	7
53	////	4
54	/	1
	Total (N)	**50**

The frequency tally allows for easier interpretation of the data. In fact, if the tolerances (specification limits) are known, e.g., 48 minimum and 53 maximum, you can readily see how many readings fall outside these limits.

The frequency tally is a simplified form of the histogram.

Histograms

A histogram is a special type of frequency tally in which the readings are grouped into classes and tallied accordingly. Each class has a midpoint and all readings within the class are recorded and graphed as thought they were actually on the midpoint. Bars are drawn for each class to make a bar chart of the data. Each bar is of equal width so that each class is proportional to the total histogram area.

If the number of class is too large or too small, the histogram yields a distorted picture of the data. The bar chart mentioned earlier presents the data in such a way that the interpretation is simplified even further.

The histogram gets its name from the fact that historical information is being displayed. It does not tell us what our process may do next, but it does tell us what it's done in the past. This helps us to get to know our process better, and getting to know the process is the first step towards gaining control of it.

The steps for constructing a histogram are as follows:

1. Determine the range of the data.
2. Determine the total number of readings.
3. Make a frequency tally of the data.
4. Determine the number of equal size classes to use, if necessary to further group the data. The number of classes is *estimated* as the square root of

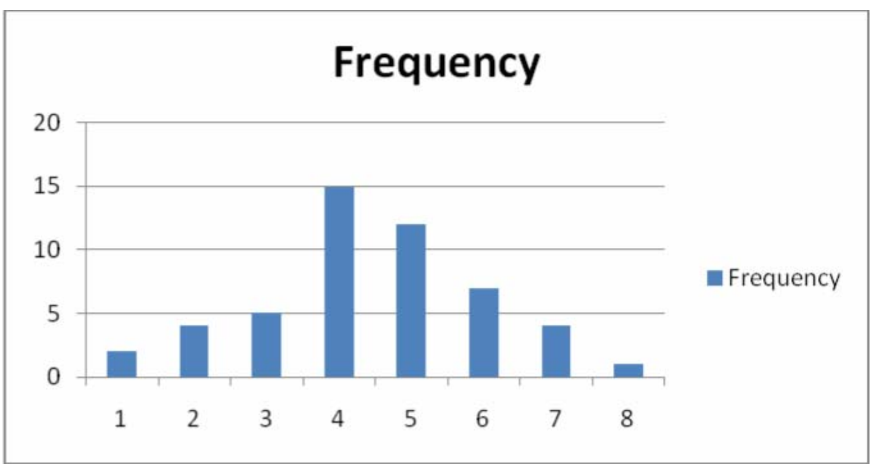
the number of data points ($\sqrt{N}$)

5. Group the data into the equal size classes.
6. Select a scale for the vertical axis of the bar chart. These are the frequencies for each class.
7. Construct the histogram.

If a data point falls exactly on the limit or boundary of a class, simply record the value in the next higher class each time and it will not affect the analysis. If you are not consistent in this, it will affect your analysis.

Histograms are usually tall in the middle and shallow towards the ends. A shape of this type is indicative of a bell-shaped distribution, commonly called a normal distribution. In fact, if we trace a smooth curve over the tops of each bar, we obtain the bell shape of the normal curve.

An example of a histogram is: (note, this is done in excel®)



Histogram

The following are the steps in constructing a histogram of the data from the preceding frequency tally.

1. The range of the data is from 47 to 54, or 7 units.
2. The total number of readings is 50.
3. From the √N, the number of classes, is 7 - 9.
4. Groups the data into equal size classes. In our case, there are 8 classes in the frequency tally.
5. Construct the histogram.

Histograms can be easily created using excel® or other statistical software. Histograms are useful in analyzing historical information about a process. However, we must remember that the histogram is a "snapshot" of the process at a particular time. It will not show trends, take time factors into account, or separate the two types of variability.

One of the real values of histograms is that they can be used to compare a process from one time period to another. Remember, it will not detect shifts that may occur in the data, it only records them after the fact.

Histograms can show how a process is centered within specifications or how a distribution approaches a target or specification.

Descriptive Statistics

1) Numbers are tools, not rules. 2) Numbers are symbols for things; the number and the thing are not the same. 3) Skill in manipulating numbers is a talent, not evidence of divine guidance.

Ashley-Perry Funny Quotes

Introduction

We have seen that data can be better represented by frequency distributions or histograms. Often these can be bulky, time consuming and sometimes misleading. What we need is a way to characterize the data in a more concise, more exact manner. Rather than a pictorial presentation, a number or numbers which adequately describe the data will be much more useful. Descriptive statistics sounds intimidating, but it really is nothing more than using numbers to describe other numbers or distributions of numbers. This can be accomplished by computing two general categories of measures: the measure of central tendency and the measure of variability.

Measures of Central Tendency

The basic measures of central tendency, or measures of location, are the mean, the median and the mode. In SPC the most commonly used measure of central tendency is the mean or arithmetic average.

Mean

The mean is simply the sum of the values divided by the number of observations. In terms of a formula,

$$\overline{X} = \sum_{x=1}^{x=n} \frac{xi}{N}$$

where,

$\overline{X}$ = the mean or average (called X-bar)

Xi= an individual value or observation

Σ = the sum of values or observations

n = the number of individual values or observations in the sample or subgroup

For example, suppose a department had been working overtime during the past five weeks. The weekly overtime hours were 87, 80, 92, 78 and 73. The mean overtime per week would be calculated

$$\overline{X} = \sum_{i=1}^{i=n} \frac{xi}{N} = (87+80+92+78+73) / 5 = 410 / 5 = 82 \text{ hours}$$

Calculating the mean is normally straightforward and not difficult. Again, excel® does automatically using the Sum function on the toolbar.

Some of the advantages of using the mean are:

- It is a very familiar concept
- Every data set has a mean
- Every data set has only one mean

- It can be calculated

There are some disadvantages though. A few are:

- It is affected by extreme values
- It can be tedious to calculate because all the data must be entered

Median

The next most often used measure of central tendency is the median. The median is a single value from the data set that measures the central item in the data set. This single item is the middlemost or most central item in the set of numbers. It simply divides the data into two parts with half the numbers above the median and half the numbers below it.

To find the median of a set of data, first arrange the numbers in ascending or descending order. If there is an odd number of data points, the median will be the middle value. If there is an even number of values, the median is the average of the two middle values.

In the previous example of overtime hours, the median of the 5 weeks (hours of 87, 80, 92, 78 and 73) is 80. If there had been 6 weeks with overtime hours of 87, 80, 92, 78, 73 and 84, the median would be the average of 80 and 84, or (80+84) / 2 = 82.

The median is designated by the symbol $\tilde{x}$

The advantages of the median are:

- Extreme values do not affect the median as they do the mean.
- It is easy to understand.
- It can be calculated from any kind of data.

The disadvantages associated with the median are:

- The data must be arrayed before the median can be found.
- For large sets of data, arraying is time consuming.
- Some statistical procedures which use the median are more complex than those which use the mean.

Mode

The mode is a measure of central tendency which, like the median, is not actually calculated but determined by observation. The mode is simply the value which occurs most often in the set of data. For example, in the following numbers, the value 23 occurs three times. Since this frequency is greater than any of the other values in the group, the mode is 23.

20	22	23	
21	22	23	Mode = 23
21	23	24	

The mode is the measure of central tendency used least often in industry.

The advantages of the mode as a measure of central tendency are:

- It is not affected by extreme values.
- It is easy to understand.
- It can be determined from any kind of data.

The disadvantages of the mode are:

- A set of data may not have a mode.
- The data set could have one, two or more modes.

Measures of Variability

A measure of central tendency can give us much information about a set of data, but it is not sufficient to adequately characterize a data set. For example, the set of five numbers below has the same mean as the set of three numbers. However, the smaller set has a much larger spread or more variability. This is sometimes referred to as the amount of dispersion.

$$\overline{X} = (110 + 125 + 130 + 135 + 150) / 5 = 130$$

$$\overline{X} = (30 + 100 + 260) / 3 = 130$$

To increase our understanding of the pattern of data, we must also measure its variability, which is really its spread or amount of dispersion.

In these courses, we will be discussing the two most commonly used measures of dispersion. These are the range and the standard deviation.

Range

The range is the simplest of the measures of dispersion - both in its concept and its method of calculation. The range is the difference between the highest and lowest observed values. It is denoted by the symbol R.

R = highest - lowest values observed

The range has some shortcomings. Since it is determined by only two values in the data set, it ignores the nature of the variation among all the other observations, and is greatly influenced by any extreme values.

The range for the set of five numbers on the previous page is:

R = 150 (highest) - 110 (lowest) = 40

Standard Deviation

In applied statistics, the standard deviation is the parameter primarily used as a measure of variability. It is basically the square root of the sum of the squared distances of the observations from the mean divided by the sample size. You can think of it as the average difference of the observed values from the mean of the data set. (Actually, it is the root-mean-square deviation rather than the arithmetic average.)

So far in this section, we have been discussing samples as opposed to populations. This is because very seldom, if ever, in our application of statistics in the workplace will we be working with populations. Samples are what we have available and are what we use in our analyses and projections. Because the calculations for the standard deviation are slightly different for populations and samples, we will present both in our discussion here.

For populations, the formula for standard deviation is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

Where σ = standard deviaion of the distribution

x_i = individual values

μ= average of the data set or distribution

N = number of values

This formula simply represents the square root of the sum of each individual reading minus the average of the data all divided by the total number of data points.

For samples, the formula is revised slightly to account for the reduced degrees of freedom because the average is a fixed point. The denominator becomes (N-1) instead of N. The formula is:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

As the number of data points (N) increases this "sample" standard deviation becomes essentially the same as the population standard deviation. For most practical applications, once you have more than about 30 data points the difference is insignificant. The purists statisticians may argue that point, however from a practical applications stand point, go with whichever you prefer when you have large amounts of data. Again, most statistics software packages will take care of this readily.

From the five numbers used for the range example,

$\overline{X}$ = (110 + 125 + 130 + 135 + 150) / 5 = 130

and

$\sigma = \sqrt{(110\text{-}130)^2 + \;\; + (130\text{-}130)^2 + (135\text{-}130)^2 + 150\text{-}130)^2 \,/\, 5} = 13.1$

In excel® simply use the statistical add-in and the function is SDEVP and the "sample standard deviation" would simply be the same numerator divided by <u>N-1, or 4</u>. Thus: **s = 14.58.** In excel® simply use the statistical add-in and the function is SDEV.

As you can see, considering a set of data a population or a sample will result in different standard deviations. It is important to recognize which of these, sample or population, any data set represents before calculating the standard deviation. When using a statistical calculator or software, you should know which standard deviation formula is being used.

Summary

Descriptive statistics characterize a distribution through the use of numbers.

Measures of Central Tendency are:

Mean ($\bar{X}$)

Median ($\tilde{X}$)

Mode (M)

Measures of Variability are:

Range (R)

Standard Deviation (σ or s)

Using Samples to Represent Populations

"By a small sample, we may judge of the whole piece." Miguel de Cervantes from Don Quixote

Introduction

Shoppers often sample a piece of cheese before purchasing any. They decide from one piece what the larger chunk will taste like. Farmers may submit a small portion of topsoil for analysis. From the results, they decide what minerals or nutrients should be added to their farmland to improve their crops. If the farmer tested all the topsoil, there would be none left for planting. Likewise, if the shopper tasted all the cheese, there would be none left to sell. This is unnecessary because by sampling a portion, they made a judgment about the whole. This is one of the objectives of SPC -- to examine a portion of data and decide, or draw conclusions, about the entire group of data.

Actually, the entire group is referred to as a population. You may also hear it referred to as a universe or parent distribution. The definition of population is "all members or elements of a group of items." For example, the population of parts produced by a machine includes all the parts the machine has ever made. For that matter, it will also include any parts the machine makes in the future. Outside of theoretical statistics, it is difficult to deal with populations. In industrial settings, it is almost impossible. Machines often operate 24 hours per day and produce thousands of parts. Data is just not available on every part. Even if it was, the population is changing from hour to hour as more parts are produced. Usually the best we can do in industry is to take a small portion of the population. This small portion of the population is called a sample. Two things hinder us from applying that which we have learned to actual manufacturing operations.

1. We must work with samples rather than populations.
2. The distributions found in industry may not be normal.

Mathematicians have overcome these two difficulties through the Central Limit Theorem. By applying this theorem, samples have a mathematical relationship to the populations from which they were taken. We can be assured of having a normal distribution with which to work even though our samples are a very small portion of the total population. This powerful theorem is the basis for applied statistics and, in particular, Shewhart control charts which are in part 3 of this five part course.

Central Limit Theorem

1. The mean of the means of several samples is the same as the mean of the population from which the samples are taken.
2. If the population is normally distributed, the samples taken from the population will be normally distributed regardless of sample size.
3. If the population is not normally distributed, the means of the sample groups taken from that population will be normally distributed for sample sizes of 30 or more and in many cases for sample sizes of 5 or more.

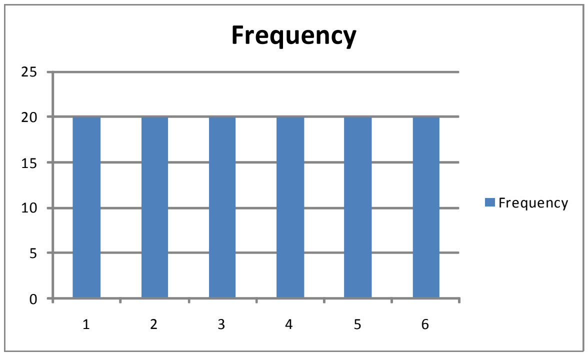
4. The standard deviation of the distribution of sample means is related to the standard deviation of the population by:
   Std. Dev. of sample means = Std. deviation of population/$\sqrt{\text{sample size}}$ or, mathematically:

$$\sigma_{\text{sample averages}} = \sigma_x/\sqrt{n}$$

This standard deviation of the sample means has a special name, the *standard error of the mean.*

The Central Limit Theorem holds true for sample sizeḍ 30, and in many cases is essentially valid for sample sizes of 5 or more. This means that subgroups of data can be taken and the averages will be normally distributed which allows us to use normal distribution statistics in most cases.

For an illustration of this concept, consider rolling dice in groups of five dice per roll. Each roll of five dice represents a sample from our process. Note that rolling a single dice (die) is a rectangular distribution bounded at 1 and 6. This means that in a120 rolls we would expect a distribution like this:



Now, consider the case of rolling the dice in subgroups of 5 for 25 rolls which would be 125 “measurements.” Average each roll of five carrying the result out to one decimal place. A normal distribution of averages becomes evident rather quickly! Try it on your own, if you need to “prove” the central limit theorem; otherwise, just accept it as proven many times over through the years!

The importance of the Central Limit Theorem in a manufacturing environment is that the critical data needed to understand and analyze the operation of any process can be obtained accurately from random samples of the population. It is not necessary, even if it were feasible, to inspect or measure the entire population.

Estimating Population Parameters from Sample Statistics

Having sampled a population, we can use the resulting sample statistics to determine the mean and estimate the standard deviation of the population. Suppose we have drawn 30 groups of 5 samples each. After calculating the mean of each group, we can average the sample means to get the overall average or $\bar{\bar{X}}_s$. The Central Limit Theorem states that the population mean or $\bar{X}_p$ or μ is the same as the $\bar{\bar{X}}_s$.

In the same manner, if the standard error of the mean, σ_x, is calculated for the sample means, it can be used to determine the standard deviation of the population. Again, the Central Limit Theorem says that $\sigma_{ave} = \sigma_{\tilde{x}}/\sqrt{n}$. By knowing the sample size n and σ_x, it is possible to determine σ_x.

Many times you will find that $\sigma_{\bar{x}}$ or the standard deviation of the sample averages was not determined. Instead, the range for each of the groups is calculated and an average range, $\bar{R}$, determined. The relationship between $\bar{R}$ for the sample groups and the population standard deviation has been established by mathematicians. The importance of this relationship is in the simplicity of the calculations. If you remember, Dr. Walter Shewhart was developing the basis for SPC in the early 1920's. At that time, computers and calculators as we know them had not been developed. As a result, the many calculations required were done by hand. By establishing the mathematical relationship between the range and the population standard deviation, the mathematician was able to predict the standard deviation of the population from a very simple calculation. The relationship developed is as follows:

$$\sigma_{ave} = \bar{R} / d_2$$

The d_2 value is a constant dependent upon sample size. It is shown below. For example, the value of d_2 for a sample size of 5 is 2.326. (Tables of these constants may be found in most statistics books, for example *Statistical Quality Control* by Grant & Leavenworth)

(n)	d_2
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.704
8	2.847
9	2.97
10	3.078

By using the Central Limit Theorem and applying the preceding relationships, we can determine characteristics of populations. This can be of major significance, because in industry the population itself is often unknown or even unknowable. Through the use of samples, the characteristics of the population are easily obtained.

Summary

- The mean of sample means is the same as the population mean.
- The dispersion of the sample distribution is related to the population dispersion by the sample size and variation
- Samples selected from a population will have normally distributed means.
- For normal populations, this is true regardless of sample size.
- For populations not normally distributed, it is true whenever sample sizes are 30 or more.

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. (In some situations it is the combination of operators, instruments and locations.)

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.*