Statistical Process Control

Jesus Cardenaz

April 12, 2023

1 Statistics Refresher

1.1 Basic Statistics and shit

1.1.1 Mean, Variance and Standard Deviation.

Mean The mean of a dataset is a way of capturing the 'center of mass' of a distribution. It is a way of calculating the expected value μ of a random variable X.

The expected value μ is a measure of the central tendency of a random variable, and it represents the *average* value that we would expect to observe if we repeated the random experiment many times. It is defined by the following equation:

$$\mu = E[X] = \sum_{x} P(X = x) \cdot x \tag{1}$$

Where x Represents values of the random variable X.

You go through all possible outcomes and you multiply the probability of that outcome times the value of the variable.

Variance The variance is a measure of how spread out the values of the random variable are from the expected value or the mean.

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = \sum_{x} P(X = x) \cdot (x - \mu)^2 \tag{2}$$

The idea is to look at the difference between each possible value and the mean, square that difference and ask for its expected value. This way, if the expected value is above or bellow the mean, it's still a positive number and the greater the difference, the bigger the number.

$E[(X - \mu)^2]$: This is the expectation of the squared difference between the random variable X and its expected value or the mean (μ). We square the difference to ensure that the value is always positive.

$\sum_x P(X = x) \cdot (x - \mu)^2$: This is the expanded form of the expected value. It involves summing up the products of the probabilities of each possible value of X and the squared difference between that value and the mean.

Standard Deviation The problem of the variance is that it's difficult to interpret this value as a 'distance from the mean' since is a squared value. So a more common way of interpreting spread is the Standard Deviation The Standard Deviation is another way of measure spread, it is the square root of the variance. It makes it easier to interpret the spread of a dataset.

It measures how much the data values deviate from the mean or the expected value of the data set. A high standard deviation indicates that the data values are spread out over a larger range, while a low standard deviation indicates that the data values are clustered around the mean.

Parameters vs statistics Parameters are numbers that describe the properties of entire populations. Statistics are numbers that describe the properties of samples. For example, the average income for the United States is a population parameter. Conversely, the average income for a sample drawn from the U.S. is a sample statistic. Both values represent the mean income, but one is a parameter vs a statistic.

Summary Value	**Parameter**	**Statistic**
Mean	μ or Mu	$\bar{x}$ or x-bar
Standard deviation	σ or Sigma	s
Correlation	ρ or rho	r
Proportion	P	$\hat{p}$ or p-hat

1.1.2 Probability Distributions

A probability distribution is the **mathematical function** that gives the probabilities of occurrence of different possible outcomes for a random variable. It describes the distribution of the values that a random variable can take and their probabilities of occurrence.

There are two main ways to represent the distribution of a set of data Frequency Distributions and Density Curves.

Frequency Distributions A frequency distribution describes the number of observations for each possible value of a variable. The frequency of a value is the number of times it occurs in a dataset. A frequency distribution is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Mainly there are two kinds of frequency distributions: Regular Frequency Distribution which tells us the number of values within certain given intervals. And Relative Frequency Distribution, which tells us the proportion of values within any given interval.

Density Curves A density curve is a **graph** that shows the probability of outcomes of a variable. It helps us visualize the overall shape of the distribution.

Properties of density curves: - A density curve mus line on or above the horizontal axis. - The total area of the curve must equal to 1.

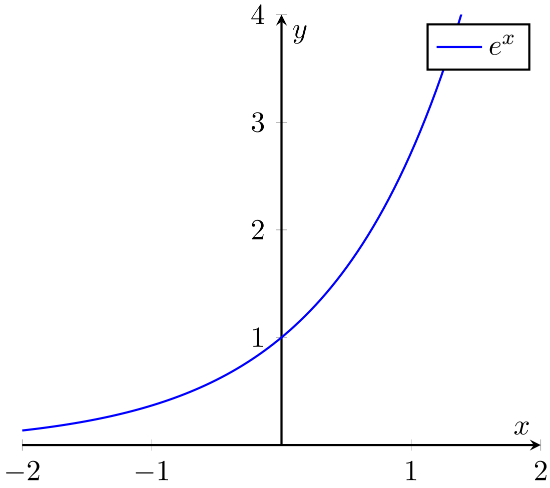
1.1.3 Normal Distribution

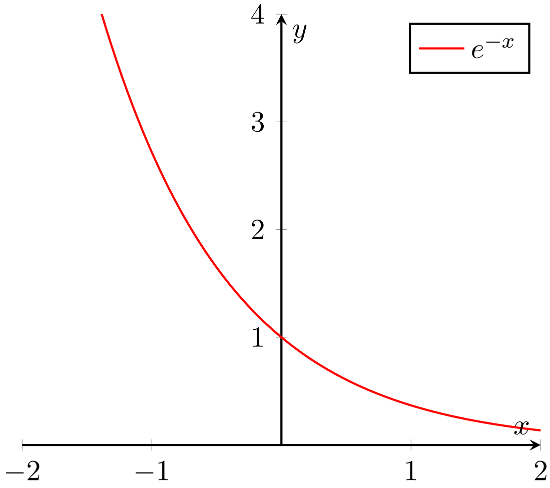
A normal distribution is a special type of density curve that is bell-shaped. In statistics, a **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{3}$$

Building up the formula To understand the formula, we're going to peel it off to the most basic form, and then build it up it step by step.

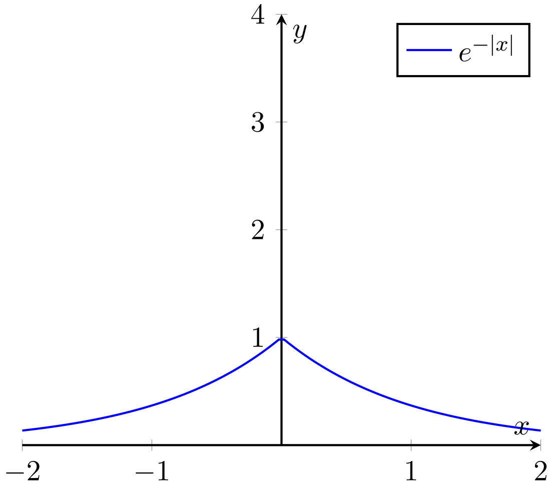
The function e^x or anything to the x describes exponential growth. Then, if we graph e^{-x} it'll describe exponential decay

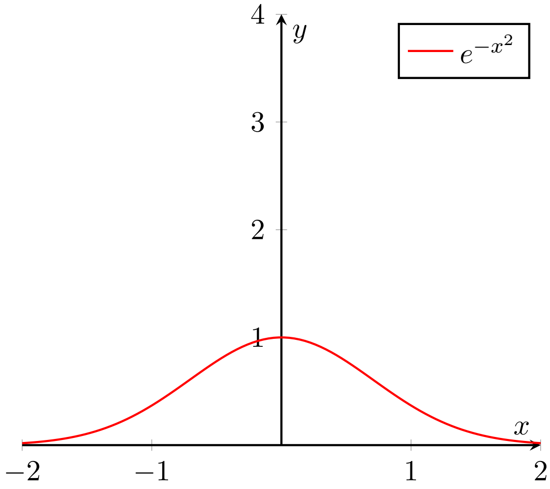




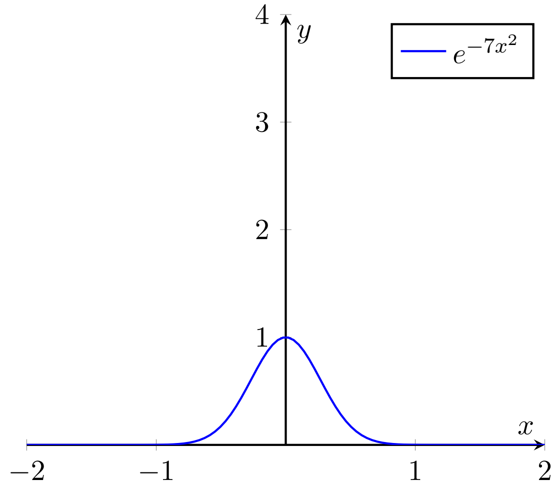
So if we want our graph to be symmetrical, we could do something like $e^{-|x|}$ But that would make a sharp point at $x = 0$, so a better solution would be to square the x value.

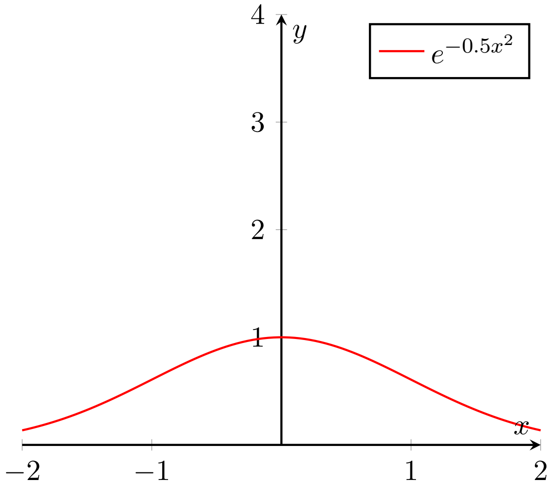
This already gave us the basic bell-shape we were looking for.





Now for the interesting part, if we trow a constant in front of the x, this allows us to stretch and squish the graph horizontally. Allowing us describe narrow and wider bell curves.





So if we want to describe the function by a value σ we would do something like: $e^{-\frac{1}{2}(x/\sigma)^2}$ But before we can define our function by the standard deviation, we want it to be a probability density function (PDF).

So we want this function to be a probability density function (PDF), that meaning that the total area under the curve of the function should be equal to one, this can be accomplished dividing by the square root of pi:

$$\frac{1}{\sqrt{\pi}}e^{-x^2}$$

But, we would like to describe our function by the standard deviation σ, so we also need to divide by that in order to the curve to still have an area of 1.

$$\frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$$

And this is already a valid normal distribution. Tweaking the value σ resulting in narrower or wider curves, would still mean that the area under the curve is one.

Standard Normal Distribution The case where the standard deviation is one, is called Standard Normal Distribution and is given by:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \tag{4}$$

Finally, if we subtract the mean μ we can characterize all possible normal distributions, and this gives us the final Standard Normal Distribution formula that we've seen.

$$\frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Properties The normal distribution is characterized by the following properties:

- The normal distribution is unimodal, which means that it has only one peak.
- The normal curve is symmetric about its mean.
- The parameters μ and σcompletely characterize the normal distribution.
- $X \sim N(\mu, \sigma)$

The notation $X \sim N(\mu, \sigma)$ represents a random variable X that follows a normal distribution with mean μ and standard deviation σ.

1.1.4 TODO: Central Limit Theorem

The central limit theorem states that the distribution of a random variable in a sample will begin to approach a normal distribution as the sample size becomes larger, regardless of the true shape of the distribution.

Consider $\frac{(X_1 + \ ... \ + X_N) - N \cdot \mu}{\sigma \cdot \sqrt{N}}$

1.2 Statistical models

A statistical model is a mathematical representation of observed data. When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically. ### Classification Models Classification is a process in which an algorithm is used to analyze an existing data set of known points.

1.3 TODO: Regression Models

Regression models are used to examine relationships between variables. Regression models are often used to determine which independent variables hold the most influence over dependent variables.

- Stepwise Regression.
- Ridge regression.
- Lassso regression.
- Elastic net regression.

1.4 Statistical Hypothesis

A statement about the nature of a population. It is often stated in terms of a population parameter. It is a formal claim about a state of nature structure within the framework of a statistical model.

The word hypothesis means a working statement. In statistics, we are interested in proving whether a working statement (the null hypothesis) is true or false.

In SPC, null-hypothesis = process in control, alter-hypothesis = process not in control.

1.4.1 Null-Hypothesis H_0

It states that the results are due to chance and are not significant in terms of supporting the idea of being investigated. The null hypothesis states that there is no relationship between the two variables being studied.

"The null hypothesis is a typical statistical theory which suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena."

In SPC, the null hypothesis usually states that the process is under control, and any observed variation is due to chance.

1.4.2 Alternative hypothesis. H_a

Represents a hypothesis of observations which are influenced by some non-random cause. The alternative hypothesis states that the independent variable did affect the dependent variable, and the results are significant in terms of supporting the theory being investigated. (Not due to chance.)

In SPC, proving that the alternative hypothesis is true means that the process is not under statistical control, and it's affected by some non-random variation (special variation).

1.4.3 Type I Errors

1.4.4 Type II Errors

1.4.5 P Values

The p-value in is the probability that the measured difference would occur due to random chance alone if the null hypothesis were true.

1.4.6 Significance Levels

TODO - Add Image: Directional Hypothesis In the Directional Hypothesis, the null hypothesis is rejected if the test score is too large or too small. Thus, the rejection region for such a test consist of one part, which is on the right side for a right-tailed test or the rejection region is on the left side from the center in the case of a left-tailed test.

TODO - Add Image: Non-Directional Hypothesis In a non-directional hypothesis test, the null hypothesis is rejected if the test score is either too small or too large. Thus, the rejection region for such test consist of two parts, one on the left and one on the right.

Why not accept the Null-Hypothesis? We assume that the null hypothesis is correct until we have enough evidence to suggest otherwise.

After you perform a hypothesis test, there are only two possible outcomes: - When your p-value is less than or equal to your significance level, you *reject the null hypothesis.* The data favors the alternative hypothesis. **Your results are statistically significant.**

- When your p-value is greater than your significance level, *you fail to reject the null hypothesis.* **Your results are not significant**

Failure to Reject the Null A lack of evidence isn't proof that something doesn't exist. You just haven't proven that it exists. It might exist, but your study missed it. That's a huge difference, and it is the reason for the convoluted wording.

Criminal Trials In a trial, we start with the assumption that the defendant is innocent until proven guilty. The prosecutor must work hard to exceed an evidentiary standard to obtain a guilty verdict. If the prosecutor does not meet that burden, it doesn't prove the defendant is innocent. Instead, there was insufficient evidence to conclude he is guilty.

Perhaps the prosecutor conducted a shoddy investigation and missed clues, or the defendant successfully covered his tracks. Consequentially the verdict in these cases is *Not Guilty.* That judgment doesn't say the defendant is proven innocent, just that there wasn't enough evidence to move the jury from the default assumption of innocence.

The hypothesis test assesses the evidence in your sample. If your test fails to detect an effect, that's not proof it doesn't exist. It just means your sample contained an insufficient amount of evidence to conclude that it exists. Like the prosecutor who missed clues, the effect might exist in the overall population but not in your particular sample. Consequently the test results *fails to reject the null hypothesis*, which is analogous to a 'not guilty' verdict in a trial. There was not enough evidence to move the hypothesis test from the default position that the null is true.

Accepting the null hypothesis would indicate that you've proven that an effect doesn't exist. You can't prove a negative. Failing to reject the null indicates that our sample did not provide sufficient evidence to conclude that the effect exists, that lack of evidence doesn't prove that the effect does not exist.

Example Suppose we wanted to check whether a coin was fair and balanced. A null hypothesis might say, that half flips will be of head and the other half will be tails. Whereas alternative hypothesis might say that flips of head and tail may be very different. For example if we flipped the coin 50 times, in which 40 heads an 10 tails results.

1.5 Statistical Measures

1.5.1 Z Score

Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.

The formula for the Z-score:

$$z = \frac{(x - \mu)}{\sigma} \tag{5}$$

The higher or lower a z-score is, the further away from the mean the point is.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A Standard Normal Distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation of 1.

- The Z-Score allows to calculate the probability of a score occurring within a standard normal distribution.

- Enables comparing two scores that are from different samples (which may have different means and standard deviations).

In SPC, the z-score is used to determine whether a data point is an outlier or falls within the expected range of values. A z-score can be converted into a p-value using a standard normal distribution table or calculator.

1.5.2 Critical Values.

In statistics, a critical value refer to specific values that are used to determine whether to reject or fail to reject a null hypothesis in a statistical test.

The critical value is the threshold value beyond which we reject the null hypothesis. It is based on the significance level (alpha level) of the test.

In SPC, the critical values are often used to define the control limits.

1.5.3 Significance Level α

The significance level defines how strong the sample evidence must be to conclude an effect exists in the population.

The significance level, also known as alpha or α, is an evidentiary standard that researchers set before the study.

1.5.4 P Values

The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

When you perform a statistical test, a p-value helps you determine the significance of your results in relation to the null hypothesis. It is a number describing how likely it is that your data would have occurred by random chance. (Null hypothesis is true). The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the P-value the stronger the evidence that you should reject the null hypothesis.

- A p-value less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis, as there is less that 5% probability the null is correct. (Results are random).
- A p-value higher than 0.05 is not statistically significant and indicates strong evidence for the null hypothesis.

The p-value calculates the likelihood of your test statistic. In other words, the p-value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by yous statistical test if the null hypothesis of that test was true.

Calculation of the P-value The calculation of the p-value depends on the statistical test you are using to test your hypothesis and the degrees of freedom of your test. No matter what test you use, the p-value always describes the same thing: *How often you can expect to see a test statistic as extreme or more extreme than the one calculated from your test*

1.5.5 Probability Density Function

It's a statistical measure used to gauge the likely outcome of a discrete value. PDFs are plotted on a graph typically resembling a bell curve, with the probability of the outcomes lying below the curve. A probability density function describes a probability distribution for a random, continuous variable. Its used to find the chances that the value of a variable will occur within a range of values that you specify. More specifically, A PDF is a function where its integral for an interval provides the probability (percentage of chance) of a value occurring in that said interval. A question that could be answered using PDF is: "What are the chances that the next IQ score that you measure will fall between 120 and 140?".

A PDF in statistics, *probability density* refers to the likelihood of a value occurring within an interval length of one unit.

A probability density function (PDF) explains which values are likely to appear in a data-generating process at any given time or for any given draw.

A cumulative distribution function (CDF) instead depicts how these marginal probabilities add up, Ultimately reaching 100% of possible outcomes.

1.5.6 Cumulative Distribution Function

CDF is used to calculate the area under the curve to the left from a point of interest. It is used to evaluate the accumulated probability.

Is a function that describes the probability that a continuous random variable X with a given probability distribution will be found at a value less than or equal to x.

1.5.7 Empirical Distribution Function

An empirical distribution function (EDF) is a cumulative distribution function (CDF) that is derived from the observed data in a sample. The EDF is used to estimate the true underlying distribution of the population from which the sample is drawn.

Suppose a given random sample of size n is $x_1, ..., x_n$ and let $x_1 < x_2 < ... < x_n$ be the order statistics; suppose further that de distribution of x if $F(x)$. The *empirical distribution function (EDF)* is $F_n(x)$ defined by:

$F_n(x) = \frac{n \leq x}{n}; -\infty < x < \infty$

More precisely, the definition is

$$F_n(x) = \begin{cases} 0 \text{ if } x < x_1 \\ \frac{1}{n} \text{ if } x_1 \leq x < x_2 \\ \frac{2}{n} \text{ if } x_2 \leq x < x_3 \\ \vdots \\ \frac{n-1}{n} \text{ if } x = x_(n-1) \leq x < x_n \\ 1 \text{ if } x \geq x_n \end{cases} \tag{6}$$

Thus $F_n(x)$ is a step function, calculated from the data; as x increases it takes a step up of height $\frac{1}{n}$ as each sample observation is reached.

1.5.8 Kurtosis

Kurtosis is a statistical measure that describes the “tailedness” of the probability distribution of a random variable. It measures the extremity of deviations or outliers, not the configuration of data near the mean. Distribution with higher kurtosis has more extreme values in the tails of the distribution, while a distribution with lower kurtosis has a more uniform distribution of values.

1.5.9 Degrees of Freedom.

1.5.10 Confidence Intervals.

1.6 Statistical Tests

Is an assumption about a population which may or may not be true. Hypothesis testing is a set of formal procedures used by statisticians to either accept or reject statistical hypotheses.

1.6.1 T-Test

A t-test is a statistical method used to determine whether two groups of data have different means.

A t-test looks at the average of each group and how much variation there is within each group, then compares these numbers to calculate a t-value. If the t-value is large, it means that the difference between the groups is likely not due to chance, and we can conclude that there’s a significant difference between them.

Here’s good video explaing this.

A T-test:

- Assumes the null hypothesis is true and then evaluates whether is that is a bad assumtion.
- The p-value tells us the probability that the two datasets would have differed due to random chance under the assumption of the null hypothesis.
- If $p < 0.05$, you can reject the null hypothesis. This corresponds to a 95 % confidence interval
- A T-tes can be paired (you tested the same group twice) or unpaired (you tested two different groups).

Depending on the t-test, the test can determine whether:

- One mean is different from a hypothesized value
- Two group means are different
- Paired means are different.

One-Sample T-test. Use a one-sample t-test to compare your sample mean to a hypothesized value for the population and to create a confidence interval of likely values for the population mean.

In example, you use the one-sample t-test when one group is compared against a standard value, like the acidity of a liquitd to a neutral pH of 7.

The formula for a one sample t-test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7}$$

Where:

- μ_9 = The hypothesized mean.
- $\bar{x}$ = Is the sample mean.
- s = The sample standard deviation.
- n = The sample size.

Assumptions: You have a random sample - Your data must be continuous - Your sample data should follow a normal distribution or have more than 20 observations.

Two-Sample T-test. You use the two-sample t-test when the two groups you're studying come from two different populations.

The formula for the two-sample t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{8}$$

Where:

- t = The t value.
- x_1 and x_2 are the means of the two groups being compared.
- s_p = The pooled standard error of the two groups.

Paired t-tests Use paired t-tests to assess dependent samples, which are two measurements on the same population, an example would be measuring before and after an experimental treatment.

TODO

Assumptions: Dependent Samples Unlike two-sample t-test, paired t-test use the same people or items in both groups. One way to determine whether a paired t-test is appropriate for your data is if each row in the dataset corresponds to one person or item.

1.6.2 Z-Test

A z-test is a statistical hypothesis test that is used to determine whether a sample mean is significantly different from a population mean, when the population standard deviation is known.

The z-test is based on the standard normal distribution. The test statistic is calculated by subtracting the population mean from the sample mean, and dividing by the standard error of the mean.

One-Sample Z test.

- H_0 : The population mean equals a hypothesized vale $\mu = \mu_0$.
- H_A : The population mean DOES NOT equal a hypothesized value $\mu \neq \mu_0$

When the p-value is less or equal to your significance level, reject the null hypothesis.

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{9}$$

Two-Sample Z test

- H_0 : Two population means are equal $\mu_1 = \mu_2$.
- H_A : Two popularion means are not equal $\mu_1 \neq \mu_2$

Again, if the p-value is less than or equal to your significance level, reject the null hypothesis.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}} \tag{10}$$

Z-Test vs T-test Z-test require you to know the population standard deviation, while t-test use a sample estimate of the standard deviation. In practice, analysts rarely use the Z-test because it's rare that they'll know the population standard deviation. It's even rarer that they'll know it and yet need to assess an unknown population mean.

When to use each? When you know the population standard deviation, use a Z-test. When you have a sample estimate of the standard deviation, the best practice is to use a t-test, regardless of the sample size.

1.6.3 Normality Test and Goodness of Fit

In statistics normality testes are used to determine if a data set is well-modeled by a normal distribution and compute how likely it is for a random variable underlying the data set to be normally distributed.

Goodness of Fit is a statistical Hypothesis used to see how closely observed data mirrors expected data. Goodness-of-Fit test can help determine if a sample follows a normal distribution, if categorical values are related, or if random samples are from the same distribution.

In assessing whether a given distribution is suited to a data-set, the following tests and their underlying measures of fit can be used: - Bayesian information criterion - Kolmogorov–Smirnov test - Cramér–von Mises criterion - Anderson–Darling test - Shapiro–Wilk test - Chi-squared test - Akaike information criterion - Hosmer–Lemeshow test - Kuiper's test - Kernelized Stein discrepancy - Zhang's ZK, ZC and ZA tests - Moran test - Density Based Empirical Likelihood Ratio tests

1.6.4 Chi-Square Test

A chi-square test is a statistical hypothesis test that is used to compare observed and expected results. It is used to determine whether there is a significant association between two categorical variables.

It is typically used to test for independence between two variables, which means that there's no relationship between them

The test is based on the chi square statistic, which is calculated by comparing the observed frequencies of the data with the expected frequencies under the null hypothesis.

If the difference between the observed and expected frequencies is large enough, the test will reject the null hypothesis

1.6.5 Kolmogorov-Smirnov test (K-S Test | KS Test)

Its a test for normality. Its a non-parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. Or to compare two samples. The test is usually recommended for large samples over 2000. For smaller samples, use Shapiro-Wilk

The test compares your data with a known distribution and lets you know if they have the same distribution. Although the test is non-parametric it doesn't assume any particular underlying distribution. It is commonly used as a test for normality to see if your data is normally distributed. It's also used to check the assumption of normality in Analysis of Variance.

1.6.6 Shapiro-Wilk Test

The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution. The test gives you a W value; small values indicate your sample is *not* normally distributed. The formula for the W value is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\left(\sum_{i=1}^{n} (x_i - \overline{x})\right)^2} \tag{11}$$

Where:

- X_i are the oredered random sample values
- a_i are constants generated from the covariances, variances and means of the sample (size n) from a normally distributed sample.

The test has limitations, most importantly that it has a bias by sample size. The larger the sample, the more likely you'll get a statistically significant result.

1.6.7 Anderson-Darling

The Anderson-Darling test is a statistical test used to determine whether a given data set is drawn from a specific probability distribution, such as the normal distribution or exponential distribution. It is a goodness-of-fit test that is based on the distance between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the theoretical distribution being tested. The test calculates a test statistic, which is a weighted sum of the squared differences between the EDF and the CDF, with greater weight given to the tails of the distribution.

The Anderson-Darling returns an statistic called *Anderson-Darling statistic (AD)*: The Anderson-Darling statistic is the *test statistic*. It's like the t-value for the t-test or the F-value for the F-test. Typically you don't interpret this statistic directly, but the software uses it to calculate the p-value for the test.

The hypotheses for the Anderson-Darling test are:

- H_0: The data follow a specified distribution.
- H_A: The data do not follow a specified distribution.

The formula for the Anderson-Darling

1.7 Analysis of Variance

I've decided that Analysis of Variance should have its own chapter.

1.8 Statistical Plots

1.8.1 Stem-and-Leaf Plot

1.8.2 Normal Probability Plot (Q-Q Plot)

The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

Usually, the Normal Probability Plot its accompanied by a Normality Test, often the Anderson-Darling or the Shapiro-Wilk tests
To determine whether the data is normal, compare the p-value to the significance level α. H_0 : The data follows is normally distributed H_A : The data does not follow a normal distribution

- P-value $\leq \alpha$: The data do not follow the normal distribution (reject H_0) If the p-value is less than or equal to the significance level, the decision is to reject the null hypothesis and conclude that your data does not follow a normal distribution.
- P-value $> \alpha$: Cannot conclude the data do not follow the distribution (Fail to reject H_0) If the p-value is larger than the significance level, the decision is to fail to reject the null hypothesis because you do not have enough evidence to conclude that your data do not follow the distribution. However, you cannot conclude that the data do follow the distribution.

1.8.3 P-P Plot

1.8.4 Kernel density estimation

1.9 Correlation and Association

1.9.1 Pearson Correlation Coefficient

1.9.2 Spearman Rank Correlation Coefficient

1.9.3 Covariance

1.10 Time Series Analysis

1.10.1 Moving Averages

1.10.2 Exponential Smoothing

2 Dr. W. Edwards Deming

William Edwards Deming (1900-1993) was an American statistician, engineer, and management consultant who is widely regarded as one of the leading thinkers in the field of quality control and management. Deming is best known for his work in Japan after World War II, where he helped to revolutionize Japanese manufacturing and transform it into a world leader in quality and productivity.

Recommended material: - Out of the Crisis

2.1 14 Points for Management

2.1.1 1. Create constancy of purpose:

Organizations should have a long-term focus and strive to achieve their core mission.

Examples: - Commit to developing your team on an ongoing basis. - Create a vision of the future. Make sure evey employee knows this vision. - Ask for process improvements and product quality improvements every week from your team

2.1.2 2. Adopt the new philosophy:

Organizations must shift from a focus on short-term profits to a focus on continuous improvement and customer satisfaction. You must no longer tolerate commonly accepted levels of mistakes, people who don't know what they're doing, defects and inadequate supervision. At the heart of this new philosophy should be the desire to put your customer's needs first.

Examples: - If you run a restaurant, is it acceptable for a customer to order only to be later told that what they ordered is out of stock? - Is it acceptable to wait over fifteen minutes to speak to a customer suppoert team member if you run a bank?

2.1.3 3. Cease dependence on inspection:

Quality should be built into the product or service, rather than relying on inspection to catch defects. Stop depending on inspections to improve quality and build quality into your processes instead. An inspection doesn't improve quality because it happens too late. At the point of inspection, the quality of the product already exists, good or bad, so all an inspection does is find an existing lack of quality.

2.1.4 4. End the practice of awarding business on price alone:

Suppliers should be chosen based on their ability to provide high-quality products or services.

2.1.5 5. Improve constantly and forever the system of production and service:

Organizations should continually strive to improve their processes, products, and services.

2.1.6 6. Institute training:

Employees should be trained in the skills and knowledge necessary to perform their jobs effectively.

2.1.7 7. Institute leadership:

Management should provide clear direction and guidance, and lead by example.

2.1.8 8. Drive out fear:

Employees should be encouraged to speak up and provide feedback without fear of retribution.

2.1.9 9. Break down barriers between departments:

Departments should work together and communicate effectively to achieve common goals.

2.1.10 10. Eliminate slogans, exhortations, and targets for the workforce:

Management should focus on providing the resources and support necessary for employees to perform their jobs effectively.

2.1.11 11. Eliminate numerical quotas for the workforce and numerical goals for management:

Performance should be evaluated based on overall improvement, rather than meeting specific targets.

2.1.12 12. Remove barriers to pride of workmanship:

Employees should take pride in their work and be given the opportunity to contribute to the success of the organization.

2.1.13 13. Institute a vigorous program of education and self-improvement:

Employees should be given opportunities for personal and professional development.

2.1.14 14. Put everybody in the company to work to accomplish the transformation:

All employees should be involved in the process of continuous improvement and strive to achieve the organization's goals

2.2 Five Deadly Diseases

2.2.1 Lack of constancy of purpose:

Organizations without a clear and consistent mission, vision, and set of values can become easily distracted and lose focus.

2.2.2 Emphasis on short-term profits:

A focus on short-term profits can lead to cutting corners, sacrificing quality, and neglecting long-term planning and investment.

2.2.3 Evaluation by performance, merit rating, or annual review:

Traditional performance evaluation systems that rely on numerical ratings or annual reviews can be arbitrary and demotivating, leading to a focus on meeting targets rather than improving performance.

2.2.4 Mobility of management:

Frequent turnover or reshuffling of management can disrupt established processes, relationships, and communication channels, leading to inefficiency and confusion.

2.2.5 Running a company on visible figures alone:

Relying solely on quantitative data can lead to a narrow focus on easily measurable outcomes, ignoring the importance of qualitative factors such as customer satisfaction, employee morale, and innovation.

2.3 Common Causes and Specials Causes of Improvement.

The central problem in management and in leadership, is failure to understand the information in variation.

A distribution only presents accumulated history of performance of a process, nothing about its capability. **A process only has a capability if it is stable**

There are two kind of mistakes when it comes to variaton: 1. Ascribe a variation or mistake to a special cause when in fact the cause belongs to the system. 2. Ascribe a variation or a mistake to the system when in fact the cause was special.

Overadjustment is a common example of mistake No. 1. Never doing anything to try to find a special cause is a common exmaple of mistake No. 2.

Supervisors commonly make the mistake of overadjustment when they direct to the attention of one of their people any mistake or defect, without first ascerting that the worker was actually responsible for the mistake. Did the worker make the mistake, or was the system responsible for it?

Difference between conformance to specifications and statistical process control. - The aim in production should not just to get statistical control, but to shrink variaton. Costs go down as variation is reduced. It is not enough to meet specifications.

2.4 Important concepts

System: A system is a network of interdependent components that work together to achieve a common goal or purpose. In his view, a system includes not only the physical components of a process, but also the people, procedures, and culture that make up the organization.

Stable System: A stable system is one where the process is in control and variation is due to common causes. In a stable system, the variation can be predicted and controlled using statistical process control (SPC) techniques.

Common Causes: Common causes of variation are inherent in the system and cannot be eliminated. They are predictable and can be managed through process improvement and statistical analysis.

Special Causes: Special causes of variation are due to factors outside the system, such as machine malfunctions, operator errors, or changes in the environment. These causes are unpredictable and require investigation and corrective action to eliminate.

Control Charts: Control charts are a tool for monitoring and controlling a process. They allow you to distinguish between common and special causes of variation, and to determine when the process is out of control.

Action on Common Causes: When common causes of variation are identified, they can be addressed through process improvement efforts, such as training, standardization, or redesign of the process.

Action on Special Causes: When special causes of variation are identified, they require immediate corrective action to eliminate the cause and prevent future occurrences.

3 Statistical Process Control

3.1 Rational Sub-grouping

3.2 Control Charts

3.2.1 I-MR Charts

3.2.2 X-Bar/R Chart

3.2.3 Other Charts

TODO. This may be wrong.

3.3 Capability Analysis

Different Standard deviations used. The calculation and definition of the different standard deviations in this sections is based upon what Minitab has in its documentation about Capability Analysis.

Definitions

Symbol	Meaning
T	= target
LSL	= lower specification limit
USL	= upper specification limit
μ	= process mean
tol	= sigma tolerance
m	= midpoint of LSL and USL
$\bar{x}$	= estimate of process mean
σ_{Within}	= within subgroup process standard deviation
$\hat{\sigma}_{Within}$	= estimate of within subgroup process standard deviation
n	= total number of observations
n_i	= number of observations in subgroup i
$C_4(n_i)$	= unbiasing constant for subgroups of size n_i (for use with sample standard deviations)
$d_2(n_i)$	= unbiasing constant for subgroups of size n_i (for use with sample standard deviations)
$\sigma_{Overall}$	= overall process standard deviation
$\hat{\sigma}_{Overall}$	= estimate of the overall process standard deviation
$P(X)$	= probability of event X
Z	= standard normal variable
x_i	= the ith observation
x_{ij}	= the jth obersvation of the ith subgroup

3.3.1 Within Standard Deviation σ_{within}

It's an estimate of the variation within the subgroups. If your data is collected properly, the within-subgroup variation should not be influenced by changes to the process inputs. such as tool wear or different lots of material. In that case, the within standard deviation represents the natural and inherit variation of the process over a short period of time. **It indicates the potential variation of the process if shifts and drifts between groups were eliminated** Within-subgroup standard deviation is also used when the sample size = 1, the formulas for sample size > 1 and sample size = 1 are different.

Subgroup Size = 1 When the subgroup size = 1, you can estimate the σ_{within} using one of the following methods:

Method 1: Average of Moving Range

$$\sigma_{\bar{x}} = \frac{\bar{R}}{d_2(w)} \tag{12}$$

Where:

- $\bar{R}$ = The average moving range is the average value of the moving range of two or more consecutive points.
- d_2 = An unbiasing constant read from a table
- w = The number of observations used in the moving range.

Now, the calculation of the moving range (MR) is not straightforward, it depends in the window size, in order to calculate the Moving Range with a window size w you have the equation:

$$MR_i = |\text{Max}[x_i, ..., x_{i-w+1}] - \text{Min}[X_i, ..., X_{i-w+1}]| \text{ for } i = w, \; ..., \; n \tag{13}$$

This equation represents the calculation of the range of a sliding window over a sequence of numbers. The sequence of numbers is represented by the vector $X = [x_1, x_2, ..., x_n]$. The window size is represented by the parameter w, which specifies the number of elements to include in each window. The equation calculates the difference between the maximum and minimum values within each sliding window of size w, and stores the resulting range values in a new vector MR

A more simpler form of the equation is presentented when the value for w is 2, which is usually the most common case for the Moving Range.

$$MR_i = |x_i - x_{i-1}| \text{ for } \; i = 2, 3, ..., n \tag{14}$$

Then, calculate the average of the moving range:

$$\overline{MR} = \frac{R_w + ... + R_n}{n - w + 1} \tag{15}$$

This equation calculates the average range over all windows of size w in a sequence of numbers. It uses the range vector R calculated by the previous equation, where R_i represents the range of the i-the window of size w.

Method 2: Median of Moving Range

$$\sigma_{\bar{x}} = \frac{\widetilde{MR}}{d_4(w)} \tag{16}$$

Where:

- $\widetilde{MR}$ = Median of the Moving Range.
- w = The number of observations used in the moving range. (AKA. The window size. Default is 2)

The median moving range is the median value of the moving range of two or more consecutive points. Use this method when the data have extreme ranges that influence average of the moving ranges

Subgroup Size $>$ 1 To estimate σ_{within} for subgroups of size bigger than one. You could use one of three different methods:

Method 1: Pooled Standard Deviation A pooled standard deviation is just a weighted average of the variances from two or more groups of data when they are assumed to come from populations with a common variance.

The formula to obtain the unbiased estimator of σ_{Within} is given by:

$$\sigma_{Within} = \frac{S_P}{C_4(d+1)} \tag{17}$$

Where the S_P is the Spooled Standard Deviation, and is given by:

$$S_P = \sqrt{\frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sum_i (n_i - 1)}} \tag{18}$$

Where:

- d = Degrees of freedom for $S_p = \sum(n_i - 1)$.
- x_{ij} = j^{th} observation in the i^{th} group.
- $C_4(d+1)$ = Unbiasing constant.
- Γ = Gamma function.

Method 2: Average of Subgroup Ranges

Method 3: Average of Subgroup Standard Deviations

3.3.2 Overall Standard Deviation $\sigma_{overall}$

Is the standard deviation of all the measurements and is an estimate of the overall variation of the process. If your data are collected properly, the overall standard deviation captures all sources of systemic variation. In that case, it represents the *actual* variation of the process that the customer experiences over time. **It is used to calculate the PP and PPK values, and other measures of the overall capability of the process**

Unbiased estimator of $\sigma_{overall}$

$$\sigma_{overall} = \frac{S}{C_4(N)} \tag{19}$$

S being the Standard Deviation for all samples with subgroups:

$$S = \sqrt{\frac{\sum_i \sum_j (x_{ij} - \bar{x})^2}{(\sum n_i) - 1}} \tag{20}$$

Where:

- x_{ij} = The j^{th} observation of the i^{th} subgroup
- $\bar{x}$ = Process mean
- n_i = Number of observation in the i^{th} subgroup.
- $C_4(N)$ = Unbiasing constant.
- $N = \sum n_i$ = Total number of observations.

Note, it is usually best practice to not use the unbiasing constant when estimating $\sigma_{overall}$, you should estimate $\sigma_{overall}$ by S directly.

If there are no subgroups, then the formula for S is just:

$$S = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}} \tag{21}$$

Where: S = Sample standard deviation N = The number of observations x_i = The observed values of a sample item and $\overline{x}$ = The mean value of the observations.

3.3.3 Between Standard Deviation

Is an estimate of the variation between the subgroups. For example, if each subgroup is collected from a different batch of items, large between-subgroup standard deviation indicates a large amount of variability between the items in different batches. **It is used calculate the between/within subgroup variation**

3.3.4 B/W Standard Deviation

Is a single value that includes both the variation between subgroups and the variation within subgroups. It's the square root of the sum of the between-subgroup variance and the within-subgroup-variance. **It is used calculate CP, CPK and other measures of the between/within capability of a process**

3.3.5 Capability Indices

CP CP is a measure of the potential capability of the process. It is a ratio comparing two vales. The specification spread ($USL - LSL$) - The spread of the process based on the standard deviation. - **Cp evaluates potencial capability based on the variation in your process, not its location**

$$Cp = \frac{USL - LSL}{TOL \times \hat{\sigma}_{within}} \tag{22}$$

Where:

- USL = Upper specification limit
- LSL = Lower specification limit
- TOL = Multiplier of the sigma tolerance *Usually six*
- $\hat{\sigma}_{within}$ = Within standard deviation

Interpretation Because CP doesn't consider the location of the process, it indicates the potencial capability that your process could achieve if it were centered.

CPL It's a measure of the potencial capability of the process based on its LSL. CPL is a ratio that compares two values - The distance from the process mean to the LSL - The one-sided spread of the process based on the within subgroup standard deviation.

Because CPl considers both the process mean and the process spread, it evaluates both the location and the variation of th eprocess.

$$CPL = \frac{\bar{x} - LSL}{(\frac{TOLER}{2})\hat{\sigma}_{within}} \tag{23}$$

CPU It's a measure of th epotencial capability of the process based on its upper specificaction limit.Compares two values: - The distance from the process mean to the USL - The one-sided spread of the process based on the variation within the subgroups.

$$CPU = \frac{USL - \bar{x}}{(\frac{TOL}{2})\hat{\sigma}_{within}} \tag{24}$$

CPK It's a measure of the potencial capability of the process and equals to the minimum of the CPU and CPL. Use CPK to evaluate the potential capability of your process based on both the process location and the process spread. Potential capability indicates the capability that could be achieved if the process shifts and drifts were eliminated.

$$CPK = min(CPU, CPL) \tag{25}$$

For more information click here.

PP It's a measure of the overall capability of the process. PP is a ratio that compares two values. - The specification spread (USL - LSL) - The spread of the process based on the overall standard deviation.

It evaluates overall capability based on the variation in the process, but not its location.

Where: - USL = Upper specification limit - LSL = Lower specification limit - TOL = Multiplier of the sigma tolerance

standard deviation

Todo. PPU

$$PPU = \frac{USL - \mu}{(\frac{TOL}{2})\hat{\sigma}_{overall}} \tag{26}$$

Todo. PPL

$$PPL = \frac{\mu - LSL}{(\frac{TOL}{2})\hat{\sigma}_{overall}} \tag{27}$$

Todo. PPK It's a measure of the overall capability of the process and its equals to the minimum of the PPU and PPL

$$PPK = min(PPU, PPL) \tag{28}$$

PPM PPM is the expected number of parts per million that are outside of the specification limits, it's based on the overall variation of the process.

$$[PPM < LSL(\text{Exp. Overall})] + [PPM > USL(\text{Exp. Overall})] \tag{29}$$

$$= \left(1,000,000 \times \left[1 - \phi\left(\frac{\bar{x} - LSL}{S}\right)\right]\right) + \left(1,000,000 \times \left[1 - \phi\left(\frac{USL - \bar{x}}{S}\right)\right]\right) \tag{30}$$

This formula is used in statistical process control to estimate the number of defective parts or products in a manufacturing process. It involves two main steps:

First, calculate the number of parts per million (PPM) that fall below the lower specification limit (LSL) or above the upper specification limit (USL), given the expected overall process capability. This is done using the cumulative distribution function (CDF) of a standard normal distribution.

Second, add the two PPM values together to get the total estimated number of defective parts.

Here is a step-by-step explanation of the formula:

- $PPM < LSL(Exp.Overall)$
  Represents the number of defective parts that fall below the lower specification limit (LSL) when the overall process is expected to be in control.
  The term "Exp. Overall" refers to the expected overall process capability, which is typically determined using historical data or engineering analysis.
- $PPM > USL(Exp.Overall)$
  Represents the number of defective parts that fall above the upper specification limit (USL) when the overall process is expected to be in control.

- $1,000,000 \times [1 - \phi((\bar{x} - LSL)/S)]$
  This term calculates the PPM of defective parts that fall below the LSL. It uses the CDF of a standard normal distribution to find the proportion of the process data that fall below the expected LSL.
  The term $(\bar{x} - LSL)/S$ represents the number of standard deviations the process mean $\bar{x}$ is below the LSL.
- $1,000,000 \times [1 - \phi((USL - \bar{x})/S)]$
  This term calculates the PPM of defective parts that fall above the USL. It uses the CDF of a standard normal distribution to find the proportion of the process data that fall above the expected USL.
  The term $(USl - \bar{x})/S$ represents the number of standard deviations the process mean $\bar{x}$ is above the USL.

Here is a step-by-step explanation of the formula:

- $PPM < LSL(Exp.Overall)$
  Represents the number of defective parts that fall below the lower specification limit (LSL) when the overall process is expected to be in control.
  The term "Exp. Overall" refers to the expected overall process capability, which is typically determined using historical data or engineering analysis.
- $PPM > USL(Exp.Overall)$
  Represents the number of defective parts that fall above the upper specification limit (USL) when the overall process is expected to be in control.
- $\frac{\bar{x} - LSL}{S} = Z_{LSL}$
  Z_{LSL} Represents the number of standard deviations the process mean $\bar{x}$ is below the lower specification limit (LSL)
- $\frac{USL - \bar{x}}{S} = Z_{USL}$
  Z_{USL} Irepresents the number of standard deviations the process mean $\bar{x}$ is below the lower specification limit (LSL)
- $\phi(\cdot)$ Is the Cumulative Distribution Function (CDF) of a standard normal distribution.
  The CDF of the standard normal distribution is used to find the proportion of the process data that fall below the expected LSL
- $\phi(Z_{LSL})$ The CDF of the standard normal distribution is used to find the proportion of the process data that fall below the expected LSL
- $\phi(Z_{USL})$ The CDF of the standard normal distribution is used to find the proportion of the process data that fall above the expected USL

Resources https://support.minitab.com/en-us/minitab/21/help-and-how-to/quality-and-process-improvement/capability-analysis/how-to/capability-sixpack/normal-capability-sixpack/methods-and-formulas/within-capability/#cp

https://support.minitab.com/en-us/minitab/21/help-and-how-to/quality-and-process-improvement/capability-analysis/how-to/capability-sixpack/normal-capability-sixpack/methods-and-formulas/overall-capability/#cpm https://support.minitab.com/en-us/minitab/20/help-and-how-to/quality-and-process-improvement/capability-analysis/how-to/capability-analysis/normal-capability-analysis/interpret-the-results/all-statistics-and-graphs/potential-within-capability/

https://www.six-sigma-material.com/Ppk.html

3.4 Design of Experiments

3.4.1 Full Factorial Designs

3.4.2 Fractional Factorial Designs

3.4.3 Response Surface Designs

3.5 Statistical Process Monitoring

3.5.1 Monitoring Process Stability

3.5.2 Detecting Process Changes

4 Quality Costs

Quality costs are a measure of the costs specifically assosiated with the achievement or non achievement of product or service quality including all product or service requirements established by the company and its contracts with customers ans society.

Quality costs repreent the difference between the actual cost of a product or service and what the reduced cost would be if there were no possibility of substandard service, product failure or manufacturing defects.

The most common format for categorizing quality costs is the Prevention Appraisal-Failure model.

- **Prevention costs:** The cost of all activities specifically designed to prevent poor quality in products or services.
- **Failure Costs:** Costs resulting from products or services not conforming to requirements or customer/user needs.
- **Appraisal Costs:** Costs associated with measuring, evaluating, or auditing products or services to assure conformance to quality standards.
- **Internal Failure Costs:** Failure costs occurring prior to delivery or shipment of the product.
- **External Failure Costs:** Failure cost occurring after delivery or shipment of the product.
- **Total Quality Costs:** The sum of the above costs, representing the difference between the actual cost of a product or service and what the reduced cost would be if there wer no defects.

The goal of any quality cost system is to facilitate quality improvement efforts that will lead to operating cost reduction opportunities.

The strategy for using quality costs is quite simple: 1. Take direct attack on failure costs. 2. Invest in the right prevention activities. 3. Reduce appraisal costs according to results achieved 4. Continuously evaluate and redirect prevention efforts to gain further improvement.

This strategy is based on the premise that: - For each failure there's a root cause. - Causes are preventable. - Prevention is always cheaper.

4.1 Taguchi Quality Loss Function (QLF) and the hidden costs of quality

A quality characteristic is whatever we measure to judge performance.

The Quality Loss Function is used to estimate costs when the product or process characteristics are shifted from the target value. This is represented by the following equation:

$$L(Y) = k(y - T)^2 \tag{31}$$

Where: $L(Y)$ is the cost incurred when the characteristic y is shifted from the target T and k is a constant depending on the process.

Loss occurs not only when a product is outside the specifications, but also when a product falls within the specifications. Although a loss function may take on many forms, Taguchi has found that the simple quadratic function approximates the behavior of loss in many instances.

Since the QLF curve is quadratic in nature, loss increases by the square of the distance from the target value.

4.2 Quality Cost Bases

Actual dollars expended is usually the best indicator for determining where quality improvement projects will have the greatest impact on profit and where corrective action should be taken, but unless the amount of work performed is relatively constant, it will not provide a clear indication of uality cost improvement trends.

The prime value of a quality cost system is in identifying opportunities for improvement and then providing a measurement of that improvement over time.

Short-range bases should be directly related by time and location to quality costs as they are being incurred and reported. They should relate the cost of quality to the amount of work performed.

Typical examples include overalll operating costs, total or direct labor costs, value added costs, and the actual average cost of delivered product or service.

For current, ongoing applications, several bases can be used. The following examples are typical indices that incorporate this feature: - Internal failure costs as percent of total production. - External failure costs as an average percent of net sales. - Procurement appraisal costs as a percent of total purchased material. - Operations appraisal costs as percent of total purchased material costs. - Total quality costs as percent of production.

There is no single perfect base, each base can be misleading if used alone and this can easily lead to confusion and disinterest. It's important to the success of quality cost use that bases for individual progressm easurements not appear unnatural to the functional area. Instead, they should be seen as complementary, for example "rework costs as percent of area labor costs".

They could aslo be used to provide indices that may have shock value simply to get the corrective action juices flowing, for example "Hey, did you know that for every dollar expended in your area, 50 cents is the cost of poor quality?"

4.3 Trend Analysis and the Improvement Process

Quality costs alone cannot do anything for a company except to illustrate what is being expended in specific areas related to quality and highlight opportunities for cost improvement.

TO put quaity costs to use, it's necessary to organize them in a manner that will support analysis. Use them to raise questions such as these: - Did you know that for every $100 in shipments, we lose $5 in internal distribution and handling failure costs?

Questions such as these immediately show the value of quality costs in direct relation to known costs expenditures.

To determine exactly where to establish short-range quality cost trend charts and goals, it is necessary to review the company's basic quality measurement system-

Real improvement depends on actions within the baisc quality measurement and crrective action system, enhance by the use of quality costs as an important suppot tool. Specific uses of quality costs, therefore, must be correlated to specific quality measurement target areas for improvement.

5 DMAIC

5.1 Define

5.2 Measure

5.3 Analyze

5.4 Improve

5.5 Control