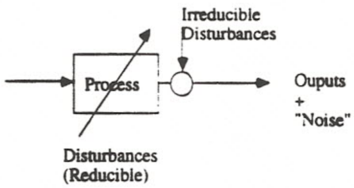
Fact Sheet for Probability Statistics and SPC

March 30, 1994

We can model a process as having noises (disturbances) that are both reducible and irreducible. A simple model could be :



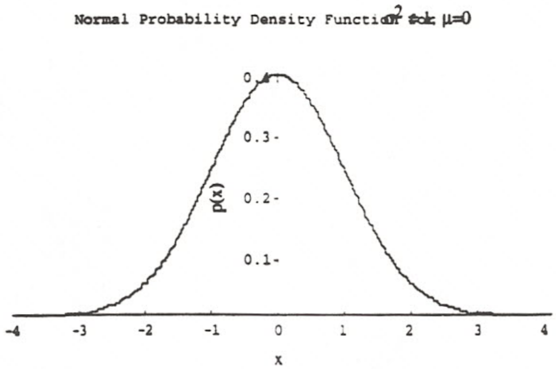
We assume that the noise or irreducible disturbance will always be present. It is also assumed that the noise is generated by a parent probability density function (pdf) that can be of any form (because of the Central Limit Theorem), but which we will assume is essentially normal.

The normal distribution is given by:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty \leq x \leq \infty$$

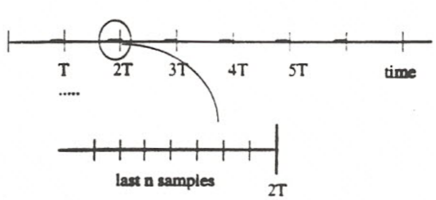
where μ is the mean of the distribution and σ^2 is the variance

The "standard" normal or Gaussian distribution is shown below



In using control charts, we seek to track the mean of a process output, and try to characterize the variation in that mean so as to determine the nature of the irreducible noise, and to detect the presence of reducible (assignable) noises.

This is done by sampling the output n times at a period **T**:



If we assume the parent distribution is normal then the following super-position principle holds:

If:

$$y = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + \ldots$$

(where c_i are constants and x_i are samples from any normal distribution)

then the mean of the distribution governing y is:

$$\mu_y = c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4 + \ldots$$

and the variance of y is:

$$\sigma_y^2 = c_1^2\sigma_i^2 + c_2^2\sigma_2^2 + c_3^2\sigma_3^2 + c_4^2\sigma_4^2 + \ldots$$

Thus if we take y as the average of n values taken at the *ith* sample time (T) for the same parent distribution, we get:

$$\bar{x}_i = \frac{1}{n}\left(x_1 + x_2 + x_3 + \ldots + x_n\right)$$

which is simply a linear sum with equal weights of $1/n$.

Assuming that the samples x_i came from a single normal distribution of mean μ_o and variance σ^2_o, the distribution of the new variable $\bar{x}$ is also normal with mean

$$\mu_{\bar{x}} = \frac{1}{n}\left(\mu_o + \mu_o + \mu_o + \ldots + \mu_o\right) = \mu_o$$

and variance

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2}\left(\sum_n \sigma_o^2\right) = \frac{\sigma_o^2}{n}$$

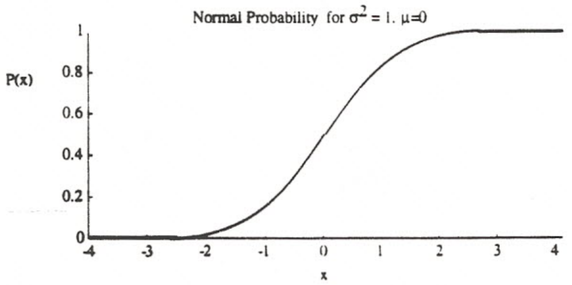
This implies that as the number of samples n taken increases, the mean of that sample remains at μ_o ;while the sample variance decreases. (In the limit as we take an infinite number of samples of x, we perfectly reproduce it and the variance of the mean goes to zero.)

Confidence Intervals.

Now we ask the question, how probable is it that x is above (or below) the mean value by a certain amount. To answer this we look at the Cumulative Probability function, which is defined as :

$$P(x) = \int_{-\infty}^{x} p(\eta)\,d\eta$$

and is plotted below:



Based on P(x), we can ask what is the probability that x is less than a certain value:

$$P(x \le a) = \int_{-\infty}^{a} p(x)dx$$

(The integral for P(x) cannot be evaluated analytically, but instead its values are tabulated in normalized form).

If instead we want to know the probability that x lies between two numbers a and b, then:

$$P(a \le x \le b) = P(x \le b) - P(x \le a) =$$

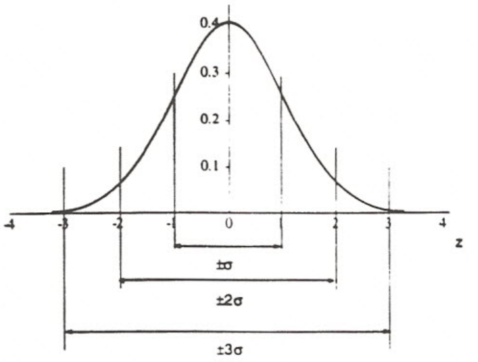
$$\int_{-\infty}^{b} p(x)dx - \int_{-\infty}^{a} p(x)dx$$

For the normal distribution we often define standard intervals of $\pm\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$. Using a normalized variable $z = \frac{x-\mu}{\sigma}$; this implies $P(-1 \le z \le 1)$, $P(-2 \le z \le 2)$ and $P(-3 \le z \le 3)$

From the *Cumulative standard normal distribution* tables this yields:

$P(-1 \le z \le 1) = P(z \le 1) - P(z \le -1) = 0.841 - (1-0.841) =$ **0.682**
$P(-2 \le z \le 2) = P(z \le 2) - P(z \le -2) = 0.977 - (1-0.977) =$ **0.954**
$P(-3 \le z \le 3) = P(z \le 3) - P(z \le -3) = 0.998 - (1-0.998) =$ **0.997**
Thus, out of 1000 values taken from a normal distribution, 997 will lie within 3σ of the mean, 954 will be within 2σ and 682 will be within σ.

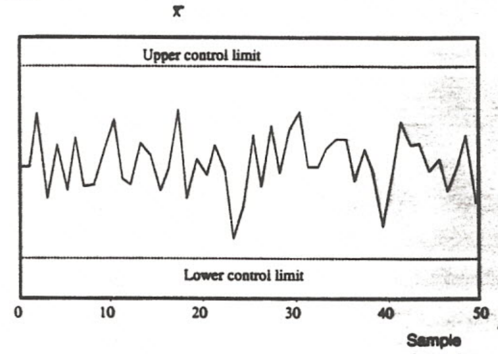
These confidence levels are in fact the area of the enclosed space on the probability function as shown below:



The $\bar{x}$ chart

Shewhart developed the $\bar{x}$ chart as a means of tracking the variation of the process and detecting changes in the output. For a "mass production" process, the process mean is expected to stay stationary, and thus the $\bar{x}$ chart is used to detect drifts or unacceptable variations in the mean. These then are an indication of "new" noises or disturbances in the process that should be reducible.

Using the definition for $\bar{x}$ based on sampling given above, this value can be plotted versus the chronological sample it represents.



The centerline of this plot is typically the "grand mean" of the process defined as:

$\bar{\bar{x}} = \frac{1}{N}\sum_{j=1}^{N} \bar{x}_j$ where $\bar{x}_j$ is a sample mean, typically taken from "prior" data. The upper and lower control limits are set most typically by bands at $\pm 3\,\sigma_{\bar{x}}$

To find $\sigma_{\bar{x}}$ from the "prior" data, we first calculate a grand standard deviation:

$\bar{S} = \frac{1}{N}\sum_{j=1}^{N} S_j$ where S_i is the sample standard deviation computed from:

$$S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_j)^2}$$

However, it can be shown (using the expectation operator) that this is a biased estimate of $\sigma_{\bar{x}}$, such that

$E(S_j) = C_4\,\sigma_{\bar{x}}$ where $C_4 = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$

Thus we set the upper and lower control limits on the xbar chart :

$UCL = \bar{\bar{x}} + 3\frac{\bar{S}}{C_4\sqrt{n}}$ and $LCL = \bar{\bar{x}} - 3\frac{\bar{S}}{C_4\sqrt{n}}$

The S-chart is similar to the xbar chart except that it tracks the sample standard deviation S_i, as defined above. Here the upper and lower control limits are defined by the standard deviation of the sample standard deviation, which is given by:

$$\sigma_S = \sigma\sqrt{1 - C_4^2}$$

Thus the S-chart is constructed with $\bar{S}$ as the centerline and with control limits:

$UCL = \bar{S} + 3\frac{\bar{S}}{C_4}\sqrt{1-C_4^2}$ and $LCL = \bar{S} - 3\frac{\bar{S}}{C_4}\sqrt{1-C_4^2}$

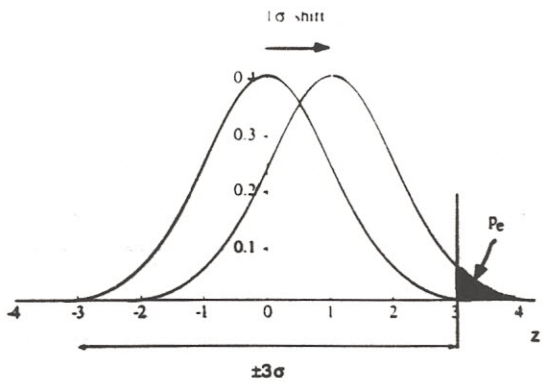
Average Run Length:

The main purpose of an xbar or S chart is to detect a change from the expected centerline value. The question then arises, how sensitive is the chart to a change once it occurs. This can be defined by the Average Run Length (ARL) which is defined as the inverse of the probability of exceeding the control limits:

ARL = $1/p_e$.

For example, if no mean shift has occurred, p_e= 1-P($x<3\sigma$) =1-0.99865=0.00135 . Thus the ARL for a point exceeding the UCL or LCL is 1/0.00135 = 740, that is we can expect to see a point outside the control limits once every 740 points. However, if the actual mean of the process has shifted, then p_e will increase and ARL will decrease. For example if a mean shift of +1 $\sigma_{\bar{x}}$ has occurred, then p_e = 1-P($x<2\sigma$) = 1-.97725 = 0.02275 and the ARL to detect this shift will be 1/0.02275 = 43.

This change can be best understood looking at the figure below:



From this figure it is evident that p_e is the area of the pdf that is outside the UCL and therefore is defined by the cumulative probability function P.

If the ARL is too long (that is too many parts would be made before a mean shift was detected) it can be shortened either by decreasing the UCL and LCL arbitrarily, or by increasing the number of samples (n) used to determine $\bar{x}$. If n is increased, then the variance of the sample mean decreases:

$$\sigma_{\bar{x}} = \frac{\bar{S}}{C_4\sqrt{n}}$$

As this happens, the same magnitude shift will cause p_e to be much larger and the shift will be more readily detected. For example, if the process shifts the same amount as shown above, but the sample size n is increased by 4 times (thereby halving $\sigma_{\bar{x}}$), then the net shift is twice as large, and the p_e = 1-0.841 = 0.16, and ARL = 6.25.