

# The Application of Process Distribution Modeling to Accurately Characterize Processes

*Is your process capable?*

Sean Hanna - Principal, [seanh@prodengmgtgroup.com](mailto:seanh@prodengmgtgroup.com)

**Keywords:** Process distribution, Process modelling, DPMO, Cpk, Process capability

**Abstract**

It's essential to know often you are shipping bad product to your customers, 100% inspection is not acceptable from a cost or quality standpoint in addition to giving you no predictive information. Quantitative process capability indices can easily give you this information and are, just as easily, totally misleading if you don't understand the process distribution from which they were derived. In the paper we describe how to characterize your process capability accurately so that you can use it as springboard for continuous improvement.

Cp	= Process Capability
Cpk	= Process Capability index
USL	= Upper spec. limit
LSL	= Lower spec. limit
$\mu$	= Population Mean
$\sigma$	= Population Std Deviation

$$\hat{C}_{pk} = \min \left[ \frac{USL - \hat{\mu}}{3 \times \hat{\sigma}}, \frac{\hat{\mu} - LSL}{3 \times \hat{\sigma}} \right]$$

Don't worry too much about the formulas yet. They key is that it is all based on the assumption that, if enough parts are sampled, an in-control process will produce a "normal" curve as shown. This curve is defined

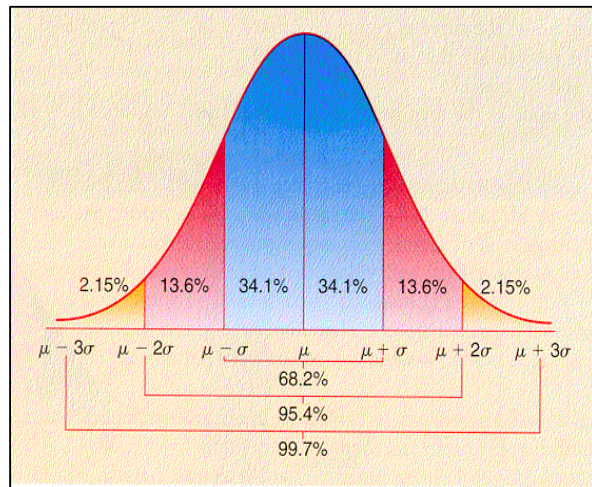
**1. Introduction**

If your business is concerned about satisfying your customers then you should have already started your journey down the well-travelled quality path:

<p>Reducing Customer returns → Improving Product Yield          Yield → Understanding Ave/Sigma → Calculating Cpk          → Controlling and improving Cpk</p>
--

If you're not yet familiar with Cpk

$$\hat{C}_p = \frac{USL - LSL}{6 \times \hat{\sigma}}$$



completely by the mean and the standard deviation. The mean is simply another term for the average, and the standard deviation is a measure of spread, or variation.

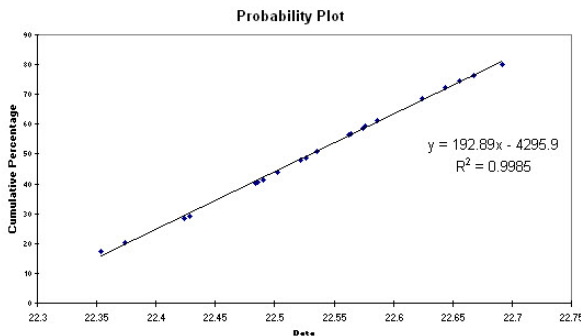
From the math of the distribution describing the normal curve, the std dev or “sigma”,  $\sigma$ , is defined such that

99.7% of the data will lie within  $\pm 3$  sigma from the mean, as shown. This information is very powerful in defining process capability and designing control charts.

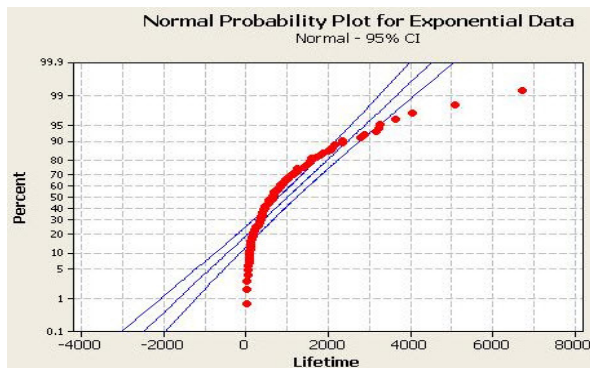
Cpk is a very useful process indicator, but is also dangerous if applied without understanding how the number was derived. Even if your Cpk for a key parameter is now  $>1.0$  or even  $>1.33$ , what does this really mean, and does it give you the information required to understand the risk of shipping defective product?

Most manufacturing processes utilize statistical tools that assume normality, and Cpk is no exception. Is the process in question normal or not? The easiest first assessment is, by observation, using the known properties of the normal distribution function:

- Make a histogram and calculate the descriptive statistics
  - Mean ~ Median ~ Mode should be approx. equal
  - Histogram should be around the mean / 65-70% of data within Mean  $\pm$  1sigma, 95% within  $\pm$  2sigma,  $<1\%$  outside  $\pm$  3sigma.
- If it is approximately “normal”, generate a normal probability plot and look for deviations from the best fit line. The slope of the line should be close to 1.0 and the data be closely grouped about the best fit line, e.g.  $r^2 > 0.95$ .



Example of normal process

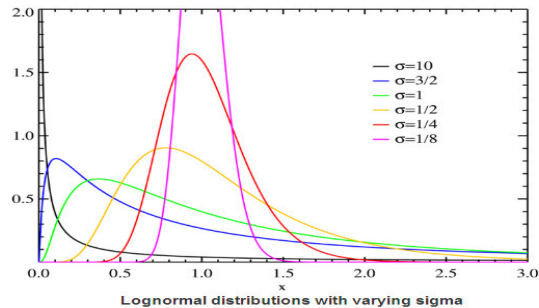
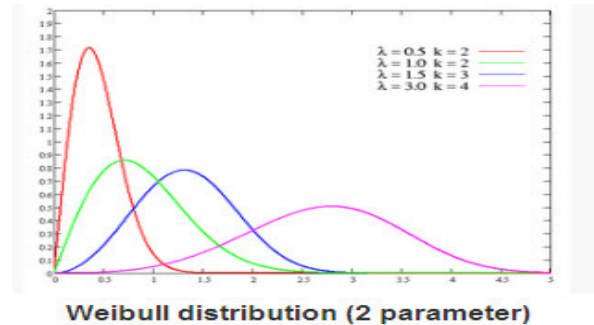


Example of exponential process plotted on normal prob. chart

If your process meets these criteria, with the plot like that on the left, your process is normal. [There are a number of statistical tests that can be applied to definitely determine the distribution in question, i.e. Kolmogorov-Smirnov and Anderson-Darling but, in most cases, the empirical analysis above will suffice.]

If the criteria are not met you have a non-normal distribution, and the most likely alternatives are lognormal and Poisson. In manufacturing one sided parameters, which have a hard stop at zero, usually produce a lognormal distribution and attribute (pass/fail) parameters generate a Poisson or binomial distribution. Weibull, exponential and uniform distributions can also show up in special cases. Unless your process output is entirely random, which would mean it is not a process at all, you will be able to approximate a distribution.

Using probability plots designed for each distribution, which are available through software packages such as Minitab, JMP or SAS, the same methodology as applied to the normal plot will tell you which distribution function best fits your data.



So, once you’ve determined the distribution your data best fits then how can you proceed to utilize the information? You’ll need a powerful software package such as those just

mentioned, at least for the first two approaches:

- Apply the real distribution - Calculate the appropriate Cpk based on the actual shape of the distribution. For non-normal cases, the median is used for the central tendency, rather than the mean. The software will plug in the distribution values (Up at the 99.865<sup>th</sup> and Lp at the 0.135<sup>th</sup> percentiles) and will define Cpk as the minimum of Cpu and Cpl.

$$\widehat{CPU}' = \frac{USL - m}{U_p - m}$$

$$\widehat{CPL}' = \frac{m - LSL}{m - L_p}$$

This corresponds to the same level of sensitivity used for traditional Cpk calculations with a normal distribution, where 99.7% of the data is within  $\pm 3\sigma$ .

- Use a transformation - Transform the data into a normal distribution using Box-Cox or Johnson transformations then use the normal Cpk calculations.
- Apply subgroups - Due to the Central Limit Theorem, the averages of these sub-groups will tend to a normal distribution even if the distribution of individual values is non-normal. However, the sub-group should be  $>5$  for this to be effective. This is one of the characteristics which make traditional control charts so powerful, a story for another day.
- NOTE: Make sure you really have a non normal distribution by stratifying the raw data. What appears to be a log normal distribution may actually be the combination of two or more normal distributions with varying and overlapping location and spread. Typical breakdown variables in a manufacturing example would be operator, machine, shift and raw material.

In the log-normal case the distribution is skewed to the left, with a tail to the right. In this common and straight forward case, the transformation is, simply, the natural log (e) of each data point, with the resulting output being analyzed in the “normal” way.

Now we have the distribution modeled, we can equate the Cpk to the DPMO (Defect Per Million Opportunity) based on the mathematics of the distribution function. In the table we are using the normal distribution and this data is available for other distributions.

NOTE: Be aware that DPMO1 allows for a standard long-term drift of the process of 1.5sigma, a cornerstone of Six Sigma metrics, at least as originally developed by Motorola. This is a controversial topic, and we should choose a long-term process shift value between 0 and 1.5, based on our process experience. If a process has an actively managed control chart and appropriate responses to out of control conditions, it’s hard to see how the process average will drift as much as 1.5sigma over time. Another reality check is to compare the calculated DMPO to actual yield. If the two are not close, there is a flaw in the logic.

Sigma	With 1.5 sigma shift of mean allowance		Cpk	With NO 1.5 sigma shift of mean allowance	
	DPMO	Yield (%)		DPMO	Yield (%)
1.5	500,000	50.0	0.5	133,614	86.64
3.0	66,800	93.32	1.0	2,700	98.73
3.5	22,700	97.73	1.17	465	99.953
4.0	6,210	99.379	1.33	63	99.994
4.5	1,350	99.865	1.50	6.8	99.9993
5.0	230	99.977	1.67	0.57	99.9999
6.0	3.4	99.99966	2.0	0.002	99.99999

Now we know our distribution and have calculated our Cpk/DPM, so we know the process capability, right? Not exactly: You need to decide if your, or your customer’s, interest is in short-term or long term process capability, or both, with the key being how sigma is calculated.

Long Term Estimate – This simply takes the standard deviation of all the individual data points to estimate the process sigma. The data is unfiltered, unless initiative is taken to remove outliers. This is the “Cpk” that has been referred to so far in this article.

To make things confusing, the trend over the past 10 years has been to re-title this parameter as PpK (Process Performance Index). It is a historic statistic, stating what was achieved, but saying nothing about future capability. Still, it is often a parameter useful to track progress over time, to update management or to satisfy a customer’s request. In addition, if the process distribution is understood and fits a curve well, some prediction in future capability can be empirically derived, at least based on the data set in question at that point in time.

Short-term Estimate – For this purpose you have to look at the data on a chronological basis. These estimates require that there is a “process” which is under statistical control, and the way to do this is using a classic control chart with samples in sub-groups (Xbar, R) or, at a pinch an individual X, moving range chart.

The estimate of the process sigma is calculated by well-known formulas based on the chart being used (Xbar, R/ Xbar, S/ X, mR), e.g.

Chart	Xbar/R	Xbar/S	X/mR
Sigma Estimate	$Rbar / d_2$	$Sbar / c_4$	$1.047 \times mRbar$
$\sigma_e$			
<b><math>d_2</math> and <math>C_4</math> are constants that vary based on sample size and are found in tables, or automatically inserted by software. Since the X/mR chart always has a sample size of 1, the constant is fixed at 1.047.</b>			

The indexes derived from these sigmas are, these days, generally known as Cpk as they quantify the potential process capability, the opportunity for special cases inflating the spread being minimized. In fact, it must be confirmed the process is in control and, if not, out-liers removed or corrective action taken, before the Cpk can be calculated from the sigma estimate. This “new” Cpk is a forward-looking indicator that tells you the absolute best the current process can perform if special causes are eliminated. It is clear that PpK and CpK are very different indices and most sets of production data will show significant differences between the two. The difference between the two is where the opportunity for improvement lies within the existing process.

It's very important that anyone interested in Process Capability or Process Performance understands how the indices are calculated and how they are named. Unfortunately, to this day, there is no consensus as some people call “Cpk” PpK, and vice-versa. When establishing internal or external reporting, the preference should be clearly defined at the start.

With the process understood, and with Cpk and DPMO in hand, you are ready to

- Set a goal for Cpk/DPMO, depending on your customer and business drivers. Certainly, key parameters with a Cpk of <1 are a high priority. Appropriate goal setting will be discussed in a future white paper.
- Establish a monitoring system for key variable process capability, including both Ppk and Cpk. Do not use software to automatically calculate the process potential capability, Cpk. It should be derived from control chart sub-groups as explained before. Use Cpk to catch process changes as they occur and use Ppk as an after the fact scorecard, typically reported on a shipment or monthly basis.
- Mission critical parameter Ppk/Cpk should be part of a dashboard or balanced scorecard reviewed regularly by upper management.