The Use and Abuse of Cp_k

by Berton H. Gunter

Starting in this column and continuing for several more, I will discuss process capability indexes. More specifically, I'll look in detail at the so-called Cp_k index and its properties. Although Cp_k is only one of several capability indexes in use, it is probably the most popular, and its properties and problems are typical of the others.

Capability indexes are widely used and have been discussed in *Quality Progress*. (See, for example, "A Systematic Approach to SPC Implementation" in the April 1987 *QP*.) Nevertheless, important statistical issues that deal with the appropriate use of Cp_k have appeared only in the technical literature and are thus unknown to many users. Unfortunately, these issues are not trivial: severe problems result when they are ignored. As a result, much of the standard usage of Cp_k indexes is incorrect and misleading. Many professionals familiar with these issues now recommend that the use of capability indexes be severely curtailed or even abandoned because of these problems.

Because many *QP* readers must deal with Cp_k and its relatives on a routine basis, a careful exposition of these issues in this column is worthwhile. They are not inherently difficult, but do require an understanding of some basic statistical ideas that are often not so carefully taught. In keeping with the spirit of this column, I'll assume only a minimal statistical background. All you need to know is how to make a histogram and what a statistical distribution is—particularly the familiar bell-shaped Gaussian (or normal) distribution.

I will look at Cp_k as a paradigm by breaking the issues into three parts. This column will define Cp_k and explain how and why it is used by studying its properties under theoretically "perfect" assumptions. The March column will show with some simple computer simulations and graphs what can happen when "perfect" assumptions don't hold, and why in many common situations they shouldn't be expected to. Finally, the third part will move from the theoretical to the practical, showing what can happen, how it relates to the earlier theory, and what practical approaches can be taken to deal with the real world contingencies.

Cp_k defined

Consider a controlled process that is producing a stable distribution of results in some key measured parameter. For definiteness, imagine a process drilling fixed-diameter holes in printed circuit boards with the parameter being the measured diameter. Figure 1 gives the distribution of the measured diameters. One can also imagine this as a histogram of millions of the diameters. Note that the distribution isn't symmetric. This is because the nature of drilling is such that there is a fixed lower limit to hole size (the drill diameters), but that upper diameters can get quite high due to wobbling in the chuck, dulling and wandering of the drills, and so forth. The target hole diameter of 60 mils and the specification limits of ± 4 mils are indicated on the distribution. Note that some of the distribution falls outside of the specifications; that is, some out-of-spec holes are being drilled.

Pretend that the true process distribution is known—something that never occurs in practice, and that it can be perfectly described mathematically—again, a convenient fiction. In practice, as will be shown later, the best one can do is approximate these situations. Continuing with the theory, however, it turns out that for most "naturally" occurring mathematical distributions, one can compute quantities μ and σ (mu and sigma) that are measures of the mean of the distribution (μ) and how spread out it is (σ). Note that these are Greek symbols, so used because they are idealized mathematical entities—or parameters—associated with the idealized mathematical distribution. They are actually computed using integration as if there were an infinite amount of data. This kind of mathematical approximation is common in science and engineering, of course. They should not be confused with averages and standard deviations, quantities computed arithmetically from a finite amount of real data. A discussion of the relationship between the parameters and the arithmetic computations will come later.

The owners and customers of the hole drilling process want to quantify the proportion of out-of-spec holes being produced as one way of benchmarking how well the process is performing. In the mathematical world of the distribution, this can be done perfectly using the underlying theory. Even in the mathematical world, however, it is convenient to make some approximations, and this is where Cp_k comes in. Cp_k works exactly right only for one theoretical distribution—the Gaussian. For all others it is an approximation.

The definition of Cp_k is:

$$Cp_k = \frac{|\mu - \text{nearer spec}|}{3\sigma}$$

The vertical lines in the numerator denote absolute value—that is, the positive amount of the difference, not the sign. Figure 2 shows μ and d_u and d_l, the distances to the upper and lower specs, respectively. The distances that are 3σ above and below the mean are also indicated.

Explanation and critique

Note that most of the distribution falls between $\pm 3\sigma$ of μ ("most" will be quantified later). This means that the distances 3σ above μ and 3σ below μ tell roughly how much room above and below the theoretical process mean there is for most of the hole diameters to fall. Of course, it would be nice to fit the distribution between the specs so that most of the holes fall within specification. This depends on three things: how close the process mean is to the target of 60 mils; how spread out the process distribution is compared to the specifications; and what the shape of the distribution is. Cp_k considers the first two of these issues, but ignores the third.

The denominator in the Cp_k index indicates how much room is needed on either side of the mean to contain most of the distribution. It is a measure of how variable the process is. Note that if the distribution were perfectly symmetric, the same amount of room would be needed on both sides; when the distribution isn't symmetric—as is often the case—there are problems. This is because Cp_k ignores shape.

The numerator measures the amount of room from μ to the specification on the side where there is less room. It is a measure of how well the process is centered. For example, if μ were equal to 64, it would mean that there is no room at all above the process mean before out-of-spec holes would be produced. Hence, Cp_k can be described as a way to measure the ratio of the amount of room needed to the amount of room available to produce product within specifications. If this ratio equals 1, then roughly speaking, there is just enough room to do the job. If the ratio is less than 1, more room is needed, and a higher proportion of out-of-spec holes will be produced. If Cp_k is greater than 1, there is extra room, and a higher proportion of holes will be within specifications.

Of course, as already noted, there is a problem with the use of the word "most" here: exactly how much "most of the distribution" really is depends on the kind of distribution. When the distribution isn't symmetric, the definition already begins to have problems because it wants to treat upper and lower specifications the same. Even if the distribution is symmetric, there can still be a lot of differences in what most is, depending on how thick the tails of the distribution are compared to the middle. In fact, the only time that most is what it is usually thought to be is in the case when the theoretical distribution is Gaussian. In this case, $Cp_k = 1$ means that almost 99.7% of the distribution will fall within the specifications. Of course, not only is this a very special case, but it is also a purely mathematical result. Future columns will show that it is doubtful whether the Gaussian is even all that good an approximation of what happens in most real processes.

Given these caveats, one might wonder why there is so much interest in Cp_k in the first place. Answer: it's an attempt to move away from zero defects (that is, whether product is in spec) and toward never-ending reduction in variability as a criterion for excellence. Figure 3 illustrates the situation. Using percent defective as the criterion for quality, both 3a and 3b represent "perfection": there are zero defects—nothing is out of spec. From the point of view of real quality, however, the much narrower variation in 3b means that the customer will see greater consistency of performance, higher reliability, and better fit and function from 3b than from the product distribution of 3a. Although percent defective doesn't distinguish between the situations and therefore fails to catalyze the desired improvement, Cp_k does: the distribution in 3b has a much higher Cp_k than that of 3a.

Summing up

That demonstrates the theoretical whys and wherefores of the definition of Cp_k. Even from this purely abstract perspective, it's clear that despite some important desirable characteristics, the widespread, uninformed use of Cp_k to benchmark all processes is of questionable validity. The next column will examine in greater quantitative detail how bad the theoretical situation can get, and how things can get even fuzzier with actual process data.

Berton H. Gunter is a contributing editor of *Quality Progress* and a statistical consultant based in Hopewell, NJ.

Figure 1.
The (nonsymmetric) distribution of hole diameters

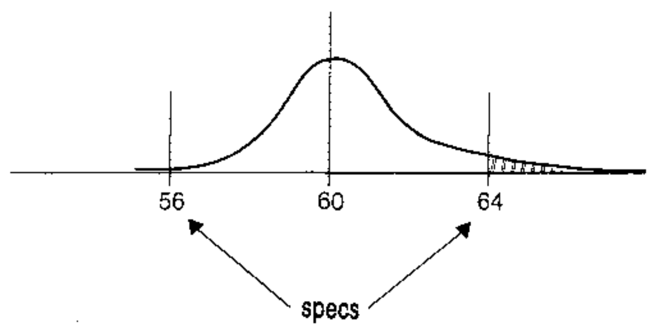


Figure 2.
Note that a portion of the distribution within 3σ of μ falls outside the upper specification.

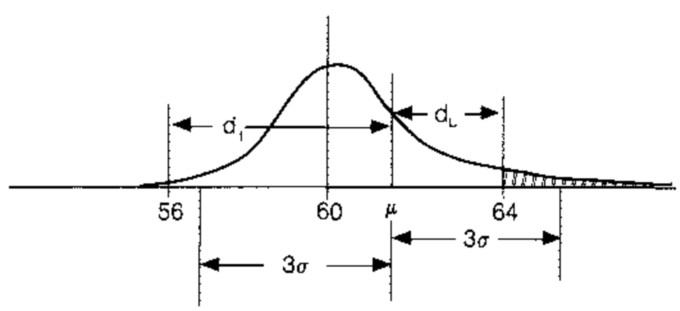
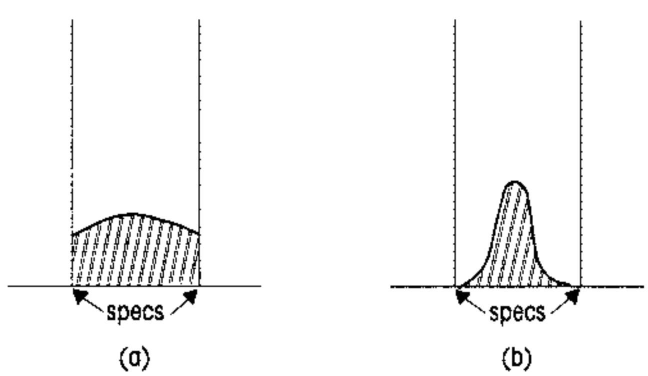


Figure 3.
Both a and b have zero defects, but the Cp_k of b is greater than that of a. Product quality is also better in b than a.

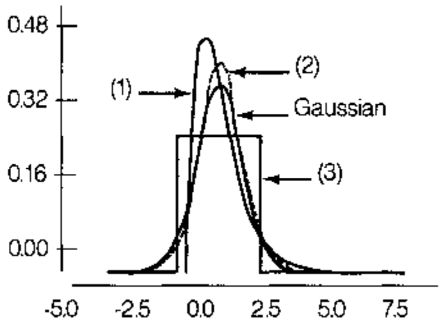


The Use and Abuse of Cp_k, Part 2

by Berton H. Gunter

In the January column, I briefly reviewed the definition and meaning of Cp_k. The discussion remained mostly in the mathematical realm. That is, I pretended that I was working with processes whose distributions were truly Gaussian (normal) and whose mean and spread were truly known. These were convenient fictions that were useful to help understand the underlying concepts, but we must now consider what happens in the real world when we attempt to reconcile practical realities with the idealized projections.

Figure 1. Four different shaped process distributions with the same μ and σ—and hence the same Cp_k.



The first assumption that I'll examine critically is the Gaussian one—that is, what happens when the processes are not well described as bell-shaped. Reasonable questions to ask at this point are whether non-Gaussian distributions ever occur in practice, and even if so, whether the central limit theorem effect will take care of things, as with control charts.

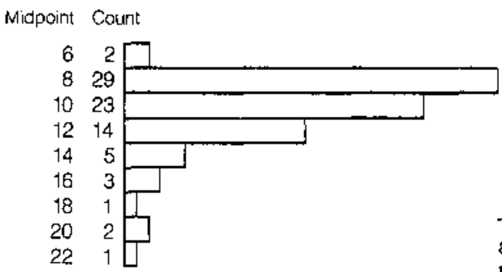
The answer to the second question can be given immediately: no. There is no central limit theorem effect operating here, and it turns out that we must be very concerned about non-Gaussian distributions. The details are too mathematical to be discussed here (interested readers can check the bibliography), but the example I'll give later should make the severity of the difficulties perfectly clear.

As to whether non-Gaussian distributions arise in practice, there are many statisticians who would say that non-Gaussianity is the norm and that Gaussianity is the rare occurrence. Although this might be somewhat pessimistic, there are certainly many situations where non-Gaussian process distributions are to be expected. Here are three commonly occurring situations where this is the case (see Figure 1).

Situation 1:
Skew distributions where a natural one-sided boundary exists (curve 1 in Figure 1).

This situation occurs with dimensional measurements like out-of-round, surface roughness, coating thickness, and so forth that cannot be less than zero. Similar situations occur for amount of contamination, and especially time until failure on a life test (for which skew distributions like the Weibull and gamma—never the Gaussian—are always used). In such situations, the process data bunch up around low values, while the tail strings out to the right. (Remember the light bulb that burned for 14 years? That was one of the points way out in the right-hand tail of the light bulb life distribution.) This can also occur for situations like the hole diameter of a drilled hole (discussed in the January column), where due to the nature of the process, a minimum hole size exists (the drill diameter) near which "most" hole diameters will cluster, while very large holes occasionally occur due to vibration, dull tools, or "crooked" setups.

Figure 2. Histogram of surface roughness in microinches for a machined part. The distribution is quite skew. $Cp_k = 2.4$.



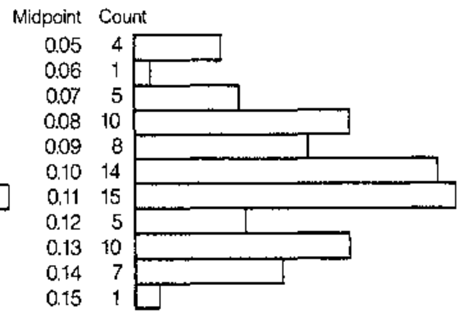
Situation 2:
Heavy-tailed distributions where deviation from perfect control causes less consistency than the Gaussian (curve 2 in Figure 1).

Many applied statisticians (myself included) believe this to be more the norm than the "normal." (In fact, a whole branch of statistics, robust/resistant methods, has been developed to deal with this.) The Gaussian distribution is concentrated around its center, but in many—perhaps most—real industrial processes, it is practically impossible to maintain such consistency. Natural and unavoidable variation (Deming's common causes) from small machine-to-machine or line-to-line differences, slight deviations in setups, environmentally induced variability, and human variation all conspire to spread out process results more to the tails and away from the center. Such results can also occur if a process tends to run most of the time in good control and occasionally, but regularly, somewhat less consistently. Detecting and characterizing this sometimes subtle effect can be very difficult, requiring a large amount of data from a stable process (hundreds or even thousands of observations might be required). Nevertheless, it is surprisingly true that subtle deviations of this sort from perfectly Gaussian behavior—differences that might be practically undetectable—can severely compromise the effectiveness and accuracy of Cp_k procedures.

Situation 3:
Short-tailed distributions where results are distributed more or less evenly between limits (curve 3 in Figure 1).

This is the standard 100% inspection situation where the process is too variable and the bad product in the tails must therefore be inspected out. However, it also occurs in other situations without 100% inspection. Such uniform distributions will occur, for example, when one combines output from several different machines, lines, or testing setups that are each internally consistent, but which are not properly maintained to agree. (Think of several Gaussian distributions lined up next to each other.) This is a common industrial situation. Another possibility is when there is excess process variability, but the design of the equipment itself cuts off the variability at some limits. An example might be making pellets of some type (candy, drugs, plastic) on a machine that passes the pellets over a screen with holes through which small ones drop, and then through an opening that prevents large ones from passing (a kind of automated sorting built into the machine). Again, as we shall soon see, this can invalidate the results of Cp_k calculations.

Figure 3. Histogram of the reciprocals of the surface roughness data. Distribution appears much more Gaussian. $Cp_k = .97$.



I believe most quality practitioners would agree that these three situations are common; most can probably also name several more in which non-Gaussian results could be expected. It is easy to mathematically examine how such situations can compromise the Cp_k procedures. The three non-Gaussian distributions and the Gaussian in Figure 1 all have exactly the same (true) mean and (true) standard deviation—and hence the same Cp_k! (For the statistically inclined, the heavy-tailed distribution is student's t with 8 df and the skew distribution is chi-squared with 4.5 df, scaled and centered to have $\mu = 0$ and $\sigma = 1$.) The approximate proportion of each distribution that is greater than $\pm 3\sigma$ from μ expressed in parts/10,000 is:

Gaussian: 30 (1/2 below, 1/2 above)
Skew: 140 (all above)
Long-tailed: 40 (1/2 below, 1/2 above)
Uniform: 0

Whether one looks at these differences as important or not depends on the application. As I will show next time, however, these mathematical subtleties will have graver consequences when it comes to the practical aspects of collecting process data and computing Cp_k from the data. It is here that the real dangers of Cp_k become manifest. The following example—provided by Gary Stork of Ford Motor Co.—well illustrates the dangers. A full discussion of the issues inherent in it will come in the next column.

Figure 2 is a histogram of 80 measurements of surface roughness in microinches on a machined part. The specification is 32 microinches, max. Note that the distribution is highly skew, as one might expect for this kind of data. The process was in control when these measurements were taken. Because these are real data, not an idealized situation, we must estimate the true underlying μ and σ. For this example, we use $\overline{X}$, the average of the data, and s, the standard deviation of the data, respectively, for μ and σ. Other choices might be better, but this is the conventional way of doing things. In this case, $\overline{X} = 10.44$ and $s = 3.053$, giving $Cp_k = 2.4$. Based on Gaussian theory, this would imply that less than one part per 10 million would be out of spec!

However, the distribution is clearly not Gaussian, so this conclusion is highly suspect. One approach to dealing with non-Gaussian data is to transform the data so that in the transformed scale, the data do look Gaussian (one can use analytic and/or graphical techniques to find appropriate transformations; a simple graphical approach was used here). In this case, a reciprocal transformation was found to be a good choice, and Figure 3 gives a histogram of the reciprocals of the original values. The transformed data now look and behave Gaussian (for the technically minded, the correlation with the normal scores is 0.994). $\overline{X}$ and s for the transformed data are 0.1025 and 0.0244, respectively, giving a Cp_k of 0.97, predicting that about one part per 500 would be out of spec![1]

The difference in predictions is truly remarkable. Of course, we have no idea which is correct, but I think most of us would be inclined to believe that the prediction on the transformed scale is probably going to be a lot closer to the truth than that made with the raw data—which predicts that essentially no parts will ever be out of spec. Clearly, this example shows how mechanical application of Cp_k to all situations can lead to serious errors. As I will show next time, when sampling variability is properly taken into account, the situation gets even worse.

Footnote

1. Note that the specifications must also be transformed when computing Cp_k.

Bibliography

Chan, L.J., S.W. Cheng, and F.A. Spiring, "A New Measure of Process Capability: Cp_m," *Journal of Quality Technology*, Vol. 20, No. 3, 1988, pp. 162-175.

Kane, V.E., "Process Capability Indices," *Journal of Quality Technology*, Vol. 18, No. 1, 1986, pp. 41-52.

Bert Gunter is a contributing editor of *Quality Progress* and a statistical consultant based in Hopewell, NJ.

The Use and Abuse of Cp_k, part 3

by Bert Gunter

In parts 1 and 2 of this series, I looked at the theoretical properties of Cp_k. Interpreting Cp_k as an accurate indicator that most (greater than 99.5%) of the process output falls within specification depends critically on the true, but usually unknown, distribution of the data. On the basis of these properties, it seems reasonable to conclude that Cp_k should be viewed only as an approximate measure of the true percentage within specifications.

But it *is* an indicator, right? That is, even if we can't pin down the exact percentage that will fall within specifications, it ought to provide a relative score on which to rank process quality. Hence, if the Cp_k of supplier A's process is greater than the Cp_k of supplier B's, that should mean supplier A will have a larger percentage within the specification than supplier B. Or so it seems.

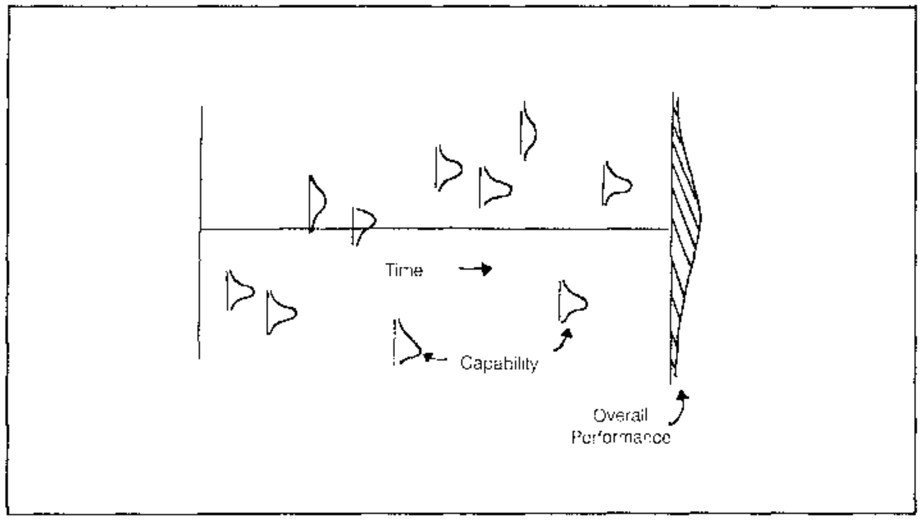
Unfortunately, appearances can be deceiving. The problem is that we have to use data to compute each Cp_k, and this means that we don't know the true values. We have only estimates (translation: guesses) based on the data. This means that there is uncertainty about the true Cp_k of each supplier. Thus, the best that can be done with the data is to make the usual kind of statistical statement that A exceeds B with some probability, or equivalently, provide some kind of confidence interval for the difference. To be specific about these matters, we never know the true μ or σ. We have only the average, $\bar{X}$, and standard deviation, s, of the data to use in Cp_k formulas. In practice Cp_k is a statistic, and to deal with it sensibly, we must understand how much uncertainty is associated with it. This turns out to be unknown in most cases.

Sources of uncertainty in Cp_k: sampling procedures

Control chart data are often used to calculate Cp_k. As usual, let's assume the data are collected in rational subgroups of size n, m subgroups in all, to give a total of N = mn data values. There are two ways to calculate $\bar{X}$ and s—and hence Cp_k—from these data. One way is to follow the control chart method given in QC texts in which the grand average is calculated from the subgroup averages and $S = \bar{R} \div d_2$. The other way is just to plug all N data values into a calculator or computer and press the average and standard deviation buttons. Will these two procedures both give essentially the same answer for Cp_k? If not, which is correct?

The answer is: unless the process is in a perfect state of statistical control, you will probably get quite different answers! Moreover, depending on exactly what you want to do with the Cp_k information, both answers are wrong!

Figure 1.
Control chart vs. calculator estimates of σ. The control chart calculation ($\bar{R} \div d_2$) captures only the spread of the within subgroup capability; the calculator method captures a weighted average of both the within and between spread.



The reason for this is as follows. The control chart and calculator procedures will give exactly the same answers (within rounding) for the overall process $\bar{X}$. But unless the process is perfectly stable, the s calculation will be different for the two procedures. Figure 1 shows why. In that figure we imagine each little distribution over a short period of time as forming a rational subgroup. Because the process is not stable, these subgroups shift over time. The $\bar{R} \div d_2$ calculation gives the average small within-group spread. It is a measure of the process capability—how things could be if all the assignable sources of variation that caused the process to jump around over time were eliminated. The calculator method gives a weighted average of both the within-group and between-group variation. The actual performance variation that will be seen in the overall process output—that is, the variation we actually must live with—is just the sum of the within- and between-group variation (variances).[1]

Of course, since s appears in the Cp_k denominator, the smaller it is, the larger Cp_k will be. So which procedure you use to calculate Cp_k might depend on which side of the supplier/customer fence you're on. If you're the customer, you want to see a Cp_k that accurately reflects all the variability that you must live with. If you're the supplier, you might want to "cheat" by making Cp_k larger with a value that reflects how you would do if you could get rid of all the assignable-cause variability. In most cases, it's the former—the actual overall results—that we're interested in.

Sources of uncertainty in Cp_k: sampling error

Once we're sampling and calculating s in the way we need, there is one more fact of life to be recognized: the variability in the data just due to the luck of the draw—sampling variability. That is, if the data could be collected again under the same process conditions that existed the first time around, we'd get somewhat different data, and thus different $\bar{X}$'s, s's, and Cp_k's. In other words, the Cp_k we have collected is only one of the many that might be obtained by sampling the same number of items under the same conditions.

As I mentioned earlier, this means that there is uncertainty about true process performance. This uncertainty is unavoidable and any reasonable approach to using capability indexes to measure quality performance must deal with it. The question is: how much uncertainty is there, and how do we deal with it using good statistical procedures? Unfortunately, this has not yet been figured out. Moreover, exact mathematical derivations might be of little practical use because, again, it appears that they are very sensitive to unverifiable assumptions.[2]

The situation is not as bleak as it might appear, however: we can easily do simulations

Figure 2.
Distributions of "perfect" and "occasionally erratic" processes.
The two distributions are hard to distinguish in practice using conventional tests, yet Cp_k behavior is quite different for the two.

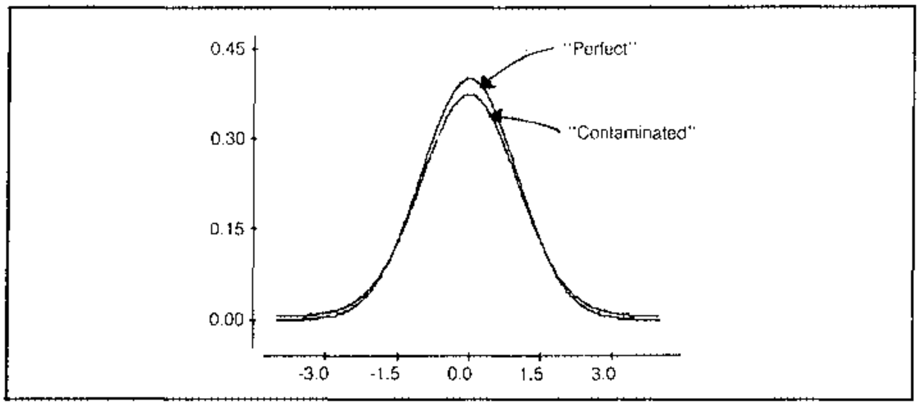
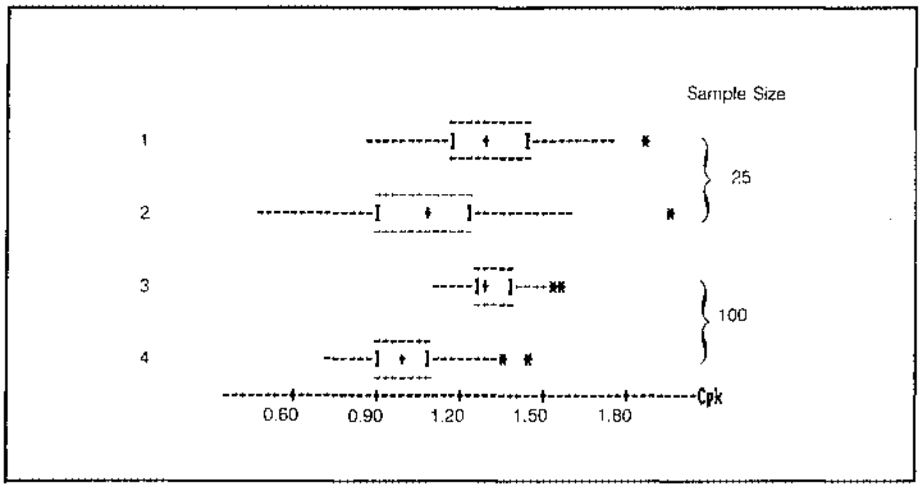


Figure 3.
Box plots of Cp_k Monte Carlo simulation. Each box plot gives 100 simulated Cp_k's from 100 samples of the sizes indicated. Numbers 1 and 3 are drawn from the "perfect Gaussian" of Figure 2; numbers 2 and 4 are drawn from the "contaminated Gaussian."



and see what happens in various cases. Figure 2 shows two idealized process distributions. The inner, more peaked distribution is Gaussian (normal) with a true Cp_k of 1.33. The slightly flatter curve represents a process that runs 90% of the time as this Gaussian and 10% of the time more erratically with a Cp_k of only about 0.5. This is one way to model the situation of processes that run well most of the time and not so well occasionally. Note how closely these curves follow one another. Unless you had a very large amount of data, it is unlikely that any of the usual tests for normality would be able to distinguish the two situations.

Nevertheless, this small amount of contamination has a considerable effect on Cp_k (whether this is good or bad is an interesting question). Figure 3 shows box plots of four computer runs comparing these situations. Runs 1 and 2 consisted of taking 100 samples of size 25 from the perfect and imperfect processes, respectively, and box plotting the resulting 100 Cp_k's from each. (Recall that the + is the median of the values, the box covers the middle 50%, or interquartile range, and the whiskers run to where most of the data should fall for well-behaved situations. The * represents possible unusual values worthy of closer examination, or as in this case, reflects the fact that the distribution of the Cp_k's is skewed to the right.) Runs 3 and 4 repeat this, but the sample size was 100, not 25.

Remember that for runs 1 and 3, the true Cp_k is 1.33, while for runs 2 and 4 a little excess variability that would ordinarily be undetectable has been mixed in. The first thing to notice is that, of course, with the larger sample size (100 as compared to 25), the bottom two box plots show less variability in the results. Nevertheless, the results still range from about 1 to 1.65 for the perfect Gaussian, and about 0.7 to 1.5 for the contaminated one. For the top two sample size 25 results, the variability is of course even greater. Since this is a small simulation (only 100 trials of each), these ranges ought to be roughly the widths of conventional (e.g., about 95%) confidence intervals. This shows that with small samples—say, under 100—the Cp_k computation doesn't give a very good idea of where things really stand. Remember, the variability of each box plot is just due to the luck of the draw and reflects no real change. As a result, it is entirely possible that a supplier with a true Cp_k of 1.5 could, through the luck of the draw, look worse from a small sample of data than another supplier with a true Cp_k of 1.25.

The second thing to notice is that the imperfect processes give, on average, a lower Cp_k, as might be expected. What is surprising is how much lower: about 25% or more in both cases. Note also that the variability of the Cp_k's for the imperfect process is also much increased, so that it is much harder to get a fix on the situation in this case, which is usually the one of interest. Again, remember that a practically undetectable amount (10%) of more variable behavior causes these changes.

All in all, even this simple simulation shows that unless a large amount of good data is taken from a stable process, Cp_k doesn't reveal a lot about what's going on, even when it's used only as a rough indicator and not as a hard measure of percentage within specification. In the July column I'll make some recommendations about what can be done to overcome these difficulties.

Footnotes

1. A delightful and readable article discussing these issues and showing how to determine within and between components of variance in a practical quality improvement case study can be found in "In the Soup: A Case Study to Identify Contributors to Filling Variability," by Lynne Hare, *Journal of Quality Technology*, Vol. 20, January 1988.

2. There are new computer-intensive procedures—particularly bootstrapping—that are now becoming available that are not dependent on so many assumptions and appear to give good results. Unfortunately, the QC community does not yet seem to be aware of them.

Bert Gunter is a contributing editor of *Quality Progress* and a statistical consultant based in Hopewell, NJ.

The Use and Abuse of Cp_k, part 4

by Bert Gunter

In the May column I discussed the fact that the σ that appears in the definition of Cp_k is generally calculated in either of two ways, via the $\bar{R}$ method of control charting or the usual textbook formula implemented on a calculator. In that column, I showed that unless a process is in a state of statistical control so that special causes that cause it to jump around are eliminated, these two calculations could yield considerably different results. I neglected to make two points concerning these matters.

First, many organizations now use SQC software to avoid hand calculation. Unfortunately, many users do not understand what the software is doing. In particular, they do not know which method the software uses to calculate σ, so that they could be misleading themselves and their customers without even knowing it.

Second, those who understand the ideas at the heart of statistical process control understand that I have just raised a red herring. Let me clarify the matter.

The results of the last three columns show that although Cp_k sounds like a good idea, it cannot be used when specifications are one-sided or when process data are not nearly Gaussian. Moreover, because the sampling distribution of the Cp_k statistic is so variable, it should not be used unless relatively large samples (100 to 200, at least) are obtained. Cp_k's based on small samples are potentially misleading.

Despite this, some Cp_k advocates say it can serve as a means of motivating quality improvement: it functions not as a guide, but as a goal for which to strive. Unfortunately, there are dangers even here.

In the Spring, 1989 *Chemical and Process Industries Division (ASQC) Newsletter*, Jim Lucas of Du Pont reported on some work of his colleague, John Herman[1], showing that arbitrary Cp_k goals can be impossible to meet when measurement error is large—often the case with the complex measurements common in modern technology. It is important to note that "measurement error" means all the sources of variability in a measurement. This means that when an unchanging standard is measured, we include as measurement error the differences obtained by different people, with different sample preparation and setup, on different days, before and after maintenance and calibration, when the temperature and humidity vary, and so forth—not merely the minor instrument error obtained by pushing the button repeatedly on a single setup. As Lucas indicates, measurement capability studies—or variance components analyses—are necessary to determine the origin and size of potential error sources. Moreover, as he again points out, it is not unusual to see measurement error as large as 25% or more of the specifications. (I once observed an operator trying to "control" a hydraulic press with a weight measurement that I found to have a larger standard deviation than did the arbitrary limits that she was supposed to keep the weights within; needless to say, this resulted in constant fiddling and near total chaos.)

To see why measurement error can make hash of arbitrary Cp_k goals, imagine that the specifications are ± 10 and that, in reality, the process is properly centered and has a standard deviation of σ_p. Moreover, suppose the total measurement error, σ_m, is 2.5 and that because of the extraordinary commitment to quality, a Cp_k goal of 1.5 has been set. As is well known, the total observed capability, σ_t, satisfies $\sigma_t^2 = \sigma_p^2 - \sigma_m^2$, so that even if the process variability were reduced to 0—an impossibility—Cp_k could never get above 1.33 (=20/[6 x 2.5]). Consider the potential for frustration and chaos in pursuing the impossible goal when measurement error is unknown or ignored! Good intentions misapplied can do more harm than good.

Is Cp_k an obstacle to never-ending improvement?

The previous discussions have described the problems that can befall the naive user of Cp_k. But I've saved the most egregious abuse for last: most of the time, Cp_k cannot be used at all! Cp_k is a meaningful measure of process quality only when a process is in statistical control. This was the red herring referred to earlier: Cp_k shouldn't be computed by either method if the process is unstable. Without statistical control, a process is unpredictable. What it does tomorrow might depend as much on the phase of the moon as on what it does today. Cp_k is meaningless because it provides only an arbitrary snapshot from which no generalization can be drawn. The Cp_k of an out-of-control process provides no information about what can be expected precisely because nothing can be expected. The goal under these circumstances, which are often the case when SPC has not previously been used, is to identify and eliminate the special causes responsible for the excess variation. In other words, the goal is to establish stability—not to produce numbers satisfying arbitrary goals.

The greatest abuse of Cp_k that I have seen is that it becomes a kind of mindless effort that managers confuse with real statistical process control efforts. Instead of helping them better understand and improve their processes, it functions as a kind of numerical hurdle that is either met and forgotten, or finessed via negotiations, well-chosen samples, and biased inspection procedures. In short, rather than fostering never-ending improvement, Cp_k scorekeeping kills it.

One example of which I recently heard is that of the earnest manager who demanded that all suppliers providing prototype parts for his department meet minimum Cp_k standards. He did so despite the fact that no processes for producing the parts yet existed and that the total number of parts might have been only a dozen or two. Another example is that of the supplier that, in order to meet the Cp_k requirements of a large customer, made sure that all parts collected for the special Cp_k measurements were made on the best machine by the most skilled operator. Because the contractual arrangements said nothing about how the data were to be collected—only that they had to be—this was entirely "legal." A final example is that of the plant that prided itself on its high Cp_k standards despite the fact that the control chart wallpaper in the QC department showed the processes to be ubiquitously out of control (but within specifications).

What's the alternative?

One of the major attractions of capability indexes is their simplicity: magically, with minimal statistical understanding required, they seem to provide a universal index of process quality. Is it any wonder that they have seduced so many? Unfortunately, life is not so simple: ignorance of the relevant statistical issues can be dangerous. As David Kearns, CEO of Xerox, said in his April 1989 *Quality Progress* interview:

> ". . . we know that most of us are still very far behind the Japanese in the use of tools and statistical techniques . . . The more I get into it, the more concerned about it I get. We've got to go faster and harder. And a lot of the things that are the least exciting to communicate—statistical techniques and tools—are the most important.
>
> ". . . This year, for the first time, we had a presentation by a team from a different company [at Xerox's Teamwork Days]: Florida Power and Light . . . [They] corroborated what we had learned from a recent Japanese review of

some of the things we're doing: we're not really utilizing the tools and the statistical processes enough. And the feeling is that if we don't we're not going to get this big step up that we need to take."

The fact is that there is no quick, simple alternative to Cp_k. As Kearns said, "Quality is not banners. It's not speeches. It's hard work." Cp_k is an attempt to take the easy way out. But there is no easy way. An understanding of process capability and where the opportunities for and obstacles to improvement lie cannot be boiled down to a single number. To make good decisions, the multidimensional nature of the situation must be understood (Figure 1); control charts must be used; appropriate statistical analyses must be performed; and designed experiments must be done. Subject matter knowledge and past experience must, of course, play a large role.

Simplicity is desirable when it clarifies, but it is no virtue when it promulgates bad choices. When the concepts underlying capability and control are not understood by decision makers, when capability studies are treated as enforced scorekeeping rather than as tools for process improvement, and when crucial statistical issues are ignored in favor of unschooled intuition, disaster lurks.

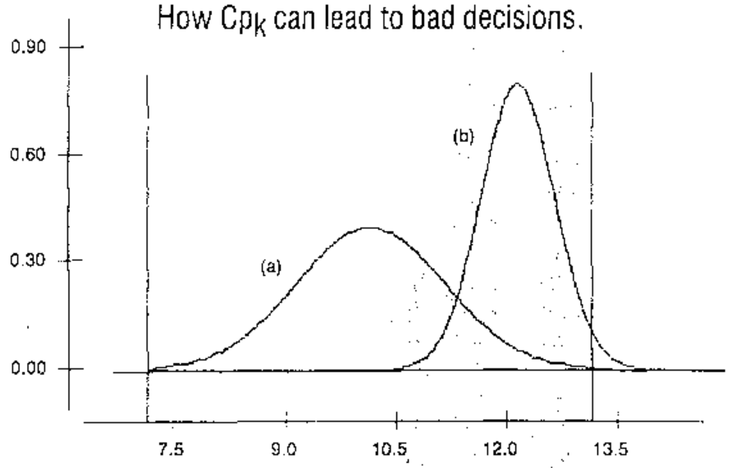
Far too often, Cp_k diverts attention from the real issues and compromises process improvement. It is time to question whether it would be wiser to abandon it and concentrate on promoting sound statistical practice.

Reference

1. John T. Herman, "Capability Index—Enough for Process Industries?," *Proceedings*, ASQC 43rd AQC, 1989.

Bert Gunter is a contributing editor of *Quality Progress* and a statistical consultant based in Hopewell, NJ.

Figure 1.
How Cp_k can lead to bad decisions.



Because Cp_k combines center and spread of a distribution, erroneous decisions can result. In the figure, process (a) has a Cp_k of 1.0, and process (b) has a Cp_k of only 0.67 (the vertical lines are specifications). But (b)'s off-centeredness could be due merely to an easily corrected measurement calibration error, while the wider spread of (a) might not be reducible. So (b) might be preferred, even though its Cp_k is higher. Both aspects of process performance—spread and centering—must be distinguished in order to make good decisions.