How many data for "process capability"?

Scott A. Hindle

Nestlé PTC Konolfingen, Nestlé-Strasse 3, Konolfingen 3510, Switzerland
E-mail: scott.hindle@rdko.nestle.com

Abstract
In considering how good a production process is, one crucial business question is: "*Is fully conforming process output expected?*". An analysis of process capability can help to answer this question. This means that data are needed, but how many? Use too few data and the risk of a wrong conclusion is too high. Too many data and it might take too long to get to a conclusion. Both imply waste. For example, not taking action on the process when action should be taken and/or taking the wrong action.

The purpose herein is to explore firstly what a proposed mathematical answer is to the "*How many data?*" question. Then, to explore how well this answer fits into real-world manufacturing processes where smart decisions need to be taken.

Keywords: Process capability, C_p, C_{pk}, degrees of freedom, predictable process, statistical control, number of data, uncertainty in standard deviation, action on a process

1. Introduction

For a manufacturer, one crucial business question is: "*Is fully conforming process output expected?*". An analysis of process capability can help to answer this question. This means that data are needed, but how many? Is "30 data", as often suggested, the right number?

The last two questions are examined using examples where individual values are collected in the analysis of process capability. Since the correct interpretation of capability statistics like C_p and C_{pk} is highly dependent on the behaviour of the process, the four examples presented in this paper start by placing their data on a control chart. (The control chart for individual values is used, also known as an XmR chart or process behaviour chart.)

2. The Uncertainty in an Estimate of Standard Deviation

2.1 Degrees of freedom and uncertainty

Given *n* data it is common to speak of *n*-1 degrees of freedom, herein d.f., in an estimate of standard deviation. The most common formula for which this is applicable is *s*, the sample standard deviation statistic:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} \qquad \text{Equation 1}$$

Symmetric functions to estimate standard deviation, of which *s* is one, are inappropriate when it comes to the analysis of process capability. Symmetric functions bury the information that is contained in the order of appearance in the original data, and must therefore be presumed inefficient until cleared (Deming, 1975). (Inefficient here refers to the burying of essential information in the ordering of the data, not statistical efficiency as introduced later (footnote #2 on page 3) when speaking of d.f.)

The origin of the use an average or median dispersion statistic, and not a symmetric function applied to all the data, in the analysis of process capability can be traced back to Walter Shewhart, the inventor of the Statistical Process Control “control chart” (Shewhart, 1931). The technical foundation of the control chart is also the same as that in the Analysis of Variance: a *within*-subgroup estimate of dispersion is used to filter out the “noise” in the data. Process data indicative of “noise” alone allow for a process to be characterised as predictable (in traditional terminology a process in “statistical control” and sometimes called a “stable process”).

Unfortunately, d.f. do not mean a lot to many people. Instead, the coefficient of variation (CV) of an estimate of standard deviation can be used to express the uncertainty in the computed standard deviation (CV is obtained from the ratio of the standard deviation of a variable to the mean of the variable). This is both easier to understand and communicate to others, and is represented by Figure 1. The relationship between the y- and x-axes is approximated by $1/\sqrt{2\times d.f.}$ (Wheeler, 2004).

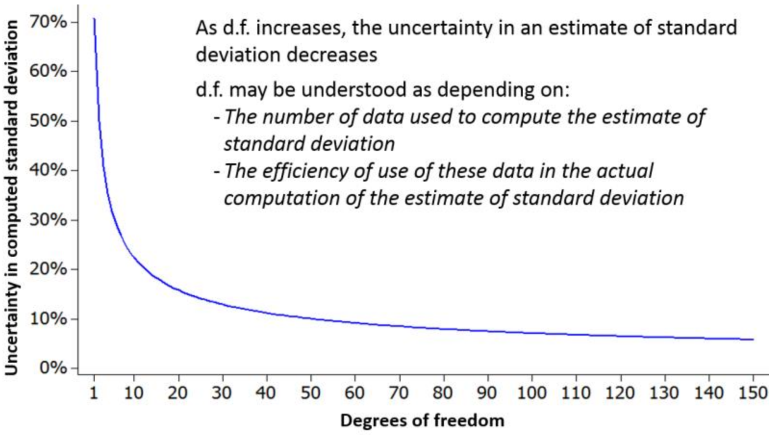

Figure 1: Plot of uncertainty in standard deviation against degrees of freedom.

Figure 1 could be understood and simplified as follows:

- With less than 10 d.f. an estimate of standard deviation lacks precision; getting more data should be seriously considered (this does not mean that more data can or must be obtained; see Example One)
- With 10 to 30 d.f. an estimate of standard deviation has started to solidify; there may still be sufficient payback from getting more data, but this payback is starting to weaken
- With more than 30 d.f. an estimate of standard deviation has effectively solidified; the need for more data would probably depend on other factors aside from the uncertainty of the estimate (unless, for example, data come along very quickly and inexpensively)

The above comments are based on the fact that to reduce the y-axis uncertainty by half the number of d.f. needs to increase four-fold. If 25 d.f. is considered insufficient, obtaining around 100 d.f. may well be required which, in quite some cases, is not be viable (probably for time and cost reasons).

2.2 Effective degrees of freedom and the average moving range method

In using an XmR chart, moving subgroups of size two are behind the estimated standard deviation, herein called SD_{within}. SD_{within} does not have a theoretical distribution characterised by a d.f. parameter, meaning that the *effective* number of d.f. are used (Wheeler, 2004)[1].

[1] The effective d.f. for the average moving range will be the d.f. for the square root of the pooled variance which would result in a CV (y-axis of Figure 1) which has the same value as the CV of the average moving range (Wheeler, 2004). The CV for a dispersion statistic is not changed by a linear transformation, hence SD_{within} has the same CV as the average moving range.

The average moving range method is used by default (in preference to the median moving range):

$$\overline{mR} = \sum_{i=1}^{n-1} \left| X_{i+1} - X_i \right| \Big/ (n-1)$$ Equation 2 (to calculate the average moving range)

$$SD_{within} = \overline{mR} \big/ d_2 = \overline{mR} \big/ 1.128$$ Equation 3 (1.128 is the appropriate bias correction factor)

SD_{within} has an *effective* number of d.f. approximated by $0.62 \times (n-1)$ (Wheeler, 2004)[2]. As an example, 30 data have an effective number of d.f. 18.0, corresponding to an associated uncertainty of ~16.7% in the calculated value for SD_{within}. (The quoted effective d.f. given herein come from Table 23 (p. 446) of Wheeler, 2004. The uncertainties are calculated using the approximation $1 \big/ \sqrt{2 \times d.f.}$)

While an uncertainty of ~16.7% may seem excessive, the reality is that it is often sufficient because:

- Many processes are not characterised as predictable (i.e. not in a state of "statistical control") so there isn't a well-defined standard deviation to estimate, no matter how many data we might collect
- High precision in SD_{within} – or lower uncertainty – is often not required, i.e. it is often not critical in reaching a sound business decision (higher precision can be important, however, when the process capability situation is somewhat borderline, i.e. the process may or may not be capable)

2.3 Quantifying process capability through C_p and C_{pk}

Process capability is often quantified using the indexes C_p and C_{pk}. These two indexes are used here with the capability requirement placed at C_p and $C_{pk} \geq 1.33$. (Many other indexes are in circulation, see for example Bothe, 1997.)

$$C_p = \frac{USL - LSL}{6 \times SD_{within}}$$ Equation 4 (depends on process variation)

$$C_{pk} = \min \left\{ \frac{\overline{X} - LSL}{3 \times SD_{within}} ; \frac{USL - \overline{X}}{3 \times SD_{within}} \right\}$$ Equation 5 (depends on both process variation and location)

(LSL and USL stand for lower and upper specification limit, respectively.)

2.4 Other factors to take into account

The uncertainty in an estimated standard deviation is just one factor. Others include:

- The ease and speed at which data can be collected
- The cost of obtaining data (some data are expensive)
- The importance of the characteristic for which data are collected

Determining how many data are needed has therefore much more to think about than what mathematical theory puts on the table. Four examples are now presented.

3. Some Examples

3.1 Example One: Thirteen data values

[2] SD_{within} therefore has a higher statistical uncertainty, or is less efficient, than *s* given the same number of data. This is a small price to pay and needs to be understood as such. *s* is inappropriate for control charting and process capability applications because it provides no leverage to examine the data for statistical control (see Example Four). Statistical control is *not* a natural state for a production process. Assuming statistical control (or process predictability) is to be avoided.

In this example, the production process is in operation every three to four weeks, and one data value per production run is judged appropriate. The specifications are LSL=8 and USL=12, and the process target is 10, the midpoint of the specifications. The X chart for the data is shown in Figure 2 (the mR part of the chart is not shown) along with a histogram and the computed C_p and C_{pk} statistics.

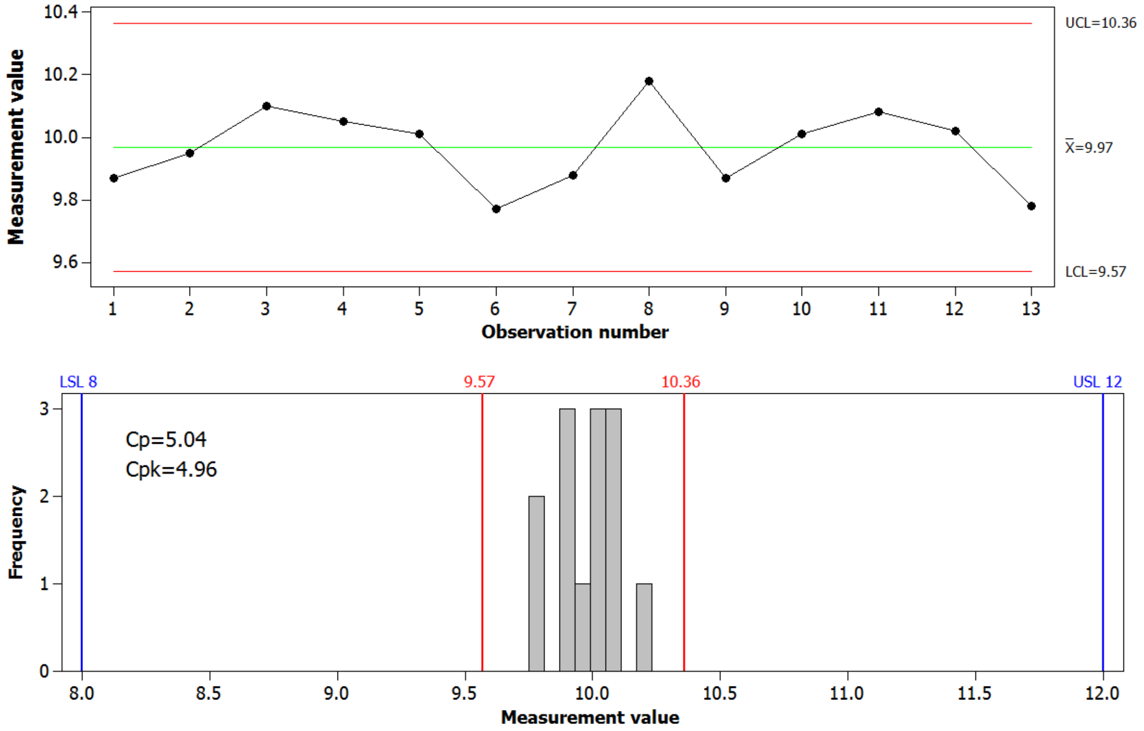

Figure 2: X chart and histogram (including specification limits, 3-sigma limits ("control limits") from the X chart and the computed C_p and C_{pk} statistics) for the data of Example One.

Thirteen data, which is not a lot of data, would represent some 9-12 months of production. With an effective number of d.f. of 7.7 (Wheeler, 2004) the uncertainty in SD_{within} is high at ~25.5%. The SD_{within} of 0.132 is hardly a precise estimate and more data would be ideal. But, do we need more precision in this case? Can we take a reasonable business decision in spite of the "high" uncertainty associated with SD_{within} and the capability statistics? From Figure 2 we ascertain that:

- No evidence of unpredictable behaviour is found in the X chart, i.e. we have some rationale to speak of a predictable process
- The variability in this process is small versus the requirement (the specifications of 8 to 12), as confirmed by the histogram and the C_p and C_{pk} statistics of 5.04 and 4.96 respectively
- There is no evidence of this process operating off-target (versus the target of 10)

While thirteen values is *not* a generically recommended number of data to analyse process capability, it seems that here we have enough data. (If not, how many more months or years would you wait before taking a decision?) Characterising this process as capable seems justified.

3.2 Example Two: Twenty data values

In this example, a production process has been operated over four days, and five values per day have been obtained. The twenty collected data $(4 \times 5 = 20)$ are behind Figure 3. (The specifications are LSL=30 and USL=35 with the target being 32.5, the midpoint of the specifications.)

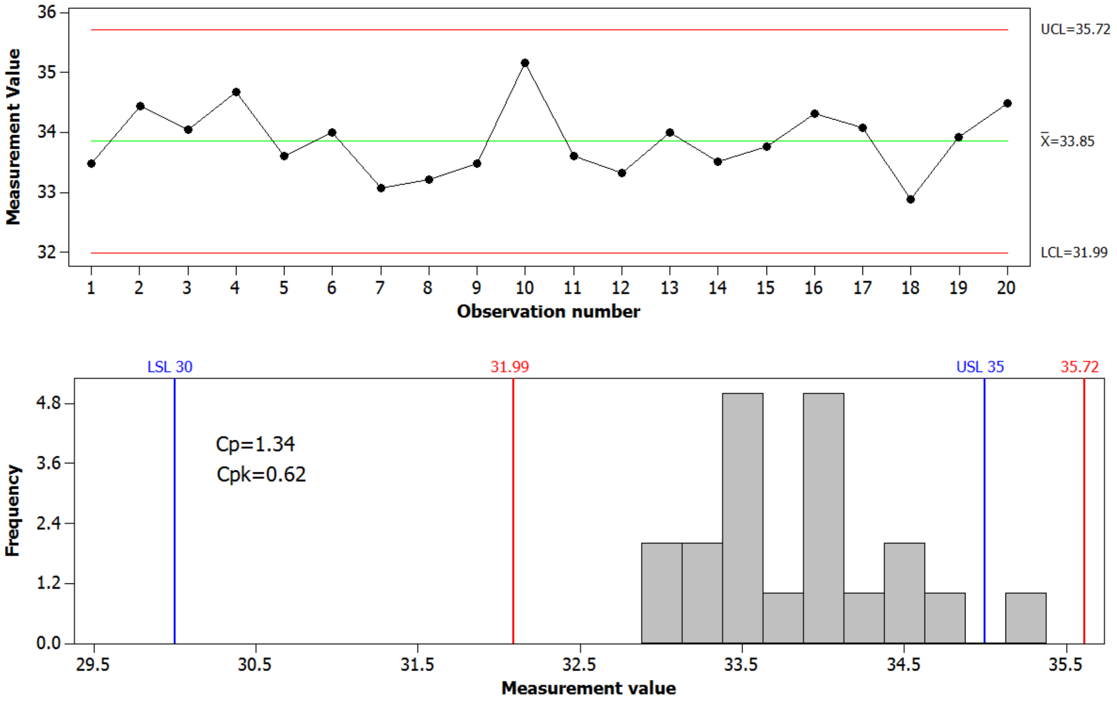

Figure 3: X chart and histogram (including specification limits, 3-sigma limits from the X chart and the computed C_p and C_{pk} statistics) for the data of Example Two.

These data provide rationale to conclude that the process:

- Can be characterised as predictable (consistent output both within and between days)
- Has been operated off-target (the histogram is well to the right of the target of 32.5)
- Has delivered one observed non-conforming unit (observation 10 on the X chart)
- Is in need of improvement (a C_{pk} of 0.62 is a strong indication of a need to rapidly improve things)

Do we not have enough data to justify action on the process (to improve things)? Waste has already been incurred and further waste should be expected unless action on the process is taken.

A first step would be to relocate the process average (to learn how to operate on-target). A second step would be to look at the capability situation again to see if the routine variation in the process – represented by $6xSD_{within}$ – needs to be reduced (if so, the need for some kind of improvement project).

3.3 Example Three: One hundred and twenty seven data values

Here, a total of 127 values were obtained over one long production run. Figure 4 shows the data which allow for the process to be characterised as reasonably predictable over this time period.

With an effective number of d.f. of 76.6 (Wheeler, 2004) the uncertainty in SD_{within} is ~8.1%. The SD_{within} of 0.178 is now a reasonably solid estimate – look again at Figure 1. (To reduce the uncertainty by half, to ~4%, one would need ~306 d.f., some 500 or so data. In many cases this number of data is just not feasible. In many of the other cases it would be unlikely that the process would be characterised as predictable after 500 data had been collected.)

With 127 values, do we have enough data to effectively analyse capability? In many cases, certainly yes. What if here, however, the team involved wanted to include at least two production runs in the analysis of capability? Then, the answer is no and a factor aside from the precision of the estimated

SD$_{within}$ takes greater importance. Good judgement along with context and knowledge of the manufacturing operation is needed to answer the question "*How many data?*" effectively. (What if the next production run were tomorrow, or six months from now? Wouldn't this change the answer?)

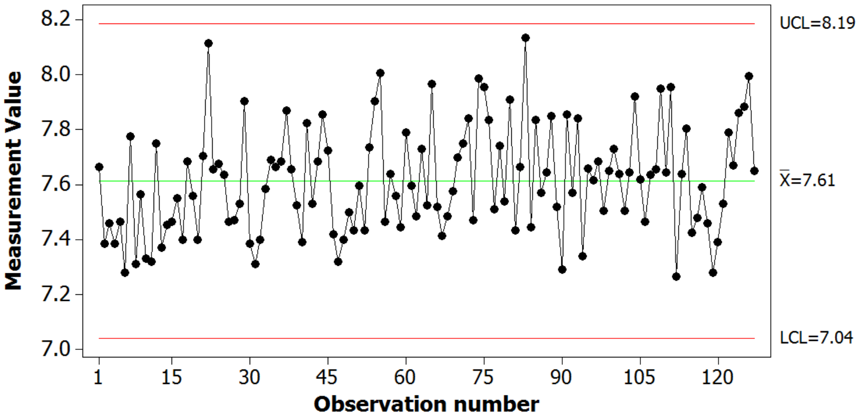

Figure 4: X chart for the one hundred and twenty seven data of Example Three.

3.4 Example Four: Thirty data values

Thirty data were recommended as a minimum to safeguard an analysis of process capability. The analysis of capability started only after thirty data had been obtained, over a period of almost two weeks and four different production runs. Figure 4 shows the data on an X chart. For the characteristic being measured, LSL=4.5, USL=5.0 and the target is 4.75.

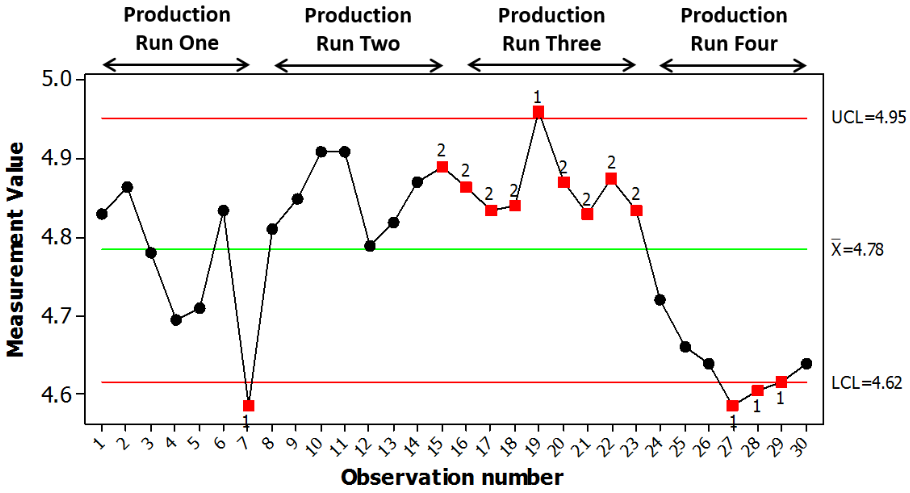

Figure 5: X chart for the thirty data of Example Four. (The detection rules used are "1" A data value outside the 3-sigma limits and "2" Eight or more consecutive values on one side of the central line.)[3]

Figure 5 leaves no option but to characterise the process as unpredictable (or not in statistical control). Predicting that practically all future values will fall in the range 4.62 to 4.95 is not credible, meaning that any computed capability statistics are estimates of what the process could be made to do (if brought into a predictable state) rather than well-defined, reliable indicators of future process performance.

[3] If "3-Standard deviation limits" (3-*s* limits) are (wrongly) used, all thirty values fall comfortably inside such limits.

If the User had looked at the data earlier he/she would have discovered the presence of assignable cause variation worthy of investigation (the X charts for the two points made below are not shown):

Point 1. Values 1 to 15 (i.e. production runs one and two) show observation number 7 as a signal of process change

Point 2. Values 1 to 23 (i.e. production runs one to three) show observation number 7 as a signal of process change and a run about the central line signalled from observation number 21

Aiming to polish an estimate of SD_{within} to furnish a better capability statistic was inappropriate here. An opportunity to learn from less than 30 data was available but not taken: The data were trying to inform the User that the process was changing when it shouldn't have been. If the User had investigated earlier the output from production runs three and four may well have been improved.

How does an analysis of capability help here? From the thirty data we have C_p=1.48 and C_{pk}=1.28. This tells us that *if* the process is brought into a predictable state and *if* it is operated on target, we can expect the process to be capable. Until the process is actually brought into a predictable state the capability statistics of 1.48 and 1.28 are purely hypothetical, i.e. what *could* be. We learn however that we do not need to make a fundamental change to the process – the challenge is to get the most from the existing process (improve its consistency over time) and not to think of introducing a new process!

4. Conclusions

Without context, 30 data is a good number of data. Figure 1 can be used to justify this. Nevertheless, context is important and context can lead to more or less data than 30 being appropriate.

When, then, do you have enough data? When the data you have collected are sufficient to justify any action you plan to take. Such actions include transferring a new technology to a factory, starting a production process in the expectation of fully compliant output, making some kind of corrective action to improve future process output, introducing new equipment or materials into the process, and so on.

Process capability provides insight that helps to take the right action on a process. Use too few data and the risk of a wrong conclusion is too high. Too many data and it might take too long to get to a conclusion. Both imply waste in that the right action, at the right time, does not take place.

Statistical theory plays a useful role in answering the "*How many data?*" question, but statistical theory alone is not enough. Judgement is also needed to get the most effective answer to the "*How many data?*" question. So, the first answer to the "*How many data?*" question will probably be "*it depends*". As more context enters the conversation a better answer should follow.

References

Bothe, D. R. (1997). Measuring Process Capability: Techniques and Calculations for Quality and Manufacturing Engineers. McGraw-Hill Inc., US.

Deming, W. E. (1975). On Probability As A Basis For Action. *The American Statistician*, 29(4), 146-152.

Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product. (pp. 302). New York: D. van Nostrand Company, Inc.

Wheeler, D. J. (2004). Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts. (pp. 80-83, 180-186 & 446). Second Edition, SPC Press, Knoxville, Tennessee.