How Much Data Do I Need to Calculate Control Limits?

One purpose of a control chart is to monitor a process to determine when a process change has occurred or there is a special cause of variation present. One way of doing this is to take data over time and, *when you have enough data*, calculate the averages and control limits. Then, assuming the process is consistent and predictable (in control), those averages and control limits are set. Future performance is judged against those set control limits.

So, when do you have enough data? Two samples are obviously not enough. Thousands of samples are probably a few too many. How do you decide when there are enough data? The month's publication takes a look at how you can determine this. And, of course, our old friend variation is a part of the process. There is always variation present - and you have to account for that uncertainty in determining how much data are enough.

In this issue:

Introduction

We will use a dataset with 200 samples to help us decide when there are enough data to set the control limits. 200 samples are definitely more than enough data, i.e., assuming the process is in statistical control. We will start with the individuals control chart and show the impact the number of samples has on the control limits. We will take a look at the entire 200 samples and then backtrack to take a look at smaller sample sizes.

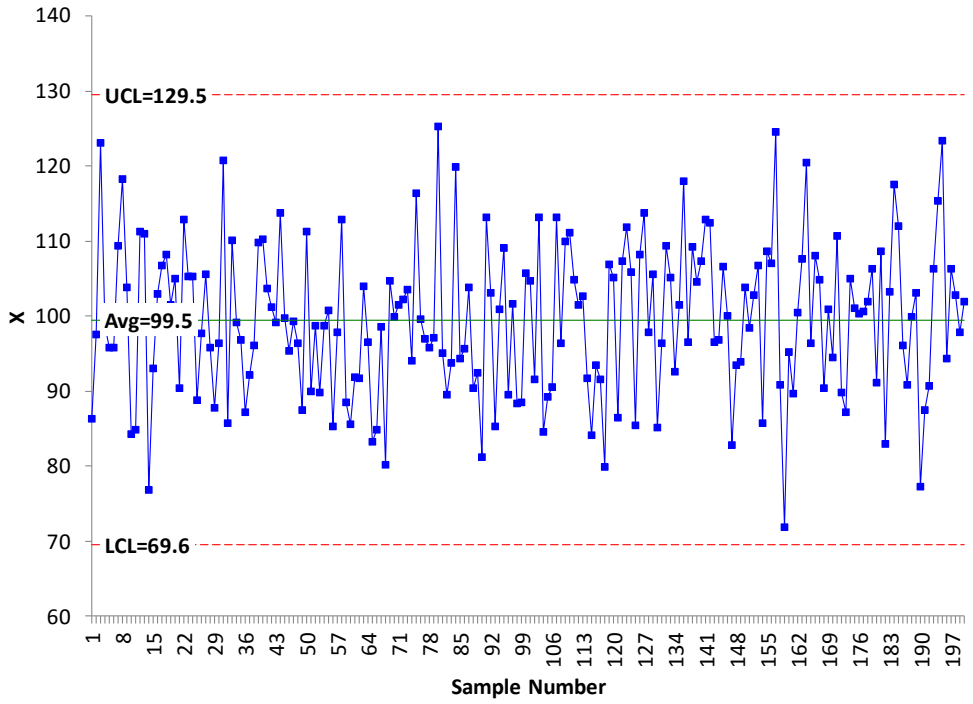
To generalize this process, two statistical terms will be introduced: degrees of freedom and the coefficient of variation (COV). Don't worry – it is done painlessly. Degrees of freedom is simply a measure of how much useful data you have. COV is a measure of the uncertainty in our data. There is a relationship between the degrees of freedom and COV that we will use to generalize the process of determining how much data you need to calculate control limits.

To complete the process, we need two things. One is a method of determining the degrees of freedom. There are simple equations to help us with this. Second, we have to determine how much uncertainty we want to live with. This will be up to you, but typically it is 15 to 20%.

Our Process

The first step in our analysis is to determine if the process (our dataset) is in statistical control. You can download the data set here if you would like. Figure 1 is the individuals control chart for our data.

Figure 1: X Chart for 200 Samples

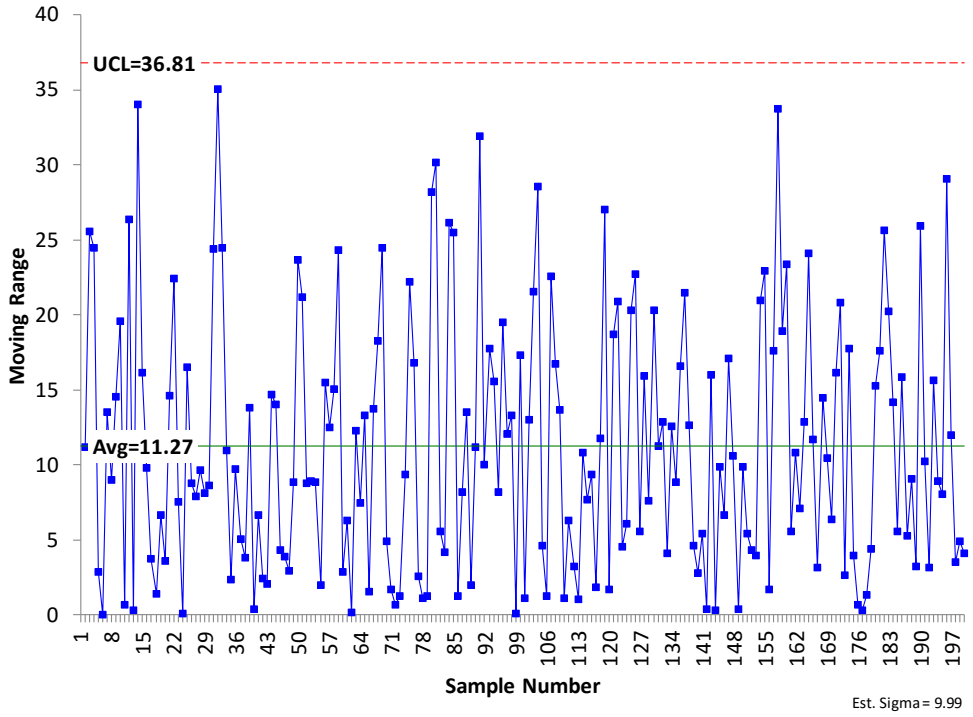


Each individual sample result is plotted. The average and control limits are calculated and added to the control chart. The process depicted in Figure 1 is in statistical control. There are no points beyond the limits and no patterns. It has an average of 99.5 with an upper control limit (UCL) = 129.5 and lower control limit (LCL) = 69.6.

The moving range chart for our data is shown in Figure 2. The moving range is the range between consecutive samples. The moving ranges are plotted on the control chart. There are 199 moving range values since there is no moving range for the first sample. The average moving range and UCL are calculated and added to the control chart. There is no LCL for this moving range chart. It is in statistical control as well. The average moving range is 11.27 with an UCL of 36.81.

The averages and control limits were calculated using the 200 samples. That is a lot of samples, so we are pretty confident in these values, particularly since the process is consistent and predictable – in statistical control. If we took another 200 samples from our process, we would not get exactly the same results – but they would be close. What if we only took 10 samples? Would the results be as close? Maybe, but probably not.

Figure 2: Moving Range Chart for 200 Samples



Think about Figure 2 – the moving range control chart. What variation is this control chart monitoring? It is monitoring the variation in consecutive samples. Thus, the average moving range is a measure of the variation in individual results – what we call the standard deviation or sigma. You can estimate sigma from the average moving range:

$$\sigma = \overline{R}/d_2 = 11.27/1.128 = 9.99$$

where d_2 is a control chart constant. For a moving range chart using the range between consecutive samples, d_2 is 1.128. Our estimate of sigma, the variation in the individual values, is 9.99. The average moving range is also used to calculate the control limits for the X chart and moving range chart:

X Chart:

$$UCL = \overline{X} + 2.66\overline{R}$$

$$LCL = \overline{X} - 2.66\overline{R}$$

mR Chart:

$$UCL = 3.27\overline{R}$$

The “quality” of our estimate of the average moving range definitely impacts how “good” our control limits are. So, how much data do we need to have “good” control limits?

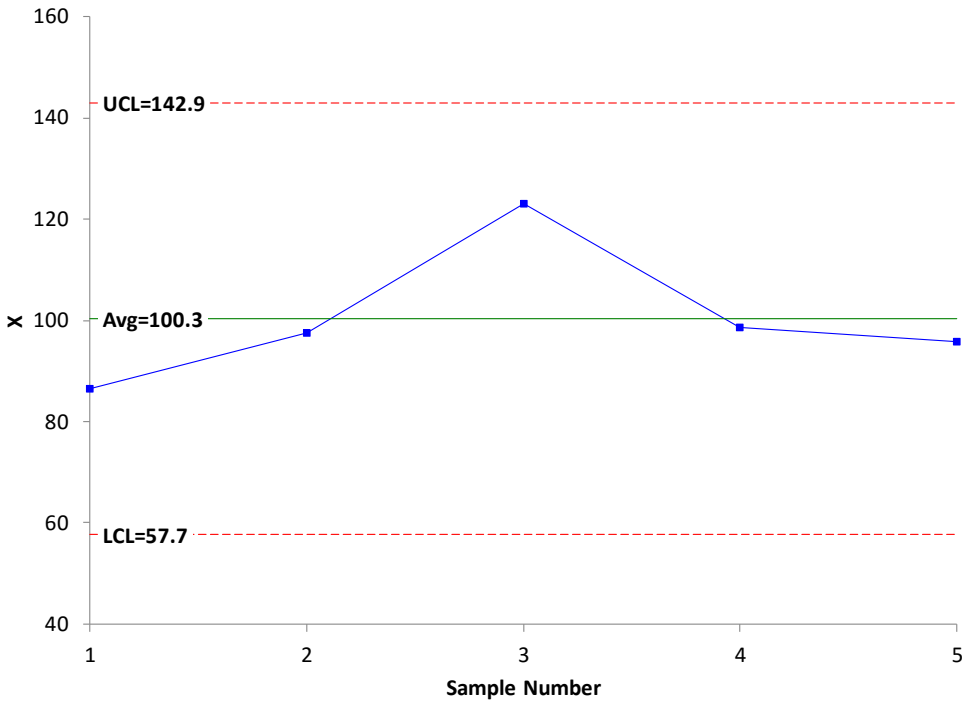
Impact of the Number of Samples on Control Limits

How do the control limits change over time as the number of samples change? One of our earlier publications stated that you can start an individuals control chart with as few as 5 points; recalculate the control limits with each new point and then lock the control limits in place after 20 points. Then recalculate after 100 points. Was that good advice? Let’s see.

Would you expect the control limits based on 5 samples to be the same as the one based on 200 samples? You probably wouldn’t. Which one has the more accurate control limits? The one with the most data – 200 samples - will give more accurate results.

Figure 3 shows the individual control chart based on the first five samples in the data set. The moving range chart will not be shown. The LCL and UCL in Figure 1 are 69.6 and 129.5, respectively. In Figure 3, the LCL and UCL are 57.7 and 142.9, respectively. The control limits based on 5 samples are much wider than the control limits based on 200 samples.

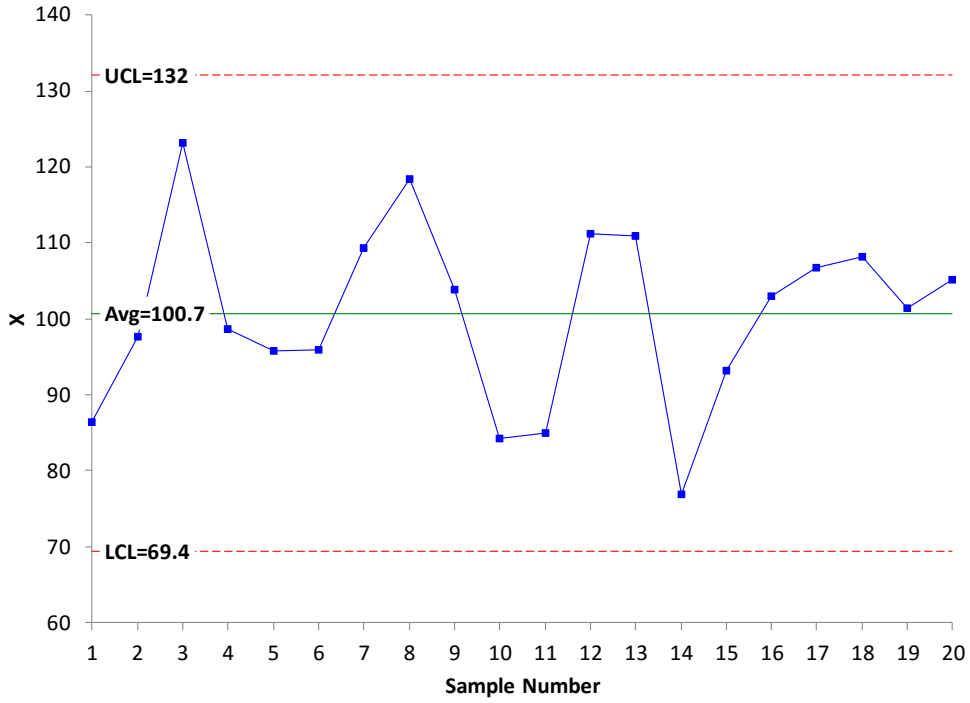
Figure 3: Control Chart Based on 5 Samples



Does this mean that you can't use a control chart with only 5 samples? Not at all. An out of control point on the control chart based on 5 samples is still a signal – a special cause of variation. You can start a control chart with 5 samples without a problem.

Let's increase our sample size to 20. Figure 4 is the X control chart for the first 20 samples. What has happened to the control limits? They have gotten tighter from the X control chart for 5 samples. For 20 samples, the LCL and UCL are 69.4 and 132, respectively.

Figure 4: X Chart for 20 Samples



What happens if we continue this process – adding more data to the control chart? What happens to the control limits? Eventually you reach the point of diminishing returns. You reach the point where the control limits don't change very much anymore. They do change, but not in a significant way.

Table 1 shows what happens as the number of samples included in the analysis increases. The columns in the table are for the average moving range, the overall average, the UCL, the LCL, and UCL – LCL. The last column shows how much difference there is between the UCL and LCL when compared to the control limits with 200 samples. For example, with 5 samples, the range between the UCL and LCL is 42.2% larger than for 200 samples.

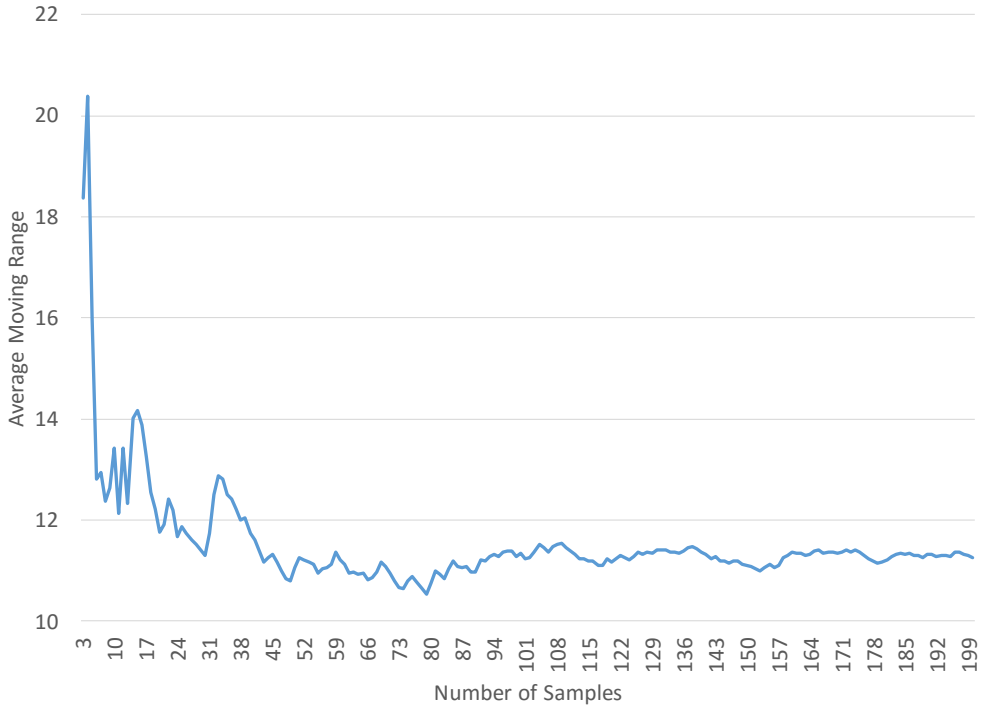
Table 1 shows that, after about 20 to 30 samples, the control limits don't change very much. At this point, there is little to be gained by continuing to re-calculate the control limits. The control limits have enough data to be "good" control limits at this point.

Table 1: Impact of Number of Samples on Control Limits

n	$\bar{R}$	$\bar{\bar{X}}$	UCL	LCL	UCL - LCL	% Diff
5	16.0	100.3	142.9	57.7	85.2	42.2%
10	13.4	101.3	137.0	65.6	71.4	19.1%
15	14.2	99.3	137.0	61.7	75.3	25.7%
20	11.8	100.7	132.0	69.4	62.6	4.5%
25	11.9	100.7	132.3	69.1	63.2	5.4%
30	11.3	100.0	130.1	69.9	60.2	0.4%
40	11.7	100.2	131.5	69.0	62.5	4.3%
50	11.1	100.3	129.7	70.9	58.8	-1.9%
100	11.3	98.7	128.9	68.6	60.4	0.7%
150	11.1	99.3	128.8	69.7	59.1	-1.3%
200	11.3	99.5	129.5	69.6	59.9	0.0%

Figure 5 shows the impact of sample size on the average moving range as the sample size increases. The chart shows how the average moving range begins to level off around 20 to 30 samples. This supports the conclusion above that 20 to 30 samples is sufficient.

Figure 5: Impact of Number of Samples on the Average Range



How do we apply this to other situations? The above analysis is for one set of data based on individuals control charts. But, what about other charts like the $\bar{\bar{X}}$-R chart? Are 30 data points enough there? Do

we have to go through the same analysis as above for each situation? No, you don't, but we need to talk about degrees of freedom and the coefficient of variation.

Degrees of Freedom and the Coefficient of Variation

Statistical terms often scare us. What do they really mean? We will keep it short and sweet here. The two terms we have to deal with are degrees of freedom and the coefficient of variation (COV). Why do we have to be concerned with these? It is because there is a relationship between the degrees of freedom and the coefficient of variation that allow you to determine when you have enough data for "good" control limits – regardless of the type of control chart you are using.

Degrees of freedom are a measure of how much useful data you. The larger the degrees of freedom, the less uncertainty you have in the results. Degrees of freedom is not the sample size (like 200 samples used in Figure 1). But, they are related. As sample size increases, the degrees of freedom increases.

Degrees of freedom are used to characterize the uncertainty in the estimated sigma – and thus the uncertainty in the calculated control limits. Degrees of freedom depend on the amount of data – and the formula being used to estimate sigma. Remember, with the individuals control chart, we calculate sigma using:

$$\sigma = \overline{R}/d_2$$

Degrees of freedom represent how much data we used to calculate the average moving range. The larger the degrees of freedom, the better our estimate of the average moving range and sigma.

The coefficient of variation (COV) is a measure of variation that describes the amount of variability relative to the mean. It is defined as:

$$\text{COV} = \text{sigma/mean}$$

COV, like degrees of freedom, is a measure of the uncertainty in sigma. There is a relationship between the two. For any estimate of sigma, the COV is inversely proportional to the square of two times the degrees of freedom:

$$COV = 1/\sqrt{2df}$$

where df is the degrees of freedom. Figure 6 is a plot of COV versus the degrees of freedom.

Figure 6 COV rapidly changes when the degrees of freedom are less than 10. In the range of 10 to 30, COV begins to level off. Above 30, the COV slowly decreases. At this point, there is a trade-off between increasing the sample size (and the degrees of freedom) and decreasing the uncertainty (COV). It takes a lot more samples to decrease the COV significantly at this point. 10 degrees of freedom corresponds to a COV of about 22%. 30 degrees of freedom corresponds to a COV of about 13%.

Figure 6: COV versus Degrees of Freedom

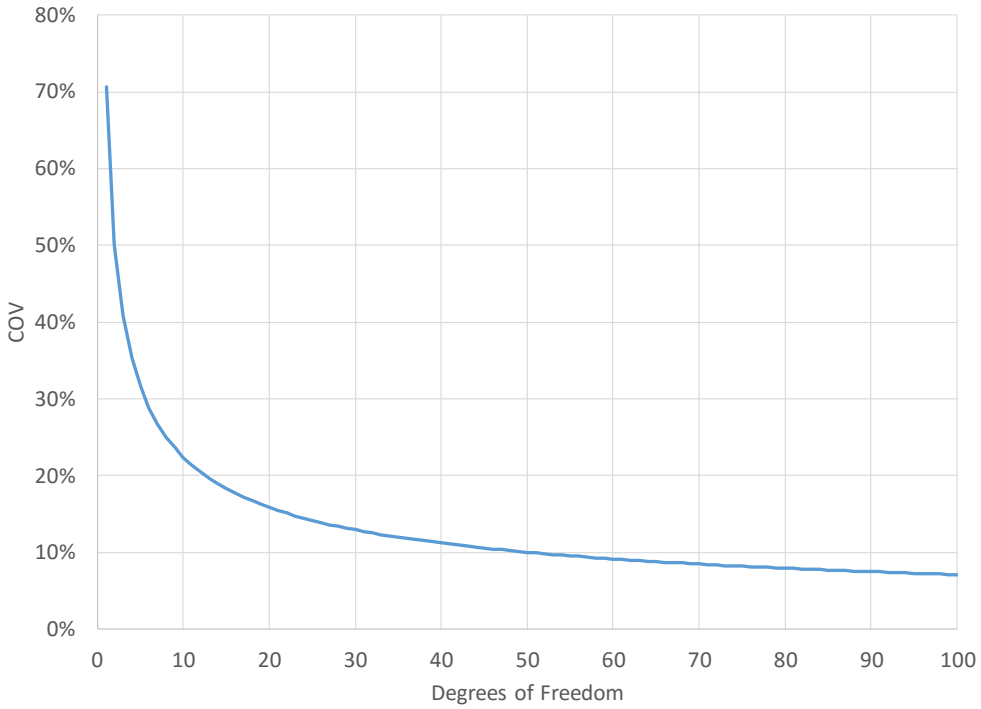


Figure 6 applies for any estimate of sigma, no matter how it was calculated. We just need a method of determining the degrees of freedom for the various methods for calculating sigma.

Determining the Degrees of Freedom

To calculate the COV, you need to be able to calculate the degrees of freedom. Dr. Donald Wheeler provides a method to easily do this in his book Advanced Topics in Statistical Process Control (www.spcpress.com). We will look at the individuals control charts and the $\overline{X}$-R control chart here.

The average moving range is used to estimate sigma with the individuals control chart. The degrees of freedom associated with the average moving range can be estimated using the following equation:

$$df = 0.62(n - 1)$$

where n is the number of samples.

For $\overline{X}$-R control charts, the average subgroup range is used to estimate sigma. The following can be used to obtain the degrees of freedom with the average subgroup range:

$$df = 0.9k(n - 1) \text{ for } n \text{ less than } 7$$

$$df = 0.85k(n - 1) \text{ for } n \text{ from } 7 \text{ to } 10$$

where k = number of subgroups and n = subgroup size. We will use these equations to find the degrees of freedom in the next section.

Using the Knowledge to Obtain "Good" Control Limits

We will start with the individuals control chart. Figure 4 was constructed using the first 20 samples. The degrees of freedom associated with the average moving range is:

$$df = 0.62(n - 1) = 0.62(20 - 1) = 0.62(19) = 11.78$$

The COV for 11.8 degrees of freedom is:

$$COV = \frac{1}{\sqrt{2df}} = \frac{1}{\sqrt{2(11.8)}} = 20.6\%$$

Figure 6 shows that this value of COV is just as the curve is beginning to level off.

Table 2 below shows the degrees of freedom associated with our sample sizes in Table 1 as well as the COV and control limits for the individuals control chart. The change in control limits decreases dramatically at a sample size of 20, corresponding to about 12 degrees of freedom. The COV at this point is 20.6%.

Table 2: Degrees of Freedom and COV for Various Sample Sizes

n	df	COV	UCL	LCL
5	2.48	44.9%	142.9	57.7
10	5.58	29.9%	137.0	65.6
15	8.68	24.0%	137.0	61.7
20	11.78	20.6%	132.0	69.4
25	14.88	18.3%	132.3	69.1
30	17.98	16.7%	130.1	69.9
40	24.18	14.4%	131.5	69.0
50	30.38	12.8%	129.7	70.9
100	61.38	9.0%	128.9	68.6
150	92.38	7.4%	128.8	69.7
200	123.38	6.4%	129.5	69.6

Looking at Figure 6, it appears we may want to go a little further than 20.6% on the COV. Suppose we decide that 15% is a good value. You can rearrange the COV equation to calculate the degrees of freedom you need for a COV = 15%:

$$df = 1/(2COV^2) = 1/(2(.15)^2) = 22.2$$

A COV of 15% corresponds to 22.2 degrees of freedom. Now, you can use the degrees of freedom formula for the individuals chart to determine how many samples you need:

Rearranging df = 0.62(n - 1):

$$n = (df/.62) + 1 = (22.2/.62) + 1 = 36.8$$

With an individuals control chart, you will need about 37 to 38 samples to get a COV of 15%.

Suppose you are subgrouping your data and using an $\overline{X}$-R control charts. You are taking five samples per hour and forming a subgroup with those 5 samples. In this case, n is 5. Suppose you want to find how many subgroups you need to have a COV of 15%. This occurs at about 22.2 degrees of freedom – same as for the individuals control chart above. The following calculations show how to determine how many subgroups you need.

$$df = 0.9k(n - 1) \text{ for n less than 7}$$

$$22.2 = 0.9(k)(5 - 1) = 3.6(k)$$

$$k = 22/3.6 = 6.2$$

About six subgroups of size 5 (30 total samples) will give you a COV of 15%.

Suppose you want a COV of 10%. That corresponds to 50 degrees of freedom. The number of subgroups required for a COV of 10% is about 14, a total of 80 samples. You almost triple the number of samples you need while only reducing the COV by a third.

So what COV should you use to determine when you have enough data? That decision is up to you. But, from Figure 6, something in the range of 15% to 20% probably is a good trade-off between samples needed for "good" control limits and minimizing the uncertainty.

Summary

This publication has looked at how much data you need before you have "good" control limits. It was shown how the averages and control limits change for an individuals control chart as the sample size increases – and that there is point where adding additional data to the control limit calculations has very little impact. It was also shown that there is a critical relationship between the coefficient of variation (COV) and the degrees of freedom. This relationship can be used to determine how many degrees of freedom you need for a certain COV. And, from this, how to determine how many samples you need for the individuals and $\overline{X}$-R control charts.

In the end, the decision is really yours. But the guidelines presented in this publication will help determine how much data you need when working with individuals and $\overline{X}$-R control charts. It is based on finding the data that will give you an appropriate value for COV. This is usually in the range of 15 to 20%.

As mentioned before, our earlier publication said you can start an individuals control chart with as few as 5 points; recalculate the control limits with each new point and then lock the control limits in place after 20 points. Then recalculate after 100 points. Was that good advice? For the most start, it is good advice. I might change the value to 30 points to lock the limits the first time.

And, this same approach will work in determining how many samples you need for a process capability analysis. Very often, people use 30 samples for a process capability analysis. What is the uncertainty for 30 samples? You now know how to determine that.

Quick Links

Visit our home page
SPC for Excel Software
SPC Training
SPC Consulting
SPC Knowledge Base
Ordering Information

Thanks so much for reading our publication. We hope you find it informative and useful. Happy charting and may the data always support your position.
Sincerely,

Dr. Bill McNeese
BPI Consulting, LLC