Consistency Charts

SPC for measurement systems

Donald J. Wheeler

What happens when we measure the same thing and get different values? How can we ever use such a measurement system to measure different things? By thinking of measurements as the product and the measurement procedure as the process we can use the techniques of SPC to address the problems of complex measurement systems.

A CONSISTENCY CHART

Churchill Eisenhart, a famous statistician who worked at the National Bureau of Standards, once wrote that a measurement system cannot be regarded in any logical sense as measuring anything at all until it has attained a "state of statistical control." As I showed in "Three Questions for Success" (***QDD***, March 1, 2011) a process behavior chart is an operational definition of how to get the most out of any process. The measurement process is no exception.

A traditional way of evaluating a measurement process has been to repeatedly measure the same thing and then to analyze the resulting measurements. When such repeated measurements are placed on an *XmR* chart we end up with what I call a Consistency Chart. This simple chart allows you to make judgments about the consistency, precision, and bias of the measurement system. This use of process behavior charts is distinctly different from the use of an Average and Range Chart to evaluate the effects of operators and instruments upon a measurement operation as described in "A Better Way to Do R&R Studies" (***QDD***, February 1, 2011). Some examples follow.

X	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*
9,999,591	—	9,999,599	5	9,999,593	4	9,999,595	4	9,999,601	3
9,999,600	9	9,999,597	2	9,999,598	5	9,999,598	3	9,999,603	2
9,999,594	6	9,999,599	2	9,999,599	1	9,999,592	6	9,999,593	10
9,999,601	7	9,999,597	2	9,999,601	2	9,999,601	9	9,999,599	6
9,999,598	3	9,999,602	5	9,999,600	1	9,999,601	0	9,999,601	2
9,999,594	4	9,999,597	5	9,999,599	1	9,999,598	3	9,999,599	2

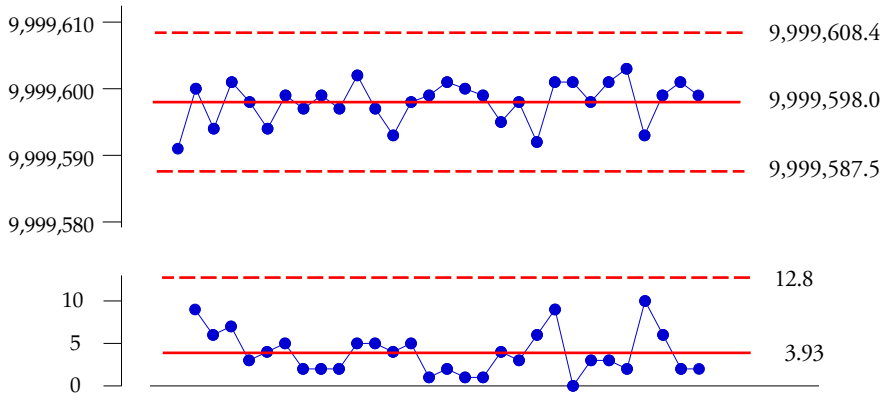

Figure 1: A Consistency Chart for 30 Weighings of NB10

Around 1940 the National Bureau of Standards obtained a ten-gram weight made out of chrome-steel alloy. This standard weight was designated NB10 and was weighed once each week by one of two individuals. The values shown below come from for the first 30 weeks of 1963. The values are shown in micrograms, so that a full 10 grams would be listed as 10,000,000.

The *XmR* Chart for these 30 weighings is shown in Figure 1. There we see that the procedure for weighing NB10 shows a reasonable degree of consistency. Since ten grams is approximately the weight of two nickels, we can see that the scale used here is rather good. It is measuring this small weight to one part in 10 million, and it is doing so consistently, week after week. However, the average value for the *X* chart is not equal to a full 10 grams. The fact that the *X* chart is not centered on 10,000,000 represents a bias. We would estimate this bias to amount to 402 micrograms. However, since this is the 10 gram standard being measured by the master measuring technique, it is difficult to say where the bias resides. Is the standard itself off by 402 micrograms, or is the scale biased by 402 micrograms, or is it a combination of both problems? We cannot determine the answer from these data alone.

Note that bias is always a relative concept. A measurement system is said to be biased if repeated measurements of the same thing yield a different average than is obtained when that same item is measured by a master measuring technique. As this example makes clear, in practice there is no such thing as an "actual value." There is only the value you obtain when using a master measurement technique repeatedly.

Next, the Consistency Chart shows that this measurement system is consistent over time, and the range chart allows us to estimate the measurement error for a single determination of the weight of this standard (commonly known as the precision of the measurement). The average moving range is 3.93 micrograms. When we divide by the d_2 bias correction factor of 1.128 we obtain an estimate of the standard deviation of the measurement system of:

$$\textit{Estimated Standard Deviation of Measurement System} = 3.48 \text{ micrograms}$$

Since this value estimates the square root of the average of the squared deviations from the average, it is virtually guaranteed to create headaches if you try to explain it to others. For this reason I use the Probable Error. The Probable Error is the median amount by which a measurement will differ from the average of repeated measurements of an item. The Probable Error is estimated by simply multiplying the estimated standard deviation by 0.675.

$$\textit{Probable Error of Measurement System} = 2.3 \text{ micrograms}$$

Both the estimated standard deviation for the measurement system and the probable error contain the same essential information, but the Probable Error is much easier to explain and use. In this case, the measurements were made to the nearest microgram. This is the measurement increment used. However, the Probable Error of tells us that half the time these measurements will err by two micrograms or less, and half the time they will err by three micrograms or more. Thus, the Probable Error effectively defines the inherent uncertainty in any measurement. As in this case, we generally prefer for the measurement increment to be about the same size as, or slightly smaller than, the Probable Error. The Probable Error provides the easiest way to characterize the precision of a measurement system. It tells us how aggressively to interpret the measurements, and it answers the question about how many digits to record.

So, what did we learn from the consistency chart? We found this measurement system to be

consistent (that is predictable), possibly biased by 402 micrograms, and we estimated the inherent uncertainty in these measurements to be about 2 micrograms.

Clearly, in the case of repeated measurements of the same thing, we should like to see a reasonable degree of predictability when the measurements are placed on a consistency chart. But what happens when the measurement system does not display any reasonable degree of consistency?

A RUBBER RULER

Consider the case of a vision system created to measure the effective diameter of the steel insert for the rim of a steering wheel. The inserts were formed by bending a mild steel rod into a circle and welding the ends together. The vision system consisted of a back-lit plate with locating pins to hold the insert in place, a video camera mounted above the plate, and a computer that would count the pixels inside the image of the insert. Once the number of pixels was known the computer would convert the area into an effective diameter (in inches) for the insert.

When shown this wonderful, new, fancy measurement system Richard Lyday decided to check it for consistency by repeatedly measuring the same insert. Since positional variation was part of the measurement system, he measured the part, took it off, reloaded it, and measured it again. After 30 such repeated measurements carried out over the course of an hour he had the data and chart in Figure 2.

X	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*
13.383	—	13.404	0.027	13.481	0.025	13.280	0.250	13.659	0.079
13.383	0.000	13.431	0.027	13.508	0.027	13.582	0.302	13.632	0.027
13.354	0.029	13.453	0.022	13.530	0.022	13.332	0.250	13.682	0.050
13.429	0.075	13.429	0.024	13.506	0.024	13.605	0.273	13.655	0.027
13.404	0.025	13.305	0.124	13.506	0.000	13.580	0.025	13.659	0.004
13.431	0.027	13.506	0.201	13.530	0.024	13.580	0.000	13.634	0.025

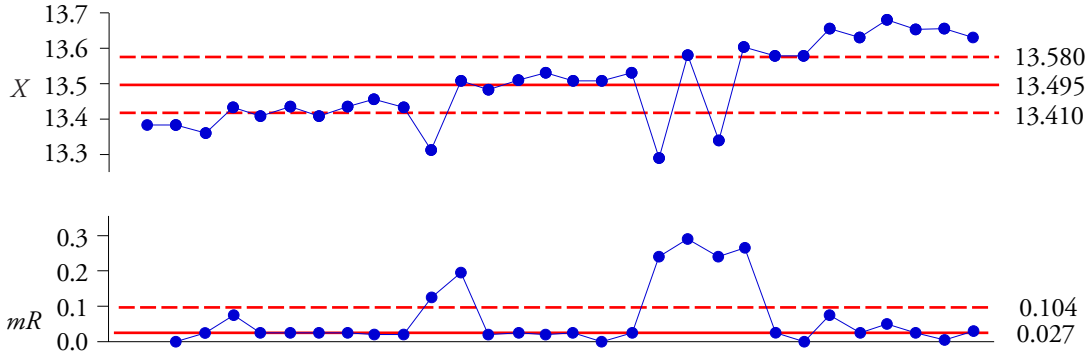

Figure 2: *XmR* Chart for Vision System Measurements

This measurement system is a rubber ruler. Since the steel insert could not have grown a quarter-inch in diameter, the trend on the *X*-Chart must be interpreted as a problem with this measurement system. This trend was explained when they discovered that the pixel size in the image shrank as the camera warmed up. Since the computer did not actually determine the size of each pixel, it was fooled by this pixel drift into computing larger diameters over the course of this hour.

In addition to the trend, there was also the problem of the stalagmites and stalactites on the

XmR Chart. These occurred when the vibration of the camera was sufficient to blur the image so that the computer lost count. With an 800-ton press in the building, and with the camera mounted on a wooden roof truss, this vibration was going to be part of the operating environment for this measurement system.

Many analytic procedures and laboratory tests are rubber rulers. The complexity of the procedures, plus the variation in the elements of the test, combine to give test results that vary over time, or even from test to test. So how can we use a rubber ruler to measure anything? The common solution is to effectively recalibrate the ruler every time you use it.

In the case of the Vision System this would require that the computer be reprogrammed to measure both the insert and, at the same time, a fixed reference circle on the back-lit plate. Since this device actually measures areas, the measured area for the reference circle could be compared to the known value for the area of that circle, and the measured area for the insert could then be scaled up or down accordingly.

The problem with the vibration could be solved by comparing the area for the reference circle with the previous value for the area for the reference circle. A difference in area that would result in a moving range that exceeded the upper range limit would be a reliable indicator of a bad reading.

Thus, while this vision system is a rubber ruler, the consistency chart has identified two inherent problems with the vision system that could be remedied by appropriate adjustments to each measurement.

The effective diameters are reported to the nearest 0.001 inch. But how good are these values really? For the vision system, the median moving range in Figure 2 is 0.027 inches. Dividing by the d_4 value of 0.954 we obtain an estimate of the standard deviation of the reported values of:

$$\textit{Estimated Standard Deviation of Vision System} = 0.028 \text{ inches}$$

This value results in a Probable Error of:

$$\textit{Probable Error of Vision System} = 0.019 \text{ inches}$$

So, while these values are reported to the nearest mil, they are good to the nearest 19 mils. Half the time the adjusted value will err by 19 mils or more, and half the time the adjusted values will err by 19 mils or less. Thus, there is no point in reporting the effective diameter to the nearest mil. This instrument is reporting one digit too many. The readout should be changed to show only two decimal places.

CONTINUED TRACKING

In the vision system example we see how the Consistency Chart may be used for a spot check of a measurement system. Since this spot check revealed problems with the measurement system, they knew better than to use the measurements to operate their process. The next time I saw this vision system it was covered with dust, so presumably they had gone back to a physical measurement of the inserts.

In the NB10 example we see how a Consistency Chart may be used to track the consistency of a measurement process over time. As long as the measurement system is operated consistently the observed values should continue to fall within the limits on this chart. The data for Weeks 31 to 60 are shown in Figure 3, along with the Consistency Chart for Weeks 1 to 60. The limits

shown are the limits computed using Weeks 1 to 30. For simplicity the initial 9,999 that is common to every number has been dropped from the graph in Figure 3.

X	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*	*X*	*mR*
9,999,597	2	9,999,594	17	9,999,601	10	9,999,596	5	9,999,589	6
9,999,600	3	9,999,594	0	9,999,598	3	9,999,598	2	9,999,590	1
9,999,590	10	9,999,598	4	9,999,593	5	9,999,596	2	9,999,590	0
9,999,599	9	9,999,595	3	9,999,594	1	9,999,594	2	9,999,590	0
9,999,593	6	9,999,595	0	9,999,587	7	9,999,593	1	9,999,599	9
9,999,577	16	9,999,591	4	9,999,591	4	9,999,595	2	9,999,598	1

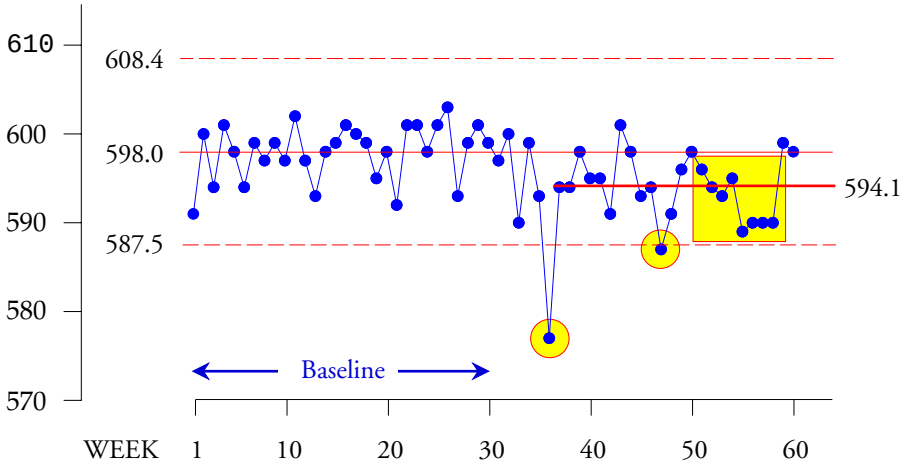

Figure 3: A Consistency Chart for Weighings of NB10, Weeks 1 to 60

Clearly something happened in Week 36. Since presumably the chrome steel weight did not suddenly become 20 micrograms lighter one week and then gain this weight back the following week, we have to conclude that the signal in Week 36 pertains to the measurement system used. Following the upset in Week 36 the running record is no longer centered on the original central line of 598.0. Another point falls below the lower limit in Week 47. Then in Weeks 51 to 58 we have a long run below the central line. Taken together, there is a shift in the values recorded for NB10 following Week 36. The average weight recorded for weeks 37 through 60 is 9,999,594.1 micrograms, which is 4 micrograms less than the average for weeks 1 to 30. This additional four microgram bias is probably related to a change in the measurement system.

A FURTHER EXAMPLE

When a Consistency Chart is created using a known standard it is possible to evaluate the bias of your measurement system. While this was illustrated with the NB10 data, the following example will illustrate further details of the comparison between the observed average and the accepted value. Each Monday morning Test Method 56 is used to test a known standard that has an accepted value of 40. The results are placed on a Consistency Chart. The data and chart for January through June are shown in Figure 4.

Date	1/7	1/14	1/22	1/28	2/4	2/11	2/18	2/25	3/4	3/11	3/18	3/25	
Value	37.7	40.6	41.0	39.5	40.9	42.6	41.7	38.2	42.1	38.5	36.4	40.6	
mR	–	2.9	0.4	1.5	1.4	1.7	0.9	3.5	3.9	3.6	2.1	4.2	
Date	4/1	4/8	4/15	4/22	4/29	5/6	5/13	5/20	5/28	6/3	6/10	6/17	6/24
Value	40.7	39.8	39.8	38.9	39.4	38.4	41.3	36.7	41.5	36.8	40.4	39.9	39.3
mR	0.1	0.9	0.0	0.9	0.5	1.0	2.9	4.6	4.8	4.7	3.6	0.5	0.6

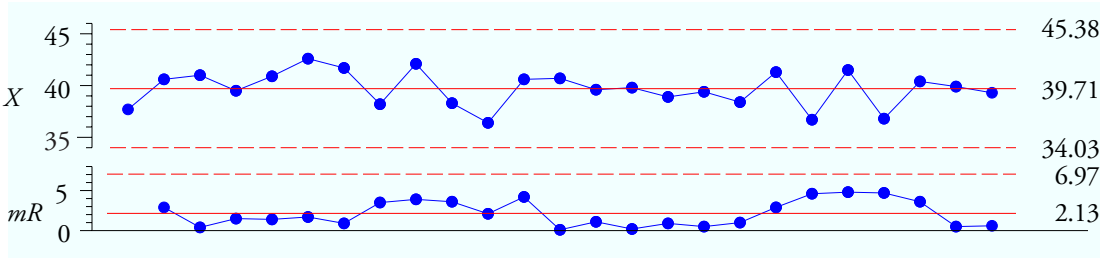

Figure 4: Consistency Chart for Test Method 56

Here we find that Test Method 56 has been operated consistently over the past six months. The average moving range of 2.13 units translates into an estimated Probable Error of 1.3 units. Inspection of the data in Figure 4 will show that the measurement increment used was 0.1 unit. Since a measurement will err by 1.3 units or more at least half the time, there is no point in recording the values to a tenth of a unit. The measurements obtained by Test Method 56 could be rounded off to the nearest whole number without any appreciable degradation in the quality of the measurements.

Is Test Method 56 biased? A simple way to graphically answer this question would be to shift the central line and limits for the *X* chart to be centered on the accepted value for the known standard. This is done in Figure 5.

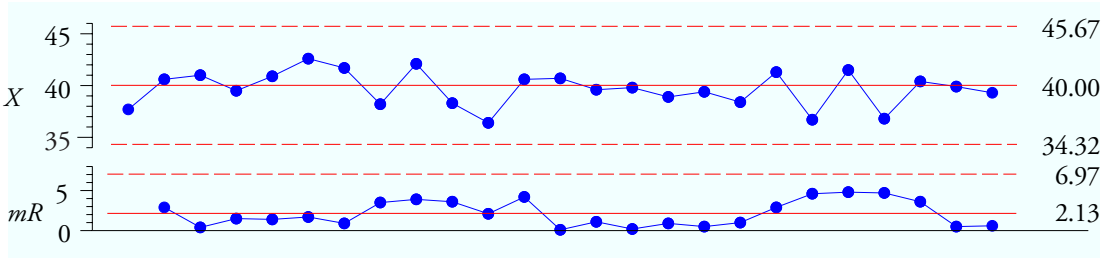

Figure 5: Consistency Chart for Test Method 56 Centered on the Accepted Value for the Standard

If the shift in limits shown in Figure 5 does not result in points outside the limits, any long runs above or below the central line, or any other run-test signals, then there is no detectable bias in the measurement system. Moreover, the chart in Figure 5 could be used with additional tests of the standard. Any signals found on this chart would indicate a change in the measurement system that results in biased measurements.

For those who are compulsive, we could easily obtain a 90% interval estimate for the expected value of repeated measurements of the standard. The observed average from Figure 4 is 39.708. The global standard deviation statistic is 1.694, and it has 24 degrees of freedom. The

appropriate t-statistic is 1.711. So our 90% interval estimate for the expected value of repeated measurements is:

$$\textit{Average} \pm t_{.05}\frac{s}{\sqrt{n}} = 39.708 \pm 1.711\frac{1.694}{\sqrt{25}} = 39.708 \pm 0.580 = 39.13 \text{ to } 40.29$$

Since this interval contains the accepted value of 40, we can say that Test Method 56 displays no detectable bias in the neighborhood of 40.

THE QUETELET FALLACY

A traditional way of assessing measurement error is to collect 30 measurements of a single item and then use the average to characterize bias (when the item measured is a known standard) and to use the standard deviation statistic to characterize the precision of the measurements. This approach dates back to the 1840s when it was used by the Belgian sociologist and statistician Adolphe Quetelet.

Unfortunately, this approach makes a very strong assumption about the data. It assumes that the data are homogenous. When this assumption happens to be correct this approach will work. However, when the data are not homogenous this approach breaks down and the results are misleading. Since this approach does not examine the data for homogeneity, its use is known as the Quetelet Fallacy.

This fallacy was recognized by Sir Frances Galton in 1875 and a search for a way around this problem was sought. By 1925 the foundations of modern statistical analysis were firmly in place. In order to avoid the Quetelet Fallacy the data would be organized and analyzed in a manner that would allow any lack of homogeneity to show up *without contaminating the estimate of dispersion used*. This approach always uses the within-subgroup, or short-term, variation in place of a global measure of dispersion.

For the NB10 data in Figure 1 the Consistency Chart shows these data to be homogeneous. The average characterizes the bias as being 402 micrograms, although we do not know if this bias belongs to the measurement system or the standard itself. The average moving range gives us an estimate of the Probable Error of 2.3 micrograms. Since these data are homogeneous, we might have blindly come to the same conclusion about this measurement system using Quetelet's approach. (The global standard deviation statistic of 3.15 micrograms results in an estimated Probable Error of 2.1 micrograms.)

The Consistency Chart in Figure 2 shows the vision system data to be very nonhomogeneous. Since the part used was not a known standard, no characterization of bias was available. The median moving range of 0.027 inches gives a short-term estimated standard deviation of 0.028 inches. Using Quetelet's approach the global standard deviation statistic is 0.114 inches, which is four times too large.

But didn't the interval estimate for checking for bias with the data from Figure 4 use the global standard deviation? Yes, it did. However, these data are homogeneous. *The only time a t-test makes sense is when the data are homogeneous. If the data are not homogeneous, there is no well-defined parameter to test or estimate*. To understand this, figure out what is the average diameter you are likely to get for future tests of the test insert in Figure 2.

The foundation of all of modern statistical *analysis* is the use of within-subgroup or short-

term variation to filter out the noise in order to detect potential signals within the data. When you use any global measure of dispersion prior to having qualified your data as being homogeneous you automatically become a disciple of Quetelet, and you will be liable to suffer the effects of the Quetelet Fallacy. Many scientists have fallen into this logical inconsistency. Many techniques used in industry (such as Levey-Jennings charts) make this mistake. Let the user beware. Trust no one who is a follower of Quetelet.

SUMMARY

When characterizing a measurement system in an absolute sense there are three properties of interest. These are consistency, bias, and precision. The Consistency Chart allows you to assess each of these characteristics.

Lest you think I deliberately used the worst part of the NB10 data, Figure 6 adds the points for Weeks 61 to 100 to those already shown. In addition to the problem in Week 36, there were other problems in Weeks 63, 85, 86, 87, 88, and 94.

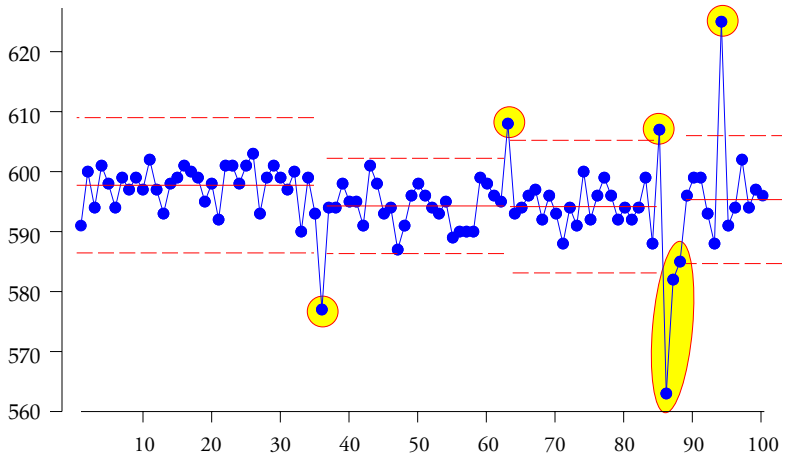

Figure 6: Consistency Chart for NB10 Weeks 1 to 100

If technicians weighing standards at the Bureau of Standards cannot operate a measurement system predictably, what do you think is happening in your labs and production operations? If you assume that your measurement systems are being operated predictably, you will probably be wrong. The only way to establish and maintain measurement consistency is by means of a Consistency Chart. Anything else is just wishful thinking.