

Rational Sampling

More about the conceptual foundation of process behavior charts

Donald J. Wheeler

While the computations for a process behavior chart are completely general and very robust, the secret to using a process behavior chart effectively lies in the art of rational sampling and rational subgrouping. Rational subgrouping was the topic last month. Here we shall look at the broader topic of rational sampling and the bearing it has upon the effective use of process behavior charts.

THE FIRST AXIOM OF DATA ANALYSIS

No data have meaning apart from their context. Outside of the statistics classroom, data are always generated by a process or system. They may be the result of an experiment where we manipulate inputs and observe outcomes, or they may be the simple by-product of ordinary operations where the values are observed as they occur. Regardless of the origin of your data, you have to know their context before you will know how to analyze and interpret them. Thus, the collection of data, the analysis of those data, and the interpretation of the results of the analysis will all depend upon the context for the data.

Rational sampling refers to collecting, organizing, analyzing, and interpreting data in a manner that makes sense in the context of those data. It is the structures within the data that define the interesting questions to be asked by the analysis. With experimental data the structure comes from the organization of the experiment. With observational data the structure comes from the who, what, when, where, why, and how of the data collection process.

Experimental data are structured to facilitate certain comparisons. This makes their analysis rather straightforward. Observational data are much less structured and this makes their analysis more complicated. As we saw last month, some data sets can be organized in various ways, and these organizations are not always equally useful. This makes the analysis of observational data more of a skill that has to be developed than a set of computations to be mastered. As before, I will illustrate the principles of rational sampling by means of a series of examples.

BALL JOINT SOCKET REVISITED

One day about five months into the production run of ball-joint sockets I happened to see Dave the supervisor. He pulled me in into his office and showed me the average and range chart in Figure 1. Now the operator was keeping the chart, and they were still operating with one mold, on one press, with one shift, while making 4000 parts each day. Since this chart started with the return from the Christmas break, they began by sampling four parts each hour for a couple of days, then as the process appeared to be operating as expected, they shifted to

collecting four parts twice each day. The chart in Figure 1 shows a process that is being operated predictably and within specifications, and this has been the case for the past five months.

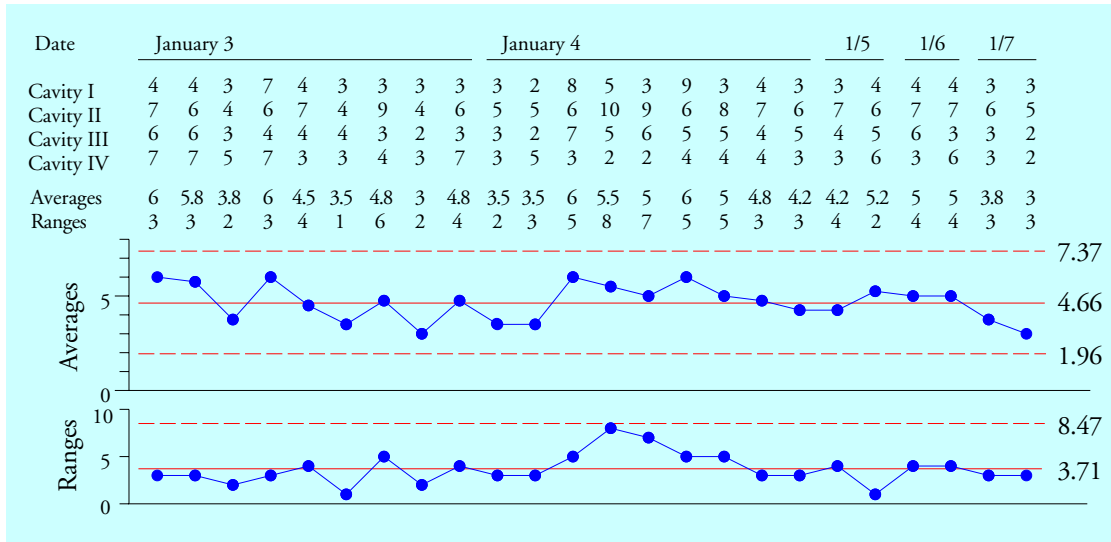


Figure 1: Average and Range Chart for Ball Joint Socket Production Data

The organization in Figure 1 is similar to the first organization used last month where each subgroup consisted of parts from all four cavities. This organization was used here because it was more convenient for the operator than other organizations. The objective for the charts had changed from certification to production. Originally this subgrouping was problematic because it assumed the cavities to be alike when they were actually quite different. After Dave adjusted the mold so that the four cavities were essentially the same, the operator could subgroup all four cavities together without getting stratified subgroups. While it is generally a bad idea to subgroup parallel operations together, they got away with doing it here because they took the time to make the four cavities alike.

How can they justify looking at only eight out of 4000 parts each day? The key to selecting the right subgroup frequency lies in understanding how quickly the process can change. Here, after the certification run, and with five months of production behind them, they felt that one sample every four hours of operation was sufficient. This choice of subgroup frequency is a judgment call based on process knowledge. When the process changes slowly, then two subgroups a day is a reasonable subgroup frequency.

THE PRODUCT NOT MEASURED

If the product stream is indeed homogeneous, then any portion of that stream will look like any other. If the product stream is changing, then any portion of that stream will eventually look different from some other portion. And, as we learned last month, the average and range chart compares the homogeneity within the subgroups with the homogeneity between the subgroups. So when the process appears to be operated predictably, the product not measured can be reasonably characterized by the product that was measured. Not only are the 96 data on the chart within specs, but the whole production run of 20,000 parts for this week can be said to be

100% conforming. The extrapolation from the data obtained to the product not measured is justified by the predictable operation of this process.

THE IMPORTANCE OF RATIONAL SAMPLING

In 45 years of teaching statistics I have repeatedly seen students lose sight of the purpose of the analysis by getting lost in the computations. One of the best ways to overcome this problem is through the use of good graphs. Perhaps the most egregious example of this occurred when a company in Connecticut did an experiment in which they made 30 batches of compound, molded these batches into parts, and coated the parts. Each of the 30 batches was processed in a slightly different way according to the factors being studied. At the end, 40 parts were taken from each batch of parts and carefully measured for the property of interest. Thus, they had 1200 data to analyze.

For the 40 parts from each batch they calculated all of the descriptive statistics they could think of, and also drew a histogram for each batch of 40 parts. (This was back in the 80s when we still did this sort of thing by hand.) So they ended up with a one-page summary for each batch. Each of these 30 pages contained the histogram, the data, the descriptive statistics, the levels of the experimental factors, and the relevant processing information about each batch.

After taking the time and trouble to collect 1200 experimental data they could not find any clear signals in the data, so they contacted me for consultation about their "analysis." When I agreed they sent me the stack of 30 pages. As an afterthought, they hurriedly placed the statistics for these 30 batches on the average and standard deviation chart shown in Figure 2 and attached this chart to the stack of pages.

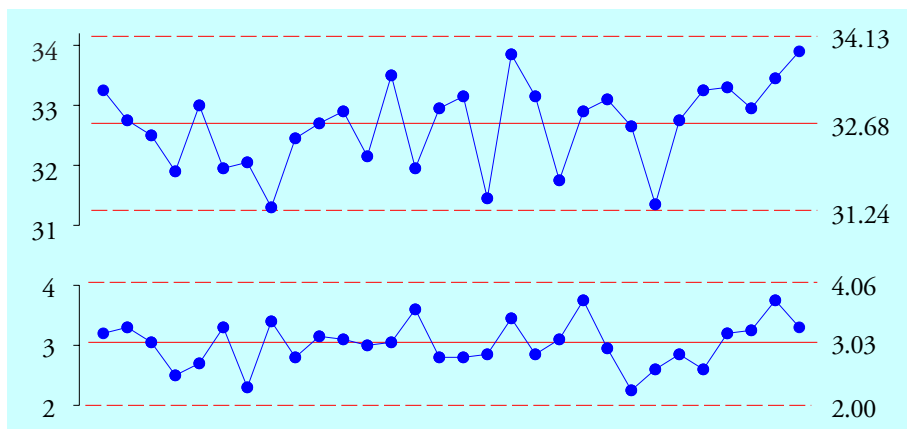


Figure 2: Average and Standard Deviation Chart for thirty Subgroups of Size 40

How do I know this chart was an afterthought? Because the order of the subgroups was the same as the order of the pages in the stack. Since this stack of 30 pages had been shuffled many times, there was no connection between the context for the data and the order of the subgroups. So, I started rearranging the subgroups according to both the levels of the factors studied in the experiment and the processing information recorded on each page. I did not find anything until I plotted the data in the order of production of each batch of compound, as shown in Figure 3.

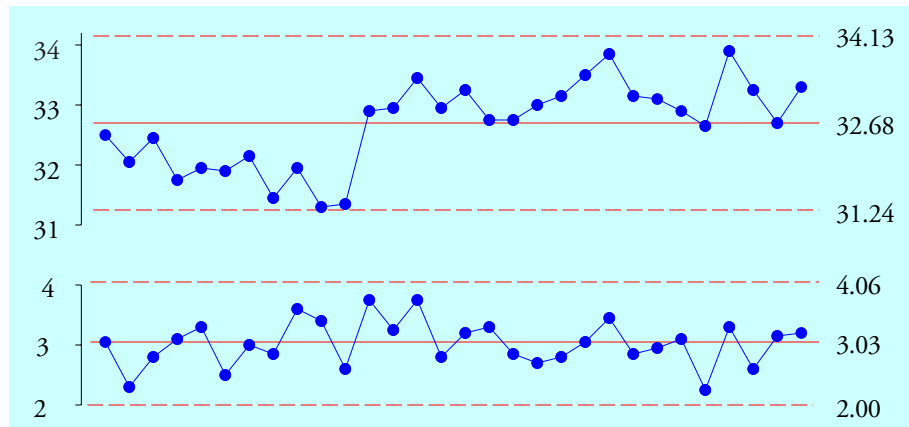


Figure 3: Average and Standard Deviation Chart for thirty Subgroups of Size 40 in Production Order

I flew up to Connecticut, showed them Figure 3, watched their mouths fall open, and flew home, where I sent them a bill for their failure to plot the data in time order sequence. By only looking at their data in terms of their experimental factors, and failing to look at the data in the broader context they had missed the story in their data. The predominant signal in these data is that the first eleven batches are detectably different from the batches produced later. Since this shift did not correspond to any of the experimental factors in the study, we had to conclude that it was due to some unknown assignable cause of exceptional variation in the production operations. So, while they had assumed that the batches were all equivalent, their experiment and their analysis had been totally corrupted by this lack of homogeneity. By analyzing the data only in terms of the experimental variables they missed the signal within their data. Good experimenters are always alert for factors outside the experiment as well as those in the study.

Data have no meaning apart from their context, and plotting the data in a way that respects this context should be an essential part of every analysis.

SUBGROUPS OF SIZE ONE

As we saw last month, the key to having an effective and sensitive chart is to have subgroups that are internally homogeneous. This will generally argue against using large subgroup sizes such as those used in the previous example. In writing about this problem Shewhart wrote on page 314 of *Economic Control of Quality of Manufactured Product*:

“Obviously, if the cause system is changing, the sample [subgroup] size should be as small as possible so that the averages of samples do not mask the changes. In fact single observations would be the most sensitive to such changes. Why then do we not use a sample size of unity? The answer is that if we do, we are faced with the difficulty of choosing the standard deviation to be used in the control charts. ... In fact, the sensitivity of the [process behavior chart] will increase, in general, with the decrease in subgroup size until the size of the subgroup is such that the data in any given subgroup come from a constant system of chance [i.e. Common] causes. In the absence of any *a priori* information making it possible to divide the data into rational subgroups, there would be some advantage therefore in reducing the subgroup size to unity.”

Thus, from the very beginning, the philosophy behind the process behavior chart has been pushing us in the direction of the Chart for Individual Values. Initially, the question of how to measure the within-subgroup variation prevented us from reaching this goal. While Shewhart found a way around this problem for certain types of count-based data (np -charts, p -charts, c -charts and u -charts are all Charts for Individual Values), his solution does not generalize to other types of data.

In 1942, W. J. Jennett provided an answer to the problem posed by Shewhart. In his work with the MO Valve Company Ltd., Jennett proposed using a technique that had been used in studying the ballistics of artillery shells in the 1890s. This technique was known as the method of successive differences. By using the differences between successive values as a measure of dispersion, Jennett effectively created the Chart for Individual Values and a Moving Range, or XmR Chart.

And, as Shewhart observed, using individual values makes the question of subgrouping moot. If you do not know how to organize your data into rational subgroups, then try using subgroups of size one. It is hard to mess up the subgrouping on an XmR chart. Nothing can hide. Every point has to sink or swim on its own. However, because we will still be using pseudo-subgroups of size two defined by the successive individual values, we are not totally free from the requirements of rational subgrouping. We will still need for successive values to be logically comparable.

METHANOL VALUES

The percentages of methanol in successive batches of alcohol are shown in Figure 4 along with their XmR chart. Here we get one number per batch, and we want to analyze and interpret each value as they become available. Waiting until we have four or five batches before we plot a point is completely unsatisfactory.

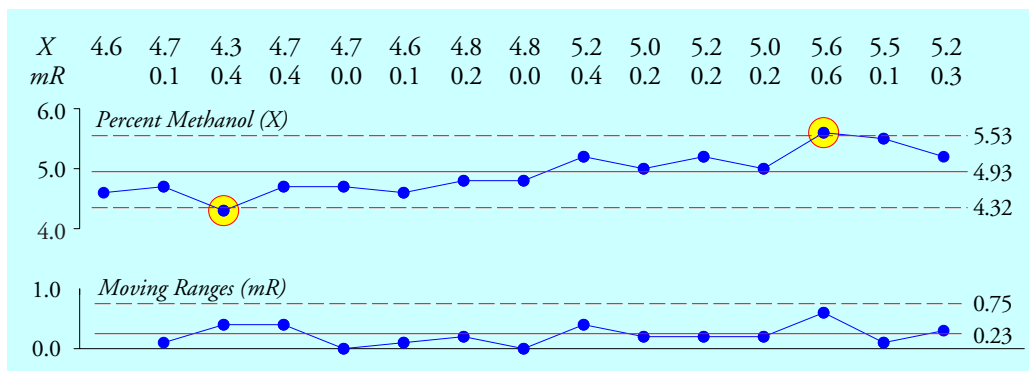


Figure 4: XmR Chart for Percentages of Methanol in Batches of Alcohol.

In Figure 4 the average is 4.93 while the average moving range is 0.23. The two points outside the Natural Process Limits confirm that the apparent upward trend in the percentages of methanol is indeed real. This production process was changing when it was thought to be operating in a steady state. The Assignable Cause of this trend will need to be identified and controlled before this process can operate up to its full potential.

So what would you predict for the next batch? Do you think it will have 4.9 percent

methanol? The fact that this process is changing over time makes all predictions little more than wishful thinking. Having characterized this process as unpredictable, we have made all questions of estimation moot.

There are two basic requirements to be met when creating an XmR Chart. The first requirement is that *successive values need to be logically comparable*. This means that you cannot use a single XmR chart to compare apples and oranges. Your data must consist of a stream of values that represent the same process or operation. In Figure 4 we have satisfied the first requirement by comparing the same measurement on successive batches coming from the same production process.

The second requirement is that *the moving ranges need to isolate and capture the local, short-term, routine variation that is inherent in the process generating the data*. In Figure 4 we have satisfied the second requirement by using the differences between successive batches to make allowance for the routine batch-to-batch variation. Since the observed long-term variation exceeds the limits based on this short-term variation we can be confident that the production process in Figure 4 is changing.

While the use of the time-order sequence of the data will usually be sufficient to satisfy these two requirements of the XmR Chart, there are times when a careful consideration of the structure in the data will require something other than a time-ordered sequence.

(For example, if the data of Figure 4 represented batches produced by two separate distillation units, say Unit X and Unit Y, and the values in time order formed the sequence X, Y, X, Y, X, Y, X, Y, X, Y, ... then the moving ranges would represent the difference between the two units rather than the routine variation of either unit. Here the data would need to be arranged in time order *by operating unit* before creating two separate charts for the two units.)

RATIONAL SAMPLING AND RATIONAL SUBGROUPING

Rational sampling simply refers to having an analysis that makes sense in terms of the data and the process represented by those data. Are the data collected in a manner that will allow the charts to detect process changes that are large enough to be of economic interest? Is the sample frequency high enough to detect changes in a timely manner? Is the sample frequency low enough to keep from getting overwhelmed by a flood of data in between the times when we might take action on the process? Does the data give us the appropriate information needed to take action on our process? Do we need to change our data collection or subgrouping when we change the way we operate our process?

Rational subgrouping is the art of organizing your data to answer the questions of interest. It can change as the questions of interest change. Of course, you cannot use rational subgrouping until you know the context for your data. As we learned in April, this means, among other things, that you **MUST** be able to express exactly what the measurements or counts represent in language that makes sense.

Process behavior charts are extremely versatile and flexible. Rational sampling and rational subgrouping are the key to using these charts to answer the questions about your process that are of interest to you and your organization. While it is important to use a within-subgroup measure of dispersion and to use three-sigma limits, these correct computations cannot do the job without rational sampling and rational subgrouping. The rationality of the collection and organization of the data is essential for the effective analysis of observational data.