A Tale of Two Comparisons

Are these two processes the same?

Donald J. Wheeler

Comparisons are often required as part of doing business. Are these two machines the same? is this process the same as that process? Are two operators performing in the same way? In this article we will look at two ways of making these comparisons.

The data we shall use are the maximum observed diameters for the three bearing surfaces of an automobile engine camshaft. The nominal dimension is supposed to be 1.3750 inches. The values recorded are the last two digits of 1.37xx, expressed in increments of a ten-thousandth of an inch. One camshaft is measured out of each tray of camshafts produced. The data for fifty consecutive trays are found in Figure 1. Each of the three bearings for a camshaft is produced on a different piece of equipment, so it is logical to want to compare these three operations.

Camshaft	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***
Bearing 1	49	51	51	52	56	52	51	52	50	49.5	51	52	53
Bearing 2	50	50	52	49	52	49	49	50	48	48	49	48	50
Bearing 3	50	46	52	51	42	50	50.5	44	48	49	43	49.5	50
Camshaft	***14***	***15***	***16***	***17***	***18***	***19***	***20***	***21***	***22***	***23***	***24***	***25***	***26***
Bearing 1	52	51	51	55	51	50.5	49	51	51	52.5	50	50	50
Bearing 2	49	49	51	51	52	50	50	48	49	50	50	48	50
Bearing 3	46	49	43	42	45	42	43	43	45	42	46	47	45
Camshaft	***27***	***28***	***29***	***30***	***31***	***32***	***33***	***34***	***35***	***36***	***37***	***38***	***39***
Bearing 1	53	52	50	53	52	52	51.5	51	49.5	52	51	51.5	51
Bearing 2	47	48	49	48	52	52	53	53	51	51	51.5	49	54.5
Bearing 3	42	49	49	52	46	50	51	50	51	50	52	52	54
Camshaft	***40***	***41***	***42***	***43***	***44***	***45***	***46***	***47***	***48***	***49***	***50***		
Bearing 1	51	50	50.5	51	51	51	56	50	50	52.5	57		
Bearing 2	50	48	50	47	49	49	48	50	52	48	48		
Bearing 3	52.5	54	51	51	51	49.5	52	49	49	50	50		

Figure 1: Camshaft Bearing Diameters

STUDENTS T-TEST

So how do you compare two processes or operations? A traditional approach is to use a two-sample t-test. Bearing One has an average of 51.46 and a standard deviation statistic of 1.68. Bearing Two has an average of 49.78 and a standard deviation statistic of 1.68. The pooled standard deviation statistic is 1.68 and it has 98 degrees of freedom. Therefore, our 95% interval estimate for the difference between these two average bearing diameters is:

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad t_{\alpha/2} \; s_p \sqrt{(1/n_1) + (1/n_2)}$$

which becomes:

$$[\ 51.46 - 49.78\ ] \pm 1.984\ (1.68)\ (0.2000) \quad = \quad 1.68 \pm 0.6666 \quad = \quad 1.01 \text{ to } 2.35$$

Since this interval does not contain zero we conclude that there is a detectable difference at the 5% level between the average diameters for bearings one and two. Bearing One averages 1.68 units larger than Bearing Two. You will need to adjust the aim point for Bearing One down about 1.5 units to get it to operate near the target value of 50. Bearing Two is already operating near the target value.

Now many students have been taught that you have to have similar standard deviations for the two groups before you can use the two-sample t-test. Some are even taught to test for equality of variance before using a two-sample t-test. Since tests on variances are much less robust than tests for location this is bad advice. To illustrate the effects of unequal variation upon the t-test we will compare Bearings Two and Three.

The average for Bearing Two is 49.78 units and the standard deviation statistic is 1.68 units. For Bearing Three the average is 48.2 units and the standard deviation is 3.54 units, which is over twice the size of the standard deviation statistic for Bearing Two. This results in a pooled standard deviation statistic of 2.770 units with 98 degrees of freedom. Our 95% interval estimate for the difference between these two bearings is:

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \quad &\pm \quad t_{\alpha/2}\ s_p \sqrt{(1/n_1) + (1/n_2)} \\ &= \quad 1.58 \pm 1.984\ (2.770)\ (0.2000) \\ &= \quad 1.58 \pm 1.099 \quad = \quad 0.48 \text{ to } 2.68 \end{aligned}$$

Since this interval does not contain zero we have once again found a detectable difference, and we did this in spite of the difference between the two standard deviation statistics. Clearly, Bearing Two averages 1.58 units larger than Bearing Three. You will need to adjust the aim point for Bearing Three up about 1.8 units to get it to operate near the target value of 50.

Nonhomogeneity of variation between the two groups may well make the t-test *less* sensitive, but when you detect a difference it is still likely to be a real difference, even if the two groups do not have equal variances.

It is important to be clear about what we have and have not discovered from these data. We now know that the three operations are not producing bearings with the same average diameters. We have not actually characterized any other aspects of these three operations. Just as the t-test does not require any distributional assumptions, it also does not check the data for other types of differences that may exist. It simply compares the average for one data set with the average for the other data set to see if they differ by more than can be reasonably attributed to chance.

PROCESS BEHAVIOR CHARTS

Can we detect the difference between the averages with the *XmR* Charts? Yes we can. The limits of a process behavior chart are intended for extrapolation. We can take the limits for one operation and use them with the data from another operation to ask the question "Are these two processes operating the same?" In Figure 2 the *X* chart limits for Bearing One are extended to the data for Bearing Two.

The thirteen points from Bearing Two that fall below the lower limit for Bearing One are clear evidence that the two operations are operating with different overall averages.

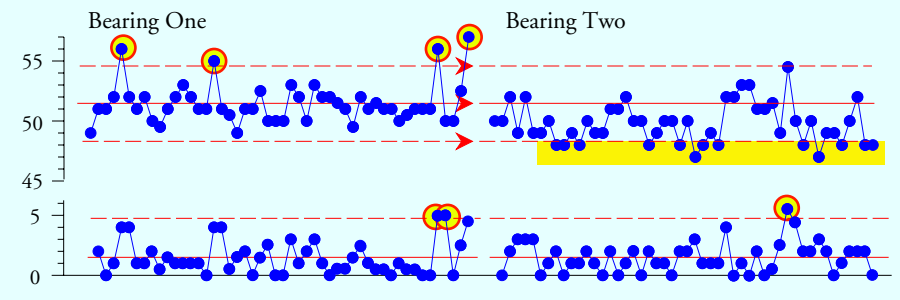

Figure 2: Comparing Bearing Two with Bearing One with an *XmR* Chart

We could have gone the other direction and use the limits for Bearing Two with the data for Bearing One. This is shown in Figure 3. Here the fact that all but four points for Bearing One fall above the central line for Bearing Two provides clear evidence that the two averages are different.

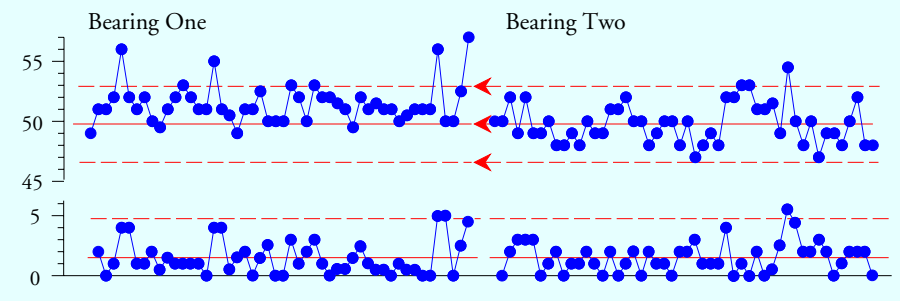

Figure 4: Comparing Bearing One with Bearing Two

However, the two *XmR* charts in Figure 4 also show that these two operations are subject to occasional upsets. These upsets undermine the interpretation of the averages as being representative of some underlying fundamental constant for each of these operations. So while you might want to lower the aim point for Bearing One, Figure 4 tells you that if you do not figure out what is causing the process upsets, no amount of adjusting the process aim is going to guarantee that your product is on target.

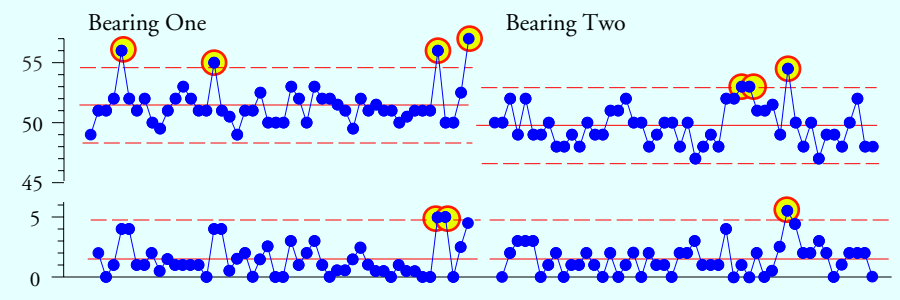

Figure 4: *XmR* Charts for Bearing One and Bearing Two

When we place the *XmR* chart for Bearing Two next to that for Bearing Three we get Figure 5. Bearing Two has 3 out of 50 points outside the *X* chart limits. Bearing Three has no points outside

the *X* chart limits? So which process is doing better?

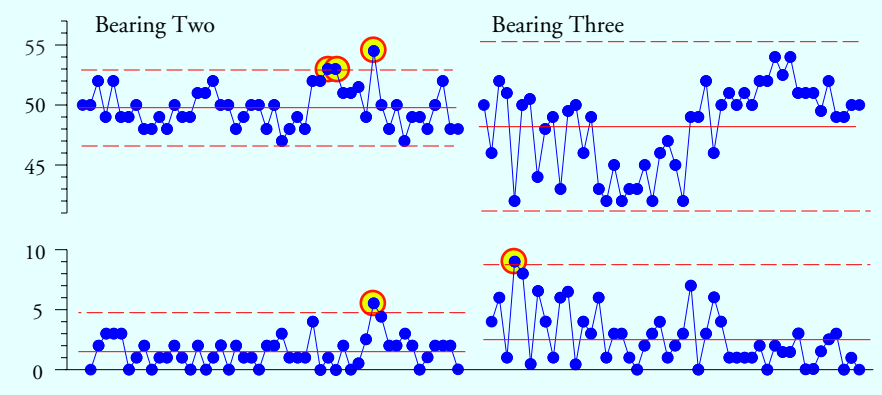

Figure 5: *XmR* Charts for Comparing Bearings Two and Three

The difference here is not so much a quantitative difference as it is a qualitative one. You cannot capture this difference with any computation.

Just what does the average of 48.2 for Bearing Three represent?

Are the first 15 camshafts consistently near the average of 48.2? No, five are substantially less and nine are closer to the target value of 50.

Are camshafts 16 to 27 anywhere near the average of 48.2? No they are all substantially smaller than 48.2.

Are camshafts 41 to 50 anywhere near the average of 48.2? No, they are all substantially larger than 48.2.

So, I repeat, what does the average of 48.2 for Bearing Three represent in terms of the underlying process? Absolutely nothing.

Likewise with measures of dispersion. For Bearing Three the variation of the first 15 camshafts is not the same as the variation in the middle, and this is not the same as the variation for the last 10 camshafts.

No computation can ever take the place of a good graph of your data. Some of the most important information may be tied up in the time-order sequence of your data, and *every one of your numerical summaries ignores this time-order information*.

Until you discover the assignable causes that are upsetting Bearings One and Two, and the assignable causes that are taking Bearing Three on walkabout, all of your t-tests, all of your statistics, and all of your analyses of these data will merely be a triumph of computation over common sense. Your computations may well summarize your data, but when the underlying process is demonstrably changing over time the data become a meaningless collection of dissimilar values and the computations have no contact with reality.

The primary question of data analysis has always been the question of whether or not the data are homogeneous. And the primary tool for examining the data for homogeneity is the process behavior chart. Start every analysis with a process behavior chart, or suffer the consequences of obtaining correct answers to the wrong questions.