2005

Integrating Several Variance Estimators

Donald Holmes
Stochos, Inc.

A. Erhan Mergen
Rochester Institute of Technology

INTEGRATING SEVERAL VARIANCE ESTIMATORS

Donald S. Holmes, Stochos Inc. 14 N. College Street, Schenectady, N.Y. 12305.
(518) 372-5426, dsholmes@stochos.com
A. Erhan Mergen, Rochester Institute of Technology, College of Business
Decision Sciences, 107 Lomb Memorial Drive, Rochester, N.Y. 14623-5608.
(585) 475-6143, emergen@cob.rit.edu

ABSTRACT

It is crucial to understand the proper variance estimator in statistical process control (SPC) when you are trying to answer the questions on the process behavior both in the short term and the long term. In this paper we will discuss various variance estimators including their calculation using range and/or range squares, their potential use in SPC, their similarities and differences.

Keywords: Statistical process control, mean square successive difference, range, capability variance, performance variance.

DISCUSSION

Range is a measure of process width which is fairly simple to calculate: the process width is the distance between the largest and the smallest values. Let's expand the width concept by looking at the distances between every possible pair of values, i.e., X's (e.g., $X_4 - X_1$). You can see, of course, that the absolute value of every possible pair is just every possible range that can be formed for the given data set. Averaging every possible width should provide better insight into the process width than just using one range (after all, we don't use a single X to estimate the process center!).

Suppose we take the following easy set of X's to demonstrate this approach:
1, 2, 3, 4, 5, i.e., n=5.

			Process
Average ($\overline{X}$)	=	3.0	center
Range (R)	=	4.0	width
Std. dev. (s)	=	1.58	

6*s = width = 9.48 which is different than 4.0.
Variance (s^2) = 2.5

Note that the width measure that uses only two X's is 4.0 whereas the one that uses all X's is 9.48. That is a major disparity which may, as you say, be helped some by converting R to an estimate of standard deviation by dividing a correction factor d_2 (see any book on SPC, for example, one by Montgomery [5, pp.210-211]) and then multiplying by 6 to get the estimated width. Estimated process width would then be:

$$= \left(\frac{R}{d_2}\right)6 = \left(\frac{4}{1.128}\right)6 = 21.3 \qquad (1)$$

This is considerably different than the value of 9.48.

Next let's take a look at the ranges of every possible pair of all X's rather than just one.

Table 1. Ranges

X	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

The zeroes on the main diagonal furnish no information and can be ignored – we know that the range of a variable with itself is zero. Thus there are (5x5) – 5 (in general n^2 - n or n(n-1)) many ranges to consider. The average range is then the sum of all off diagonal elements divided by n(n-1), which in this case 40/20 = 2. Now this value adjusted to estimate width is $\left(\frac{2}{1.128}\right)6 = 10.6$ - which is much closer to the 6s width of 9.48. That is to say, using all of the ranges (not just one) we can get an estimate of process width without calculating the average of the X's at all.

We can actually estimate the variance (s^2) without using the average value. This can be done by taking the ranges of possible pairs of data in the sample, squaring the ranges, averaging the squared ranges and then dividing the average squared range by 2 to get the variance. Let's use the above data set to demonstrate this.

As you see, this estimate of variance is identical to the one that we obtained using the average. It is relatively easy to prove that this is generally true – that is, not just peculiar to this data set.

Table 2. Squared Ranges

X	1	2	3	4	5
1	0	1	4	9	16
2	1	0	1	4	9
3	4	1	0	1	4
4	9	4	1	0	1
5	16	9	4	1	0

$$\Sigma R^2 = 100 \qquad (2)$$

$$\overline{R}^2 = \frac{\Sigma R^2}{n(n-1)} = \frac{100}{20} = 5 \tag{3}$$

$$s^2 = \frac{\overline{R^2}}{2} = \frac{5}{2} = 2.5 \tag{4}$$

Eliminating the zero ranges on the main diagonal was based on the fact that these are inescapable facts which provide no information about the average width. Thus one is averaging (n)(n)-n values (or n(n-1)). The information argument, extended, would be that the matrix of R^2 above is symmetrical and the top half provides no more information than the bottom half. So, the average of the R^2 values obtained by dividing by n(n-1) is to be multiplied by ½ (i.e., divided by 2) to get the variance estimator.

Another way to show that this "divide by 2" is a valid approach is to calculate R^2 for any pair and show that it is simply the R^2 divided by two. Let's use first and the last observation from the above set, i.e., X_1 and X_5 and n=2.

Let $R_{15} = |(X_1 - X_5)|$ and $R_{51} = |(X_5 - X_1)|$. The sample variance then would be

$$= \frac{\left(\left[X_1 - \frac{X_1 + X_5}{2}\right]^2 + \left[X_5 - \frac{X_1 + X_5}{2}\right]^2\right)}{n-1} \tag{5}$$

$$= \frac{\frac{(X_1 - X_5)^2}{4} + \frac{(X_5 - X_1)^2}{4}}{1} \tag{6}$$

$$= \frac{R_{15}^2}{4} + \frac{R_{51}^2}{4} \text{ or } = \frac{R_{15}^2}{2} \tag{7}$$

So each of the sample size two R^2s are divided by two to get a variance and then they are averaged. What has been shown is that the variance measure of process width is just the average of a bunch of squared ranges (R^2) divided by two. One may ask the importance of this. Answering that question brings us to the next issue: Dynamic (order dependent) vs. static (order independent) variances (and thus standard deviations).

The usual definition for variance (i.e., average of squared distances between each X and the average of the X's) is time independent. That is to say, the variance of the data set 1, 2, 3, 4, 5 is the same for the data set 1, 4, 3, 5, 2 which is the same for the data set 3, 5, 2, 4, 1, etc. from the average. Referring to the range squared matrix above, it is calculated using all of the entries other than those on the main diagonal. The value of this variance is 2.5.

The next definition of variance is one which utilizes the first diagonal below the main diagonal. It is called the Mean Square Successive Difference (MSSD) estimator for the variance (see, for example, Neumann, et al. [6], Hald [1, pp.357-360], Holmes and Mergen [3]. Since it is the result of just one specific diagonal, rather than all the values, the result depends upon the order of the data.

Examples

Remember that the regular variance for the data set above is 2.5 (as shown above).
X's: 1, 2, 3, 4, 5

Case 1:
X's in order of occurrence: 1, 2, 3, 4, 5
Character of "time series": Pure linear trend
Squared ranges: 1, 1, 1, 1

Variance estimate (MSSD): $$\frac{\frac{\Sigma R^2}{n-1}}{2} = \frac{\frac{4}{4}}{2} = 0.5 \quad (8)$$

Note that this is the first diagonal below the main diagonal in the Range squared matrix above.

Case 2:
X's in order of occurrence: 1, 3, 4, 2, 5
Character of "time series": Nearly random
Squared ranges: 4, 1, 4, 9

Variance estimate (MSSD): $$\frac{\frac{\Sigma R^2}{n-1}}{2} = \frac{\frac{18}{4}}{2} = 2.25 \quad (9)$$

Case 3:
X's in order of occurrence: 1, 5, 4, 2, 3
Character of "time series": Nearly random
Squared ranges: 16, 1, 4, 1

Variance estimate (MSSD): $$\frac{\frac{\Sigma R^2}{n-1}}{2} = \frac{\frac{22}{4}}{2} = 2.75 \quad (10)$$

Case 4:
X's in order of occurrence: 1, 5, 2, 4, 3
Character of "time series": Highly cyclical
Squared ranges: 16, 9, 4, 1

Variance estimate (MSSD): $\frac{\frac{\sum R^2}{n-1}}{2} = \frac{\frac{30}{4}}{2} = 3.75$ (11)

Use of These Variances

As you can see from these examples, if the data series is nearly random, the MSSD variance estimate is approximately the same as the regular variance. In the SPC sense, if the process is "in control" (i.e., random), the two estimates are similar. On the other hand, if there is a smooth trend in the data, the MSSD estimator is smaller than the regular estimator. That is to say, if the process is subject to unexpected trends (not "in control"), the MSSD variance provides an estimate of the potential reduction in process variance which can be achieved if the process were brought "into control". In a generic sense, this variance is an estimate of the "capability" variance of the process. In contrast, the regular variance provides an estimate of the "performance" variance of a process (see, for example, Holmes and Mergen [3]), i.e., current total variance in the process. Should the MSSD variance indicate that there are cycles in the data; the data must be segregated prior to looking for the capability measure of variance.

The MSSD variance estimate can be used, for example, to determine the potential capability of a process which is not in statistical control (i.e., not stable) should the non-random causes be removed (Holmes and Mergen [4]). This potential capability estimate of the process can then be compared to the current capability estimate (i.e., current performance of the process), which uses the regular variance estimate. The proper actions can then be taken to improve the process capability.

Another use of MSSD variance estimate would be in testing the rationality of the subgroups formed in control charts. A comparison of the MSSD variance estimator with the regular variance estimator in each subgroup can be tested for significant difference (Holmes and Mergen [2]).

The standard deviation estimator, such as $\frac{\bar{R}}{d_2}$, is sort of an unconventional estimator in the sense it does not take the squared differences; that's why the average R value should be divided by a bias reduction factor, d_2. Another estimate which is similar to this is $\frac{\bar{s}}{c_4}$. This standard deviation estimate sums the sample standard deviations, averages it and then divides the average by a bias reduction factor, c_4, to estimate the process standard deviation. This has to be done because this estimate is not based on the sum of the sample variances, but rather on the sum of the sample standard deviations

CONCLUSION

In this paper we introduced several variance estimators that can be used in SPC. It is important that the proper variance estimator is chosen to deal with the question at hand. Failure to do this may lead to an erroneous conclusion about the process variability.

REFERENCES

[1] Hald, A. *Statistical theory with engineering applications*. New York, NY: Wiley, 1952.

[2] Holmes, D.S. and Mergen, A.E., "Testing control chart subgroups for rationality." *Quality and Reliability Engineering, International*, 1989, 5(2), 143-147.

[3] Holmes, D.S. and Mergen, A.E., "An alternative method to test for randomness of a process." *Quality and Reliability Engineering International*, 1995, 11(3), 171-174.

[4] Holmes, D.S. and Mergen, A.E., "Estimating potential capability of an unstable process." *Proceedings of the 2004 North East Decision Sciences Meeting*, March 2004, 290-292.

[5] Montgomery, D.C. *Introduction to statistical quality control*, New York, NY: Wiley, 2001.

[6] Neumann, J.V., Kent, R.H., Bellinson, H.R. and Hart, B.I., "The Mean square successive difference." *Annals of Mathematical Statistics*, 1941, 12(2), 153-162.