Separating the Signals from the Noise

Donald J. Wheeler

The second principle for understanding data is that while some data contain signals, all data contain noise, therefore, before you can detect the signals you will have to filter out the noise. This act of filtration is the essence of all data analysis techniques. It is the foundation for our use of data and all the predictions we make based on those data. In this column we will look at the mechanism used by all modern data analysis techniques to filter out the noise.

Given a collection of data it is common to begin with the computation of some summary statistics for location and dispersion. Averages and medians are used to characterize location, while either the range statistic or the standard deviation statistic is used to characterize dispersion. This much is taught in every introductory class. However, what is usually not taught is that the structures within our data will often create alternate ways of computing these measures of dispersion. Understanding the rolls of these different *methods* of computation is essential for anyone who wishes to analyze data.

Perhaps the most common type of structure for a data set is to have k subgroups of size n where the n values within each subgroup were collected under the same set of conditions. This structure is found in virtually all types of experimental data, and in most types of data coming from a production process. To illustrate the alternate ways of computing measures of dispersion we shall use a simple data set consisting of $k = 3$ subgroups of size $n = 8$ as shown in Figure 1.

Subgroup One	*4*	*5*	*5*	*4*	*8*	*4*	*3*	*7*
Subgroup Two	*2*	*4*	*3*	*7*	*5*	*4*	*2*	*5*
Subgroup Three	*3*	*6*	*6*	*4*	*5*	*4*	*6*	*6*

Figure 1: Data Set One

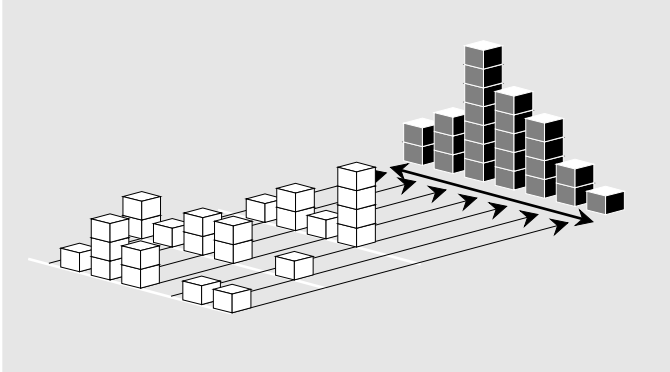


Figure 2: Method One for Estimating Dispersion

DATA SET ONE WITH METHOD ONE

The first method of computing a measure of dispersion is the method taught in introductory classes in statistics. All of the data from the k subgroups of size n are collected into *one* large group of size nk and a single dispersion statistic is found using all nk values. This dispersion statistic is then used to estimate a dispersion parameter such as the standard deviation for the distribution of X, $SD(X)$.

As shown in Figure 3 the range of all 24 values is 6. The bias correction factor for ranges of 24 values is 3.895. Dividing 6 by 3.895 yields an unbiased estimate of the standard deviation of the distribution of X of 1.540.

The global standard deviation statistic is 1.551. The bias correction factor for this statistic when it is based on 24 values is 0.9892. Dividing 1.551 by 0.9892 yields an unbiased estimate of the standard deviation of the distribution of X of 1.568.

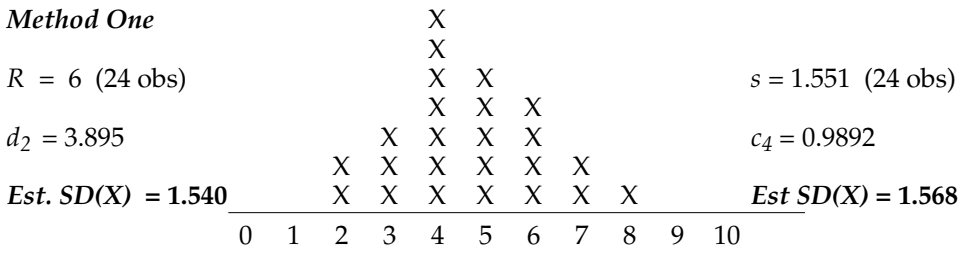


Figure 3: Method One with Data Set One

Since the original data are given to the nearest whole number, there is no practical difference between the two estimates of $SD(X)$ shown in Figure 3. Whether we use the range or the standard deviation statistic will not substantially affect our analysis.

DATA SET ONE WITH METHOD TWO

While Method One ignores the subgroups, Method Two respects the subgroup structure within the data. Here we calculate a dispersion statistic for each subgroup. These separate dispersion statistics are then averaged, and the average dispersion statistic is used to form an unbiased estimate for the standard deviation parameter of the distribution of X.

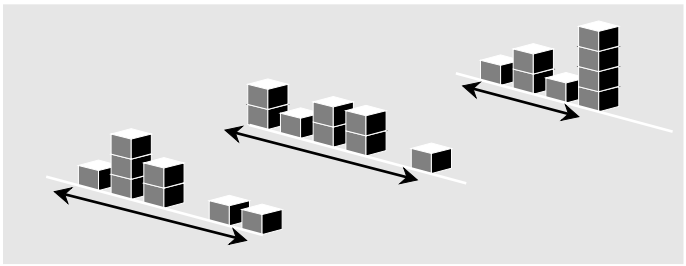


Figure 4: Method Two for Estimating Dispersion

Using Data Set One, we compute a dispersion statistic for each of the three subgroups. Since the subgroups are all the same size we can average the statistics prior to dividing by the common bias correction factor.

As shown in Figure 5, the subgroup ranges are respectively 5, 5, and 3. The average range is

4.333 and the bias correction factor for ranges of eight data is 2.847. Dividing 4.333 by 2.847 we estimate the standard deviation for the distribution of X to be 1.522.

The subgroup standard deviation statistics are respectively 1.690, 1.690, and 1.195. The average standard deviation statistic is 1.525 and the bias correction factor is 0.9650. Dividing 1.525 by 0.9650 we estimate the standard deviation for the distribution of X to be 1.580.

Method Two

Subgroup One:				X						
$R_1 = 5$				X	X					$s_1 = 1.690$
$d_2 = 2.847$			X	X	X		X	X		$c_4 = 0.9650$
	1	2	3	4	5	6	7	8	9	
Subgroup Two:										
$R_2 = 5$		X		X	X					$s_2 = 1.690$
$d_2 = 2.847$		X	X	X	X		X			$c_4 = 0.9650$
	1	2	3	4	5	6	7	8	9	
Subgroup Three:						X				
						X				
$R_3 = 3$				X		X				$s_3 = 1.195$
$d_2 = 2.847$			X	X	X	X				$c_4 = 0.9650$
	1	2	3	4	5	6	7	8	9	

Average Range = 4.333 *Average Std. Dev. = 1.525*
Est. SD(X) = 1.522 ***Est. SD(X) = 1.580***

Figure 5: Method Two with Data Set One

As before, there is no practical difference between the two estimates shown in Figure 5. Neither is the any practical difference between the estimates in Figure 3 and those in Figure 5. The four estimates obtained using the two different measures of dispersion and the two different methods are all very similar.

DATA SET ONE WITH METHOD THREE

The third method will probably seem rather strange. It is certainly indirect. Instead of working with the individual values as the first two methods do, the third method works with the subgroup averages. These subgroup averages are used to obtain a dispersion statistic, and this dispersion statistic is then used to estimate the standard deviation parameter of the distribution of X.

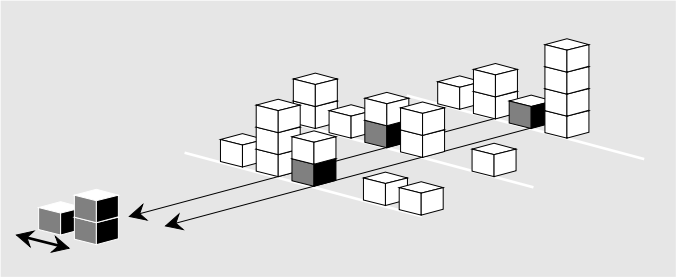


Figure 6: Method Three for Estimating Dispersion

For Data Set One the subgroup averages are respectively 5.0, 4.0, and 5.0. The range of these

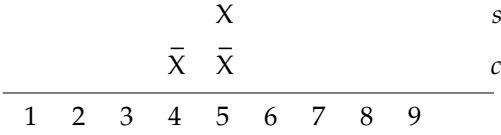
three averages is 1.00. The bias correction factor for the range of three values is 1.693. Since each of these averages represents eight original data, we will have to multiply by the square root of 8 and divide by the bias correction factor to estimate the standard deviation parameter for the distribution of X. When we do this with the values above we obtain an estimate of $SD(X)$ of 1.671.

The standard deviation statistic for the three subgroup averages is 0.5774. Dividing by the bias correction factor of 0.8862 and multiplying by the square root of 8 we obtain an unbiased estimate of the standard deviation of the distribution of X of 1.843.

Method Three

$R_{\bar{X}} = 1.000$ $\qquad$ $s_{\bar{X}} = 0.5774$

$d_2 = 1.693$ $\qquad$ $c_4 = 0.8862$



Est SD(X) = 1.671 $\qquad$ ***Est. SD(X) = 1.843***

Figure 7: Method Three with Data Set One

Once again, there is no practical difference between using the range and using the standard deviation statistic. Here the two estimates are slightly larger than before, but not by any appreciable amount.

	Range Based		Std. Dev. Based	
	Est SD(X)	c.v.	Est SD(X)	c.v.
Method One	1.540	18.1%	1.563	14.7%
Method Two	1.522	16.5%	1.580	15.6%
Method Three	1.671	50.6%	1.845	50.0%

Figure 8: Summary of Three Methods for Data Set One

As summarized in Figure 8, we have just obtained six unbiased estimates for the standard deviation parameter for the distribution of X using three different methods and two different statistics. These six values are listed along with their coefficients of variation (c.v.). The first four unbiased estimates are all quite similar because they all have similar coefficients of variation. The last two unbiased estimates are not as cozy as the first four because they have much larger coefficients of variation and therefore have more uncertainty attached.

Before we attempt to draw any lesson from this example we need to know that Data Set One has a very special property. When we place Data Set One on an average and range chart we end up with Figure 9. There we see no evidence of any differences between the three subgroups. Data Set One contains no signals. It is pure noise.

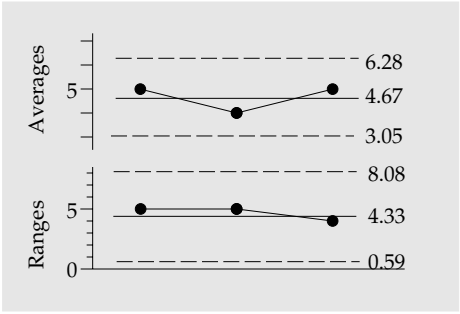


Figure 9: Average and Range Chart for Data Set One

Therefore, at this point we can reasonably conclude that when the data are homogeneous and contain no signals the three methods will yield similar values for unbiased estimates of *SD(X)* regardless of whether we use the range or the standard deviation statistic.

DATA SET TWO

But what happens in the presence of signals? After all, the objective is to filter out the noise so we can detect any signals that may be present. To see how signals affect our estimates of *SD(X)* we shall modify Data Set One by inserting two signals. Specifically we shall shift subgroup two down by two units while we shift subgroup three up by four units. This will result in Data Set Two which is shown in Figure 10. As may be seen in the average and range chart in Figure 11, these changes have introduced two distinct signals.

Subgroup One	*4*	*5*	*5*	*4*	*8*	*4*	*3*	*7*
Subgroup Two	*0*	*2*	*1*	*5*	*3*	*2*	*0*	*3*
Subgroup Three	*7*	*10*	*10*	*8*	*9*	*8*	*10*	*10*

Figure 10: Data Set Two

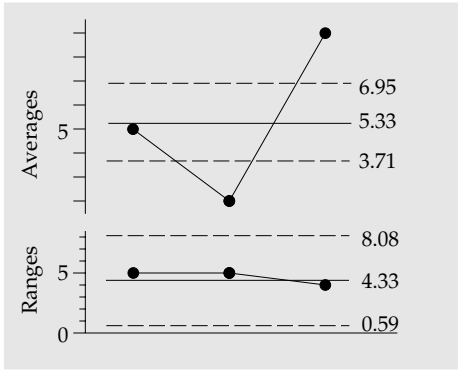


Figure 11: Average and Range Chart for Data Set Two

METHOD ONE WITH DATA SET TWO

Method One uses all 24 values in Data Set Two to compute global measures of dispersion. As shown in Figure 12, the global range is 10.0 which results in an unbiased estimate of the standard deviation parameter of 2.567. The global standard deviation statistic is 3.279 which gives an

unbiased estimate of the standard deviation parameter of 3.315.

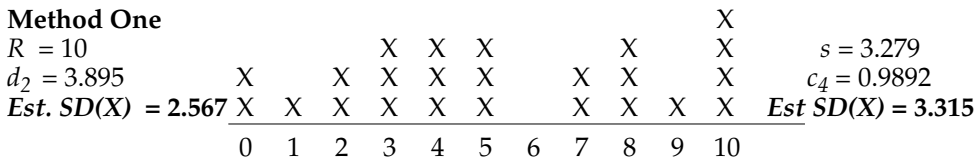


Figure 12: Method One with Data Set Two

These estimates of *SD(X)* are roughly twice the size of those found in Figure 3. Thus, the signals introduced by shifting the subgroup averages have inflated both of the Method One estimates by an appreciable amount.

METHOD TWO WITH DATA SET TWO

Using Method Two, we compute a dispersion statistic for each of the three subgroups. Since the subgroups are all the same size we can average the statistics prior to dividing by the common bias correction factor. As shown in Figure 13, the average range is 4.333 and the bias correction factor for ranges of eight data is 2.847. Dividing 4.333 by 2.847 we estimate the standard deviation for the distribution of *X* to be 1.522.

The average standard deviation statistic is 1.525 and the bias correction factor is 0.9650. Dividing 1.525 by 0.9650 we estimate the standard deviation for the distribution of *X* to be 1.580.

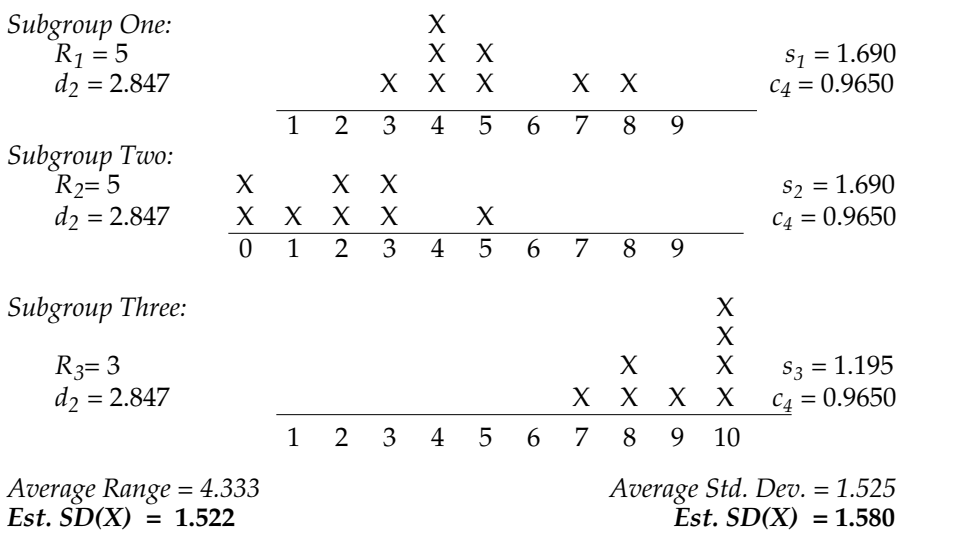


Figure 13: Method Two with Data Set Two

The Method Two estimates of *SD(X)* for Data Set Two are exactly the same as those obtained for Data Set One in Figure 5. Thus, the Method Two estimates are not afffected by the signals introduced by shifting the subgroup averages.

METHOD THREE WITH DATA SET TWO

For Data Set Two the subgroup averages are respectively 5.0, 2.0, and 9.0. The range of these three averages is 7.00. The bias correction factor for the range of three values is 1.693. Since each

of these averages represents eight original data, we will have to multiply by the square root of 8 and divide by the bias correction factor to estimate the standard deviation parameter for the distribution of X. When we do this with the values above we obtain an estimate of *SD(X)* of 11.693.

The standard deviation statistic for the three subgroup averages is 3.512. Dividing by the bias correction factor of 0.8862 and multiplying by the square root of 8 we obtain an unbiased estimate of the standard deviation of the distribution of X of 11.209.

Method Three

$R_{\bar{X}} = 7.000$ $s_{\bar{X}} = 3.512$

$d_2 = 1.693$ $\bar{X}$ $\bar{X}$ $\bar{X}$ $c_4 = 0.8862$

1 2 3 4 5 6 7 8 9

Est SD(X) = 11.693 ***Est. SD(X) = 11.209***

Figure 14: Method Three with Data Set Two

These Method Three estimates of *SD(X)* are seven times larger than values found in Figure 7. Thus, the signals introduced by shifting the subgroup averages have severely inflated both of the Method Three estimates.

When we summarize the results of the three methods with Data Set Two we get the table in Figure 15. We have obtained six unbiased estimates of *SD(X)* using three different methods and two different statistics, yet these six values differ by almost an order of magnitude!

	Range Based		Std. Dev. Based	
	Est SD(X)	c.v.	Est SD(X)	c.v.
Method One	2.567	18.1%	3.315	14.7%
Method Two	1.522	16.5%	1.580	15.6%
Method Three	11.693	50.6%	11.209	50.0%

Figure 15: Summary of Three Methods for Data Set Two

The differences left to right in Figure 15 show the effects of using the different dispersion statistics. The differences top to bottom reveal the differences due to using the different methods. Clearly, the differences left to right pale in comparison with those top to bottom. The key to filtering out the noise so we can detect the signals does not depend upon whether we use the standard deviation statistic or the range, but rather upon which *method* we employ to compute that dispersion statistic.

Method One estimates of dispersion are commonly known as the Total Variation or the Overall Variation. Method One is used for description. It implicitly assumes that the data are globally homogeneous. When the data are not globally homogeneous this method will be inflated by the signals contained within the data and the value obtained will no longer estimate *SD(X).*

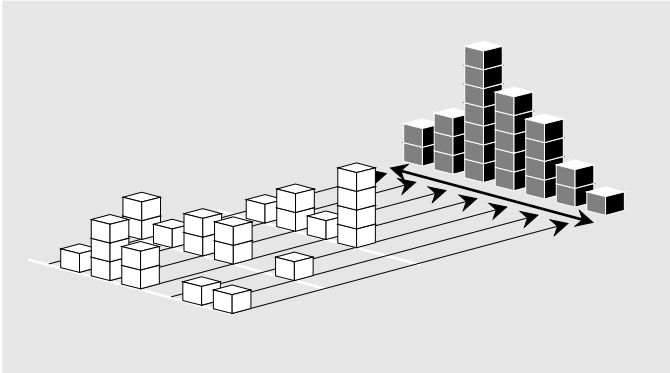


Figure 16: Total or Overall Variation

Method Two estimates of dispersion are commonly known as the Within-Subgroup Variation. Method Two is used for analysis. Whenever we seek to filter out the noise in order to detect signals we use Method Two to establish the filter. Method Two implicitly assumes that the data are homogeneous within the subgroups, but it places no requirement of homogeneity upon the different subgroups. Thus, even when the subgroups differ, Method Two will provide a useful estimate of *SD(X).*

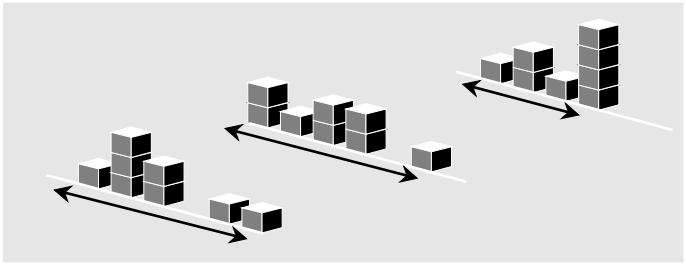


Figure 17: Within-Subgroup Variation

Method Three estimates of dispersion are commonly known as the Between-Subgroup Variation. Method Three is used for comparison purposes. It assumes that the subgroup averages are globally homogeneous. When Method Three is computed it is generally compared with Method Two; the idea being that any signals present in the data will affect Method Three more than they affect Method Two. When the subgroups differ, Method Three will not provide an estimate of *SD(X).*

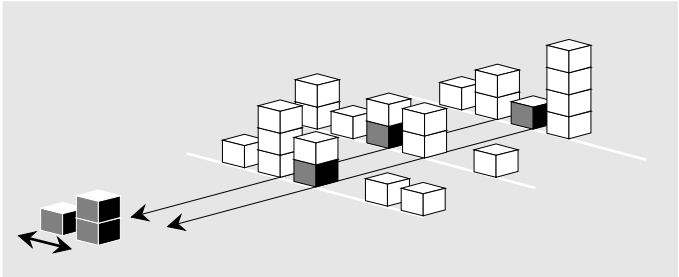


Figure 18: Between-Subgroup Variation

SEPARATING THE SIGNALS FROM THE NOISE

The essence of every statistical analysis is the separation of the signals from the noise. We want to find the signals so that we can use this knowledge constructively. We want to ignore the noise where there is nothing to be learned. To this end we begin by filtering out the noise. And for the past 100 years the standard technique for filtering out the noise has been Method Two! To illustrate this point Figure 19 shows the average chart for Data Set Two with limits computed using each of the three methods. Only Method Two correctly identifies the two signals we deliberately buried in Data Set Two.

So when it comes to filtering out the noise you have a choice between Method Two, Method Two, or Method Two. Any method is right as long as it is Method Two!

Method One is inappropriate for filtering out the noise because it gets inflated by the signals. Method One has always been wrong for analysis, and it will always be wrong. Trying to use Method One for analysis is so wrong that it has a name. It is known as Quetelet's Fallacy and it is the reason there was so little progress in statistical analysis in the Nineteenth Centrury.

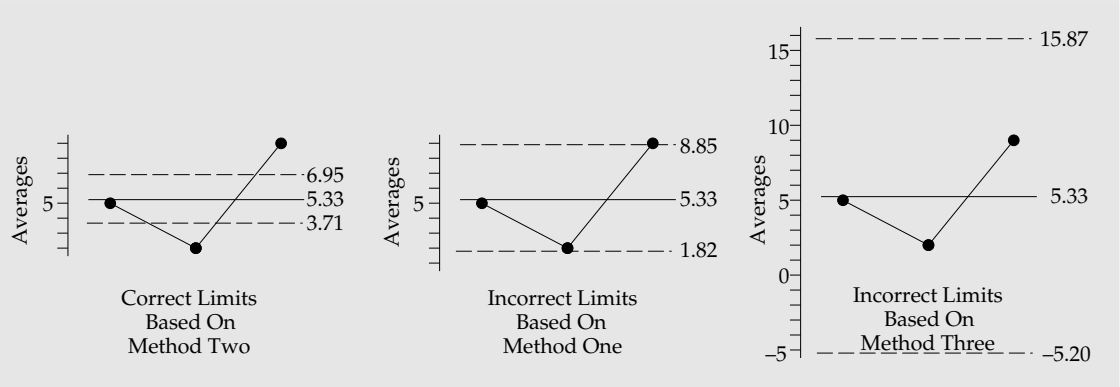


Figure 19: Average Charts for Data Set Two

Method Three is completely inappropriate for filterring out the noise because it will be severely inflated in the presence of signals. If you use Method Three to filter out the noise you will have to wait a very long time before you detect a signal. So while there are analysis techniques that make use of the Method Three (Between Subgroup) estimate of dispersion, they do so only in order to compare it with a Method Two (Within Subgroup) estimate of dispersion.

Thus, the foundation of all modern data analysis techniques is the use of Method Two to filter out the noise. This is the foundation for the Analysis of Variance. This is the foundation for the Analysis of Means. And this is the foundation for Shewhart's process behavior charts. Ignore this foundation and you will undermine your whole analysis.

Many analysis techniques from the Nineteenth Century, such a Franklin Pierce's test for outliers, are built on the use of Method One to filter out the noise. As may be seen in Figure 19, this approach will let you occasionally detect a signal, but it will cause you to miss other signals.

In fact, many techniques developed in the Twentieth Centruy also suffer from Quetelet's Fallacy. Among these are Grubb's test for outliers, the Levey-Jennings control chart, and the Tukey control chart. Moreover, virtually every piece of statistical software available today allows the user to choose Method One for creating control charts and performing various other statistical tests. Nevertheless, this error on the part of naive programmers does not make it right or even acceptable to use Method One for *analysis*.

So while there are proper uses of Method One and Method Three, they are never appropriate for filtering out the noise. The only correct method for filtering out the noise is Method Two. Understanding this point is the beginning of competence for every data analyst.

You now know the difference between modern data analysis techniques and naive analysis techniques. Naive techniques use Method One or Method Three to filter out the noise. Today all sorts of new naive techniques are being created by those who know no better. Let the user beware.

To help with this problem of identifying naive techniques Figure 20 contains a listing of 27 of the more commonly encountered within-subgroup estimators of both the standard deviation parameter and the variance parameter. There we see the hallmark of the within-subgroup approach: Each estimator is based on either the average or the median of a collection of k within-subgroup measures of dispersion. Method One and Method Three each use a single measure of disperison. Now you know the importance of using the right method, and you know what the right method will look like in practice. While this may be more than you ever wanted to know about statistics, it is essential knowledge for all seek to understand their data.

	Estimators for *SD(X)*		**Estimators for *V(X)***	
Name of Estimator	**Biased**	**Unbiased**	**Biased**	**Unbiased**
Average Range	$\frac{\overline{R}}{d_2^*}$	$\frac{\overline{R}}{d_2}$	$\left(\frac{\overline{R}}{d_2}\right)^2$	$\left(\frac{\overline{R}}{d_2^*}\right)^2$
Median Range	----	$\frac{\widetilde{R}}{d_4}$	$\left(\frac{\widetilde{R}}{d_4}\right)^2$	----
Average Moving Range	$\frac{\overline{R}}{1.414}$	$\frac{\overline{R}}{1.128}$	$\left(\frac{\overline{R}}{1.128}\right)^2$	$\left(\frac{\overline{R}}{1.414}\right)^2$
Median Moving Range	----	$\frac{\widetilde{R}}{0.954}$	$\left(\frac{\widetilde{R}}{0.954}\right)^2$	----
Average Root Mean Square Dev.	$\overline{s}_n$	$\frac{\overline{s}_n}{c_2}$	$(\overline{s}_n)^2$	----
Median Root Mean Square Dev.	$\widetilde{s}_n$	$\frac{\widetilde{s}_n}{c_1}$	$(\widetilde{s}_n)^2$	----
Average Standard Deviation	$\overline{s}$	$\frac{\overline{s}}{c_4}$	$(\overline{s})^2$	----
Median Standard Deviation	$\widetilde{s}$	$\frac{\widetilde{s}}{c_6}$	$(\widetilde{s})^2$	----
Pooled Variance	$\sqrt{\overline{s^2}}$	$\frac{\sqrt{\overline{s^2}}}{c_4'}$	----	$\overline{s^2}$

Figure 20: Some Within-Subgroup Estimators

This article is based on material found in *Advanced Topics in Statistical Process Control*, *Second Edition* © 2004 SPC Press. Used with permission.