# Avoiding Bias Correction Confusion

## When should we use the various bias correction factors?

Donald J. Wheeler

Recently I have had several questions about which bias correction factors to use when working with industrial data. Some books use one formula, other books use another, and the software may use a third formula. Which one is right? This article will help you find an answer.

Before we can meaningfully discuss different bias correction factors we need to understand what they do. To this end we must make a distinction between parameters for a probability model and statistics computed from the data. So we shall go back to the origin of our data and move forward.

A statistic is simply a function of the data. Data plus arithmetic equals a statistic. Since arithmetic cannot create meaning, it is the context for the data that gives specific meaning to any statistic. Thus, we will have to begin with the progression from a physical process to a probability model, and then we can look at how the notion of a probability model frames the way we use our statistics.

Assume that we have a process that is producing some product, and assume that periodic checks are made upon some product characteristic. These checks will result in a sequence of values that could be written as:

$$\{ X_1, X_2, \ldots, X_n, \ldots \}$$

When we compute descriptive statistics from the first $n$ values of this sequence there are two fundamental questions of interest: "How well do these statistics characterize the product produced during the time period covered?" and "Can we use these statistics to predict what the process will produce in the future?" Everything we do in practice hangs on these two questions, and the answers to these questions require an extrapolation. We have to extrapolate from the product we have measured to the product not measured. And this applies both to product produced in the past and product to be produced in the future. Therefore, we have to know when such extrapolations make sense.

Walter Shewhart provided a succinct answer to the question of extrapolation. Paraphrasing, he said: *A process will be predictable when, through the use of past experience, we can describe, at least within limits, how the process will behave in the future.* If the process has displayed predictable operation in the past, and if there is no evidence of unpredictable operation in the present, then the extrapolation from the data to the underlying process will be credible. Moreover, as long as the process continues to be operated predictably, the statistics based upon the historical data will continue to characterize the production process and the process outcomes.

However, when the process shows evidence of unpredictable operation, we are no longer justified in extrapolating from the data to the process. With the strong evidence that the process is changing that is provided by a process behavior chart, any attempt to use the historical data to predict the future can only be based on wishful thinking.

## FROM A PROCESS TO THE NOTION OF A DISTRIBUTION

When a production process is operated predictably it will be characterized by data that are homogeneous—measurements that display a consistent and recurring pattern of variation. This will result in a histogram that will essentially look more or less the same from time period to time period. This stable pattern of variation might then be approximated by some continuous probability model, *f(x)*. This conceptual probability model, *f(x)*, will be a mathematical function that can be characterized by parameters such as the center of mass, *MEAN(X)*, and the radius of gyration (also known as the standard deviation parameter) *SD(X)*. (In actual practice we neither need to draw the series of histograms, nor choose a probability model, but with a predictable process such actions would make sense.)

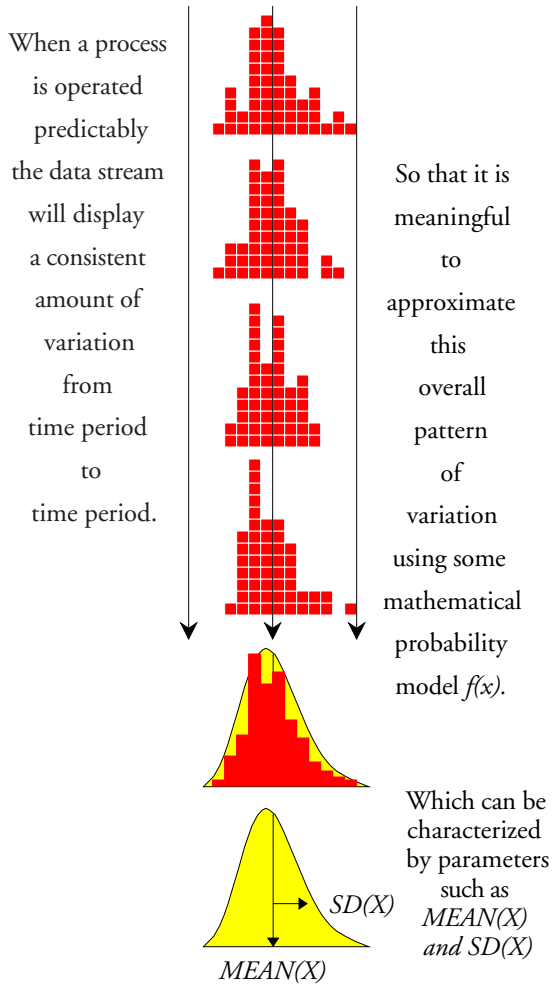


**Figure 1: When a Probability Model Makes Sense**

However, in practice, we will never have enough data to ever fully define a specific probability model in the manner implied by Figure 1. So even though we may have a predictable process, we will not be able to directly compute our process parameters. Fortunately we can still characterize our process using estimates of the process parameters based on the statistics computed from the process data. Before we look at how this is done we need to consider what happens when a process is operated unpredictably.

## WHEN THE NOTION OF PROCESS PARAMETERS EVAPORATES

When a process displays unplanned and unexpected changes it is said to be operated unpredictably. As a result the data will not be homogeneous and the histograms will be changing from time period to time period. So while we can always calculate our summary statistics, and while these summary statistics might somehow describe the past data, the idea that we can find a single probability model that will characterize the process outcomes no longer makes sense.

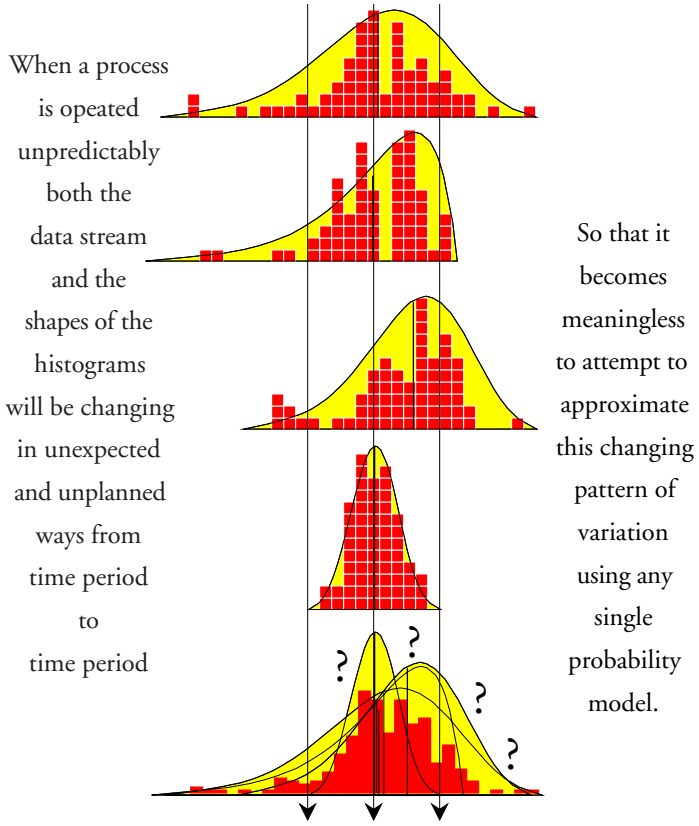


**Figure 2: When a Probability Model Does Not Make Sense**

When we cannot use a single probability model to describe the process outcomes the notion of process parameters will evaporate. Here we can no longer meaningfully talk about a process

mean, a process variance, or a process standard deviation parameter. While we may still compute our various statistics, and while may still use these statistics to characterize the process behavior as being either predictable or unpredictable, the statistics themselves will no longer represent specific process parameters. Our statistics only become meaningful estimators of process parameters when the process is being operated predictably. This is why is it crucial to make a distinction between statistics, which are functions of the data, and parameters, which are descriptive constants for a predictable process. All that follows will only make sense when we are working with a predictable process.

## ESTIMATORS FOR DISPERSION PARAMETERS

When we have a reasonably predictable process we generally want to characterize our process location and dispersion. The average statistic provides an intuitive estimator for the process mean, *MEAN(X)*. The complexity begins when we seek to characterize dispersion. First we have to decide if we need to estimate the process standard deviation, *SD(X)*, or the process variance, *VAR(X)*. Next we have to decide which dispersion statistic we shall use. While there are many possible choices here, the three most commonly used are the standard deviation statistic, $s$, the variance statistic, $s^2$, and the range statistic, $R$. Finally we have to choose whether to use a biased estimator or an unbiased estimator. Thus, we have a matrix of dispersion estimators as shown in Figure 3. The three quantities shown in the denominators are known as bias correction factors. A table of these factors is given at the end of the paper.

| Dispersion Statistic | Estimators for *SD(X)* | | Estimators for *VAR(X)* | |
|---|---|---|---|---|
| | Biased | Unbiased | Biased | Unbiased |
| $s$ | $s$ | $\frac{s}{c_4}$ | — | $s^2$ |
| $R$ | $\frac{R}{d_2^*}$ | $\frac{R}{d_2}$ | $\left(\frac{R}{d_2}\right)^2$ | $\left(\frac{R}{d_2^*}\right)^2$ |

**Figure 3: Choices When Estimating Dispersion**

So how do we choose between these various formulas? In most cases the formula is already incorporated into the technique, so you do not have to choose. But when given a choice the unbiased estimators are generally preferred.

## UNBIASED ESTIMATORS

An estimator of a parameter is said to be unbiased when it is, on the average, neither too large nor too small.

For example, in Figure 3, the variance statistic is an unbiased estimator for *VAR(X)* because the mean of the distribution of the variance statistic is equal to the parameter value.

$$MEAN(s^2) = VAR(X) = \sigma^2$$

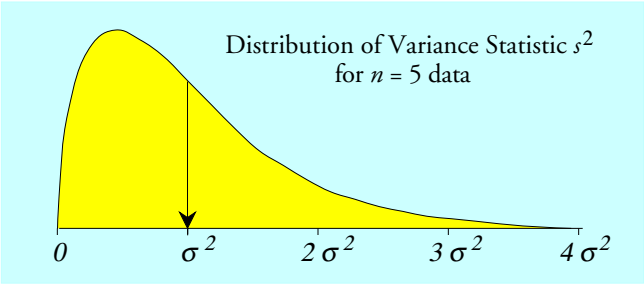


**Figure 4: The Distribution of the Variance Statistic when $n$ = 5**

Thus, the property of being unbiased is a property of the distribution of the formula for the statistic (i.e., a random variable) rather than being a property of the computed value (an observation on the random variable). An observed value for the variance statistic might fall anywhere under the curve in Figure 4, but the *mean value* of all possible observations of the variance statistic will be equal to the value of *VAR(X)*. Thus, we may write:

$$\textit{Unbiased Estimator for VAR(X)} \;=\; \textit{variance statistic} \;=\; s^2$$

BIASED ESTIMATORS

When we take the square root of the variance statistic we end up with the standard deviation statistic, *s*, which we often use as an estimator for the *SD(X)* parameter. Figure 5 shows the distribution for this estimator.

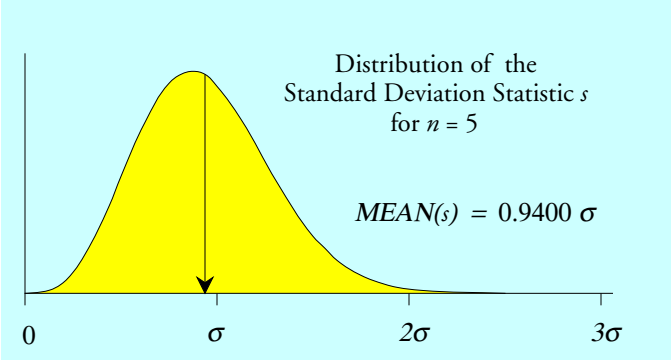


**Figure 5: The Distribution of the Standard Deviation Statistic when $n$ = 5**

The mean of the distribution in Figure 5 is only equal to 0.9400 times *SD(X)*. This means that, when $n$ = 5, the standard deviation statistic will turn out to be about 6 percent too small on the average, and so it is said to be a *biased estimator for SD(X)*.

$$\textit{Biased Estimator for SD(X)} \;=\; \textit{standard deviation statistic} \;=\; s$$

This bias is a consequence of the square root transformation. The property of being unbiased is only preserved by linear transformations (such as averaging) and it is lost whenever we perform a non-linear transformation (such as squaring or taking a square root). So while the variance statistic is an unbiased estimator for *VAR(X)*, the standard deviation statistic is a biased

estimator for *SD(X)*, and this property is inherent in the definition of an unbiased estimator. The square of an unbiased estimator will always be biased, and the square root of an unbiased estimator will always be biased.

## REMOVING THE BIAS

Whenever the mean value for the distribution of a statistic is some multiple of the value of a parameter, the statistic may be converted into an unbiased estimator for that parameter by the use of a bias correction factor. Figure 5 suggests that, for $n = 5$:

$$\textit{Unbiased Estimator for SD(X) when } n = 5 \;=\; \frac{s}{0.9400}$$

The usual symbol for these bias correction factors for the standard deviation statistic is a lower-case $c$ with a subscript of 4:

$$\textit{Unbiased Estimator for SD(X)} \;=\; \frac{s}{c_4}$$

The distribution of the range statistic for subgroups of size $n = 5$ is shown in Figure 6. On the average, subgroups of size five will have a range statistic that is 2.326 times the *SD(X)* parameter.

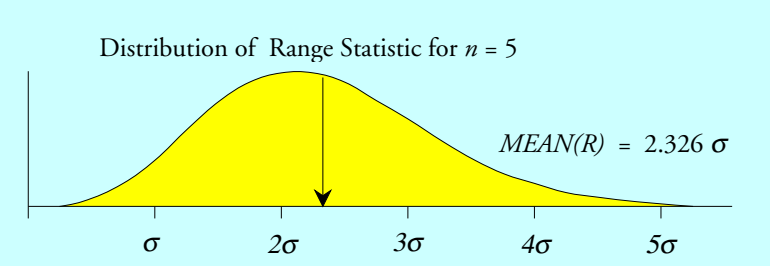


**Figure 6: The Distribution of the Range Statistic when $n = 5$**

Hence the bias correction factor for the range of five data is 2.326. Collectively we denote the bias correction factors for the range statistic using a lower-case $d$ with a subscript of 2:

$$\textit{Unbiased Estimator for SD(X)} \;=\; \frac{R}{d_2}$$

Figure 7 shows the bias-adjusted distributions of both the standard deviation statistic and the range statistic for $n = 5$. There is no practical difference between these unbiased estimators of *SD(X)*. They both provide essentially the same information with equal precision. This effective equivalence holds for subgroup sizes up to $n = 10$.

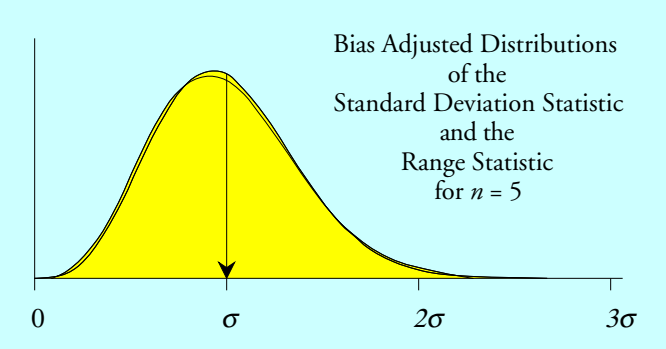


**Figure 7: Bias-Adjusted Distributions when $n = 5$**

Since the property of being unbiased is only preserved by linear transformations, we know that the square of either of the unbiased estimators for *SD(X)* given above will result in a biased estimator for *VAR(X)*. Specifically, the square of our bias-adjusted range will result in a biased estimator for *VAR(X)*.

$$\textit{Biased Estimator of VAR(X)} = \left(\frac{R}{d_2}\right)^2$$

AN UNBIASED ESTIMATOR FOR VARIANCE BASED ON THE RANGE

In 1950 another bias correction factor was defined which allowed a range or an average range to be used as an unbiased estimator for *VAR(X)*:

$$\textit{Unbiased Estimator of VAR(X)} = \left(\frac{R}{d_2^*}\right)^2$$

The following will illustrate how and why this correction factor differs from the one for estimating *SD(X)*. Figure 8 shows the distribution for the range statistic for one subgroup of size 2. There we see that the bias correction factor for estimating *SD(X)* when $n = 2$ is:

$$d_2 = \frac{2}{\sqrt{\pi}} = 1.128.$$

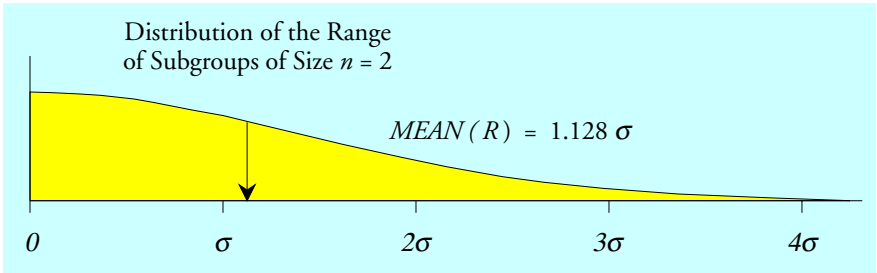


**Figure 8: Distribution of Range Statistic when $n = 2$**

When we square the range statistic in Figure 8 we end up with the transformed version of the chi-square distribution with one degree of freedom shown in Figure 9.

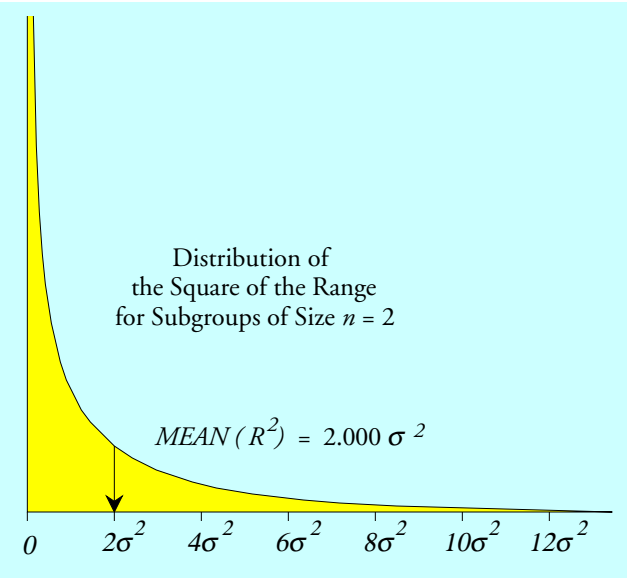


**Figure 9: Distribution of Range Statistic Squared when $n = 2$**

When $n = 2$, to obtain an unbiased estimator for *VAR(X)* based on the square of the range we will need to divide the squared range by 2.000. Alternatively, we could divide the range itself by the square root of 2 and then square the result. Thus, for $n = 2$ and $k = 1$:

$$\textit{Unbiased Estimator of VAR(X)} = \frac{R^2}{2} = \left(\frac{R}{1.414}\right)^2 = \left(\frac{R}{d_2^*}\right)^2$$

## BIAS CORRECTION FOR AVERAGE DISPERSION STATISTICS

Figure 3 and all of the preceding discussion was focused on what happens with dispersion statistics computed from one group of $n$ data. Here we will address what happens when we are working with $k$ subgroups of size $n$ and use an average dispersion statistic.

Figure 10 lists the estimators based on the average standard deviation statistic, the pooled variance statistic, and the average range statistic.

| Dispersion Statistic | Estimators for SD(X) Biased | Estimators for SD(X) Unbiased | Estimators for VAR(X) Biased | Estimators for VAR(X) Unbiased |
|---|---|---|---|---|
| *Avg. Std. Dev.* | $\bar{s}$ | $\frac{\bar{s}}{c_4}$ | $[\,\bar{s}\,]^2$ | — |
| *Pooled Var.* | $\sqrt{\overline{s^2}}$ | — | — | $\overline{s^2}$ |
| *Avg. Range* | $\frac{\bar{R}}{d_2^*}$ | $\frac{\bar{R}}{d_2}$ | $\left(\frac{\bar{R}}{d_2}\right)^2$ | $\left(\frac{\bar{R}}{d_2^*}\right)^2$ |

**Figure 10: Estimating Dispersion using $k$ Subgroups of Size $n$**

When we average $k$ dispersion statistics, each of which is based on the same amount of data, we can simply use the bias correction factor that is appropriate for a single one of the dispersion statistics to obtain an unbiased estimator. Remember, linear transformations do not affect the

property of being unbiased. Thus, when we average the first three unbiased estimators from Figure 3 we get the following:

$$\textit{Unbiased Estimators of SD(X)} = \frac{\overline{s}}{c_4} \quad \text{or} \quad \frac{\overline{R}}{d_2}$$

$$\textit{Unbiased Estimator of VAR(X)} = \overline{s^2}$$

For this reason we do not have to be concerned with the number of subgroups involved when averaging unbiased estimators. This is illustrated in Figure 11.

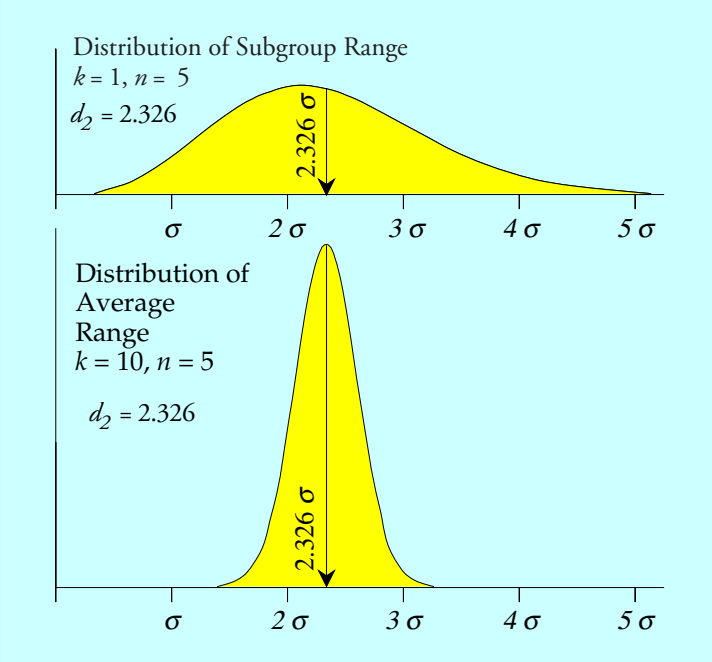


**Figure 11: Computing an Average Does Not Affect the Mean Value**

However, when using the range to estimate *VAR(X)* things are different. The fact that we are first going to compute the average range and then *square* it to obtain our unbiased estimator for *VAR(X)* makes the bias correction factor dependent upon both $n$ and $k$. The nonlinear transformation in the middle of the formula effectively creates this dependence.

When we square the random variables defined in Figure 11 we get the two distributions shown in Figure 12. Because the original distributions differ, the distributions of the squared random variables differ. Specifically, the distributions in Figure 12 have different mean values. When we take the square root of these mean values we find that the appropriate bias correction factor for a range-based estimate of *VAR(X)* is different when $n = 5$ and $k = 1$ from what it is when $n = 5$ and $k = 10$. Thus, the bias correction factors for estimating *VAR(X)* will depend upon both $n$ and $k$.

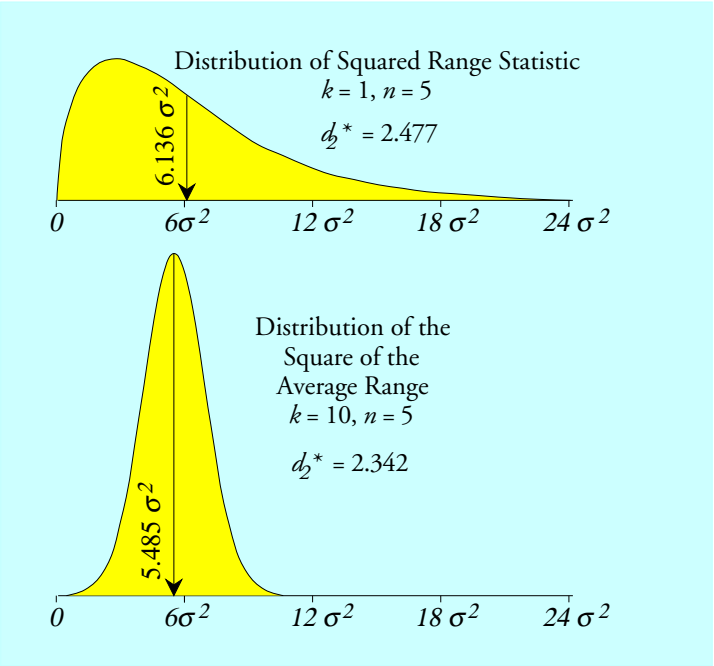


**Figure 12: Squaring the Distributions in Fig 11 Affects the Mean Values Differently**

## SO WHICH BIAS CORRECTION FACTOR SHOULD WE USE?

We prefer to use unbiased estimators simply because it always sounds better to be unbiased! However, it is important to note that in the distributions in Figures 4 through 12 an estimate can fall anywhere under the curves. Your *statistics* are not going to give you values that fall right at the mean of these distributions. So, in practice, the numbers we compute are *always biased* in the sense that they are not equal to the unknown parameter value. We have to be content with the knowledge that they are merely in the right ballpark.

To illustrate this point an artificial example is used (so that we actually know the parameter values). Ten values were obtained at random from a normal probability model having a mean of 15 and a variance of 4. They were split into two subgroups of size five and the dispersion statistics were computed.

$$\{\ 14.4,\ 12.9,\ 14.1,\ 18.2,\ 16.0\ \} \quad s = 2.046,\ s^2 = 4.186,\ R = 5.3$$
$$\{\ 15.2,\ 17.7,\ 14.8,\ 12.8,\ 13.4\ \} \quad s = 1.906,\ s^2 = 3.632,\ R = 4.9$$

The average standard deviation statistic is 1.975. The pooled variance statistic is 3.909. And the average range is 5.1. The bias correction factors are:

$$c_4 = 0.9400, \quad d_2 = 2.326, \quad \text{and} \quad d_2^* = 2.404.$$

The formulas in Figure 10 result in the nine estimates listed in Figure 13.

| Dispersion Statistic | Estimators for *SD(X)* | | Estimators for *VAR(X)* | |
|---|---|---|---|---|
| | Biased | Unbiased | Biased | Unbiased |
| Avg. Std. Dev. | 1.975 | 2.102 | 4.418 | — |
| Pooled Var. | 1.977 | — | — | 3.909 |
| Avg. Range | 2.121 | 2.193 | 4.808 | 4.501 |

**Figure 13: Nine Estimates of Dispersion using 2 Subgroups of Size 5**

*The knowledge that we have used a formula for an unbiased estimator does not convey any information about the distance between our estimate and the parameter value. Unbiased estimators are not categorically closer to the parameter value than are biased estimators. They just sound better.*

The estimators in Figure 10 are all within-subgroup estimators. In practice, these within-subgroup estimates of dispersion are essentially interchangeable, especially if we take into account their different degrees of freedom. Using one of these estimates in place of another estimate of the same parameter will not make any real difference in your analysis.

This makes the distinction about whether to use a biased or unbiased estimator a tempest in a teapot. The difference between the various estimators will always be trivial in comparison with the uncertainty in the dispersion statistics themselves. So, while we might prefer to use unbiased estimators, this is not something to obsess about. Use the formulas given as part of the technique, and stop trying to "fine tune" things with alternative formulas.

## Bias Correction Factors

Bias Correction Factors for Estimating *SD(X)*

| | $n = 2$ | $n = 3$ | $n = 4$ | $n = 5$ | $n = 6$ | $n = 7$ | $n = 8$ | $n = 9$ | $n = 10$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_4$ | 0.7979 | 0.8862 | 0.9213 | 0.9400 | 0.9515 | 0.9594 | 0.9650 | 0.9693 | 0.9727 |
| $d_2$ | 1.128 | 1.693 | 2.059 | 2.326 | 2.534 | 2.704 | 2.847 | 2.970 | 3.078 |

$d_2$* Bias Correction Factors for Estimating *VAR(X)*

| $k$ | $n = 2$ | $n = 3$ | $n = 4$ | $n = 5$ | $n = 6$ | $n = 7$ | $n = 8$ | $n = 9$ | $n = 10$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.414 | 1.906 | 2.237 | 2.477 | 2.669 | 2.827 | 2.961 | 3.076 | 3.178 |
| 2 | 1.276 | 1.806 | 2.149 | 2.404 | 2.603 | 2.767 | 2.905 | 3.024 | 3.129 |
| 3 | 1.227 | 1.767 | 2.120 | 2.378 | 2.580 | 2.746 | 2.886 | 3.006 | 3.112 |
| 4 | 1.206 | 1.749 | 2.105 | 2.365 | 2.569 | 2.736 | 2.876 | 2.997 | 3.104 |
| 5 | 1.189 | 1.738 | 2.096 | 2.358 | 2.562 | 2.729 | 2.870 | 2.992 | 3.098 |
| 6 | 1.179 | 1.731 | 2.090 | 2.352 | 2.557 | 2.725 | 2.867 | 2.988 | 3.095 |
| 7 | 1.172 | 1.726 | 2.086 | 2.349 | 2.554 | 2.722 | 2.864 | 2.986 | 3.093 |
| 8 | 1.167 | 1.722 | 2.082 | 2.346 | 2.552 | 2.720 | 2.862 | 2.984 | 3.091 |
| 9 | 1.163 | 1.718 | 2.080 | 2.344 | 2.550 | 2.718 | 2.860 | 2.982 | 3.089 |
| 10 | 1.159 | 1.716 | 2.078 | 2.342 | 2.548 | 2.717 | 2.859 | 2.981 | 3.088 |
| 15 | 1.149 | 1.708 | 2.071 | 2.337 | 2.543 | 2.713 | 2.855 | 2.977 | 3.085 |
| 20 | 1.144 | 1.705 | 2.068 | 2.334 | 2.541 | 2.710 | 2.853 | 2.975 | 3.083 |
| 25 | 1.141 | 1.702 | 2.066 | 2.332 | 2.540 | 2.709 | 2.852 | 2.974 | 3.082 |
| 30 | 1.139 | 1.701 | 2.065 | 2.331 | 2.539 | 2.708 | 2.851 | 2.974 | 3.081 |
| 40 | 1.136 | 1.699 | 2.064 | 2,330 | 2.538 | 2.707 | 2.850 | 2.973 | 3.081 |
| $\infty$ | 1.128 | 1.693 | 2.059 | 2.326 | 2.534 | 2.704 | 2.847 | 2.970 | 3.078 |