Obey Gravity

It's the law.

Donald J. Wheeler

There is no virtue in obedience when we do not have a choice. But when we have a choice it helps to understand both the law and the reason behind the law. This column is about bad choices that are being made on a daily basis by the users of statistical software. These bad choices violate the laws of physics, mathematics, statistics, and possibly even gravity itself. For enlightenment, read on.

In statistics classes we introduce the concept of probability models and their parameters, then we quickly move on to descriptive statistics and other functions of the data. As a result, in the mind of many students, there is no little amount of confusion over the roles these parameters and statistics play. As a result, the statistics end up being used in inappropriate ways and these uses end up contributing even more confusion to the whole subject.

PARAMETERS

On the mathematical plane we use probability models to represent random phenomena that exist in the real world. They allow us to develop analysis techniques that will work with real data. This mathematical modeling of the real world is a mainstay of physics and is the heart of mathematical statistics.

The first parameter for a probability model is the mean. The mean is the balance point for the probability model and is defined as follows. Let *f(x)* define the probability density function for the probability model, then the mean of that probability model is found by integrating the product of *X* and *f(x)* over all values for *X*:

$$MEAN(X) \quad = \quad \int x \; f(x) \; dx$$

The second parameter for a probability model is the variance. The variance is the rotational inertia about the mean for the probability model. It is found by integrating the product of *f(x)* and [the square of the distance between *X* and *MEAN(X)*] over all values for *X*.

$$VAR(X) \quad = \quad \int [\, x - MEAN(X) \,]^2 \; f(x) \; dx$$

A related parameter for dispersion is the standard deviation which is defined as the square root of the variance parameter. As the square root of the rotational inertia it has no intuitive interpretation, but it is useful in various mathematical operations.

$$SD(X) \quad = \quad \sqrt{VAR(X)}$$

Now when was the last time you evaluated an integral? Clearly, parameters are not values that we compute in practice. However, these parameters define the location and dispersion of a probability model in a complete and concise manner. For this reason, we are often interested in estimating these unknown parameter values. To this end we use statistics.

STATISTICS

Statistics are functions of the data. Data plus arithmetic equals a statistic. Exactly what a statistic represents will depend upon the context for the data. The most common measure of location is the average statistic. When all of our data comes from the same source, we are justified in using the average statistic to estimate the *MEAN(X)* parameter. For example, if we measured the heights of people in a college classroom we could use the average height to approximate the mean height for the population of adults as a whole. However, if we tried this in a first-grade classroom, the average height would neither represent the mean height for the six-year old kids nor the mean height for the teachers. Thus, it is the context of the data that determines when it is appropriate to use statistics to estimate parameter values. If the context is lacking the statistics will not provide useful estimates of parameters even though they may still describe properties of the data.

Common measures of dispersion are the range statistic and the standard deviation statistic. When working with a homogeneous collection of data we can obtain useful unbiased estimates of the standard deviation parameter, *SD(X),* by using either the standard deviation statistic, s, or the range statistic, *R*:

$$\textit{Unbiased Estimate of } SD(X) \;=\; \frac{s}{c_4} \quad or \quad \frac{R}{d_2}$$

where c_4 and d_2 are the appropriate bias correction factors. For unbiased estimates of the variance parameter, *VAR(X),* we need to use the standard deviation statistic and range statistic slightly differently:

$$\textit{Unbiased Estimate of } VAR(X) \;=\; s^2 \quad or \quad \left[\frac{R}{d_2{}^*}\right]^2$$

where $d_2{}^*$ is the bias correction factor for estimating variances. The point here being that the square of an unbiased estimator of the standard deviation parameter is not an unbiased estimator of the variance parameter. There are rules that must be obeyed when working with measures of dispersion. If you do not know the rules you can make mistakes without knowing it. One place that these mistakes show up is when people try to combine measures of dispersion and when they try to compare dispersion to other quantities.

COMBINING DISPERSIONS FOR DIFFERENT QUANTITIES

Consider the problem of stacking parts. Part *X* fits on top of Part *Y*. What can we say about the total height of the assembly? The mean height of the assembly will be found by adding the mean values for each part:

$$MEAN(\,X + Y\,) \;=\; MEAN(X) \;+\; MEAN\,(Y)$$

So, if we have estimates for each of these means, we simply add them up to estimate the mean height of the assembly:

$$Est.\ MEAN(\ X + Y) \ = \ Est.\ MEAN(X) \ + \ Est.\ MEAN\ (Y)$$

However, when working with dispersion, the formula we use will depend upon the correlations between X and Y. In the simplest case, where X and Y are uncorrelated, we know that the variance for the height of the assembly is:

$$VAR(\ X + Y\ ) \ = \ VAR(X) \ + \ VAR(Y)$$

Now, as every student of high-school geometry knows, this equation means that the standard deviation parameter of the assembly heights is not, and can never be, equal to the sum of the standard deviation parameters for X and Y.

$$SD(\ X + Y\ ) \quad \text{IS NOT EQUAL TO} \quad SD(X) \ + \ SD(Y)$$

So, if we want to estimate the standard deviation parameter for the assembly we will have to first estimate the variance parameter and then find the square root.

$$Est.\ SD(\ X + Y\ ) \ = \sqrt{\ Est.\ VAR(\ X + Y\ )}$$

Therefore, whenever we work with measures of dispersion for two or more quantities, we cannot simply add up the standard deviation statistics, or form ratios using the standard deviation statistics, and expect the results to be anything other than complete nonsense. We have to add (or form ratios) using the estimates of the variance parameters. However, in case after case, software packages compute ratios of estimated standard deviations of different quantities and offer them up for consumption by the users.

For example, the ratios computed in the AIAG Gauge R&R study are erroneous because they compare standard deviations of different quantities and interpret the resulting trigonometric functions as proportions when they are not proportions. (For more on this see my columns for January 2011 and December 2013.)

Also, when we compare the standard deviations with fixed quantities the same problem arises. Consider the widely used precision to tolerance ratio, P/T. Here a multiple of the standard deviation of measurement error is compared to the specified tolerance. However, the specifications are not applied to measurement error, they are applied to the product measurements, X. Now these product measurements may be thought of as being made up of the product value, Y, plus the measurement error, E.

$$X \ = \ Y \ + \ E$$

Since we assume the product values are independent of the measurement errors, the variance for X will be the sum of the variances for Y and E. Thus, the standard deviation of E does not affect the standard deviation of X in an additive manner. Therefore, the standard deviation for E does not affect any constants that apply to the product measurements in an additive manner. This lack of additivity makes any attempt to directly compare the standard deviation of E to the specified tolerance into an exercise in nonsense arithmetic. The result will not be a proportion, but rather some complex trigonometric function. Yet when computing traditional guard-bands and when computing the precision to tolerance ratio we are, in effect, assuming that the standard deviation for E affects the product measurements in an additive manner. This is what makes these traditional computations incorrect and misleading. (For more on this see my columns for June and July of 2010.)

COMBINING DISPERSION STATISTICS FOR ONE QUANTITY

On the other hand, when you possess multiple dispersion statistics, where each statistic is an independent estimate of one and the same parameter, then you can directly combine these statistics into a composite estimate. Consider the traditional situation for an average and range chart where we have *k* subgroups of size *n*, and for each of these subgroups we have a subgroup range. Here we can average the ranges, divide by the common value of d_2 that is appropriate for each of the ranges, and obtain a composite estimate of the standard deviation parameter for *X*.

$$\textit{Unbiased Est. SD(X)} \quad = \quad \frac{\bar{R}}{d_2}$$

In a similar manner we could average the subgroup standard deviation statistics, divide by their common bias correction factor, and obtain a composite estimate for the standard deviation parameter for *X*.

$$\textit{Unbiased Est. SD(X)} \quad = \quad \frac{\bar{s}}{c_4}$$

In both of these formulas we have effectively averaged multiple estimates of one standard deviation parameter. It is the fact that these statistics are all estimates of the same parameter that makes the averages above correct and appropriate.

Thus, there are times when we can add standard deviation statistics, and there are times when we cannot. This has nothing to do with the statistics themselves, but rather with what the statistics represent. This is not some sideline, or some inconsequential result. It is an essential part of the foundation of modern statistical analysis.

MODERN STATISTICAL ANALYSIS

The purpose of any statistical analysis is to separate the potential signals from the probable noise. Different techniques look for different types of signals, and they use different ways to filter out the noise, but the underlying principle is the separation of the signals from the noise.

The fundamental principles of modern statistical analysis were developed between 1875 and 1925. Prior to that time techniques like Pierce’s Criterion for identifying outliers used a global measure of dispersion, computed using both noise and signals combined, as the mechanism for filtering out the “noise.” Since the signals contaminate all such filters, these techniques proved to be very weak. While they would occasionally detect very large signals, they would miss many other signals.

Around the beginning of the Twentieth Century an alternate approach was beginning to be used. This approach divided the data into subgroups where each subgroup was thought to be relatively homogeneous, and where the potential signals were isolated to occur between the subgroups. With this structure the noise would be filtered out using the within-subgroup variation, and the variation between the subgroups would contain the potential signals. By 1925 this idea was fully developed and was incorporated into the analysis of variance by Sir Ronald Fisher and was also being used as the foundation of the process behavior chart by Walter

Shewhart. Today this approach to separating the potential signals from the probable noise has been extended and is found in many different analysis techniques. In addition to the many varieties of ANOVA, it is also the foundation of the Analysis of Means (ANOM) and related techniques. Figure 1 contains a table listing various within-subgroup estimators of dispersion parameters. (For more about between-subgroup and within-subgroup variation see my column for October 2013.)

Regardless of whether we are estimating a standard deviation parameter, or a variance parameter, and regardless of whether we are using a biased or unbiased estimator, all 26 of the estimators in Figure 1 are based on either an average or a median of multiple measures of dispersion. Twenty-three of these involve averages or medians of estimators of the standard deviation parameter. Only the last row uses estimators of the variance parameter. The use of averages or medians of within-subgroup estimators to filter out the noise is the hallmark of all modern statistical analysis techniques.

Name of	Estimators for *SD(X)*		Estimators for *VAR(X)*	
Estimator	Biased	Unbiased	Biased	Unbiased
Average Range	$\frac{\bar{R}}{d_2{}^*}$	$\frac{\bar{R}}{d_2}$	$\left(\frac{\bar{R}}{d_2}\right)^2$	$\left(\frac{\bar{R}}{d_2{}^*}\right)^2$
Median Range	----	$\frac{\tilde{R}}{d_4}$	$\left(\frac{\tilde{R}}{d_4}\right)^2$	----
Average Two-Point Moving Range		$\frac{\overline{mR}}{1.128}$	$\left(\frac{\overline{mR}}{1.128}\right)^2$	$\left(\frac{\overline{mR}}{1.414}\right)^2$
Median Two-Point Moving Range		$\frac{\widetilde{mR}}{0.954}$	$\left(\frac{\widetilde{mR}}{0.954}\right)^2$	----
Average Std. Dev.	$\bar{s}$	$\frac{\bar{s}}{c_4}$	$\left(\frac{\bar{s}}{c_4}\right)^2$	----
Median Std. Dev.	$\tilde{s}$	$\frac{\tilde{s}}{c_6}$	$\left(\frac{\tilde{s}}{c_6}\right)^2$	----
Average RMS Dev.	$\bar{s}_n$	$\frac{\bar{s}_n}{c_2}$	$\left(\frac{\bar{s}_n}{c_2}\right)^2$	----
Median RMS Dev.	$\tilde{s}_n$	$\frac{\tilde{s}_n}{c_1}$	$\left(\frac{\tilde{s}_n}{c_1}\right)^2$	----
Pooled Variance	$\sqrt{\overline{s^2}}$	$\frac{\sqrt{\overline{s^2}}}{c_4{}'}$	----	$\overline{s^2}$

Figure 1: Within-Subgroup Estimators of Dispersion Parameters

So while there are many different formulas for within-subgroup estimators of dispersion parameters, this multiplicity occurs within the framework of estimating a single dispersion parameter. This is the only context where we can work directly with estimates of the standard deviation parameter. Moreover, due to considerations of robustness, the Pooled Variance estimators are only appropriate when working with the Analysis of Variance. They are inappropriate for use with ANOM or process behavior charts. (For more on this see my columns for January and February 2010.)

As indicated earlier, when working with multiple dispersion parameters, our choices are much more constrained, and the formulas that are appropriate will depend upon the context. In general, it will be the variances that are additive, and so you will need to work with estimates of the variance parameters. In the case of correlations between the variables involved, adjustments to the sum of the variances are required, and the advice of a professional statistician should be sought. In no case should you try to make adjustments, or make comparisons, using estimates of the standard deviation parameters.

So, as long as you continue to obey gravity, you need also to obey the laws of mathematics, physics, and statistics in your computations. Rotational inertia is additive. The square root of rotational inertia is not additive. Beware of all simple comparisons and ratios involving estimates of standard deviation parameters. They have a very high probability of being incorrect.