# Diagnosing the Top-Level Coefficient of Variation: An Alternative Approach

**Daniel J. Andelin**
**Operations Research Analyst**

**United States Department of the Navy**
daniel.andelin@navy.mil

## ABSTRACT

The coefficient of variation (CV), defined as the standard deviation divided by the mean, is a useful metric to quantify the uncertainty inherent in a probability distribution, as it provides a relative (and thereby more intuitive) measure of spread than the variance or standard deviation. In the field of cost analysis, the CV can be used as a diagnostic to determine the level of risk and uncertainty built into an estimate and whether those levels fall within the expected range for a program of given scope and complexity and at a given stage in its life cycle.

The CV is easy to calculate, even for complicated distributions, and is a standard output for risk-based estimating software, such as @Risk or ACEIT's RI$K. However, it is not always intuitive to understand what factors contribute to the overall CV, or why a particular estimate may have a CV that is lower (or higher) than expected. When conducting *ad hoc* diagnostics, it is tempting to treat the CV of a parent Work Breakdown Structure (WBS) element as approximately the weighted average of its children's respective CVs. This approach is fundamentally flawed, as it neglects both the widening effect of correlation and the narrowing effect of adding distributions (due to the fact that the standard deviation adds in quadrature).

An alternate approach to diagnosing the CV is based on representing the parent CV as a function of the relative size and CV of the child elements and the correlation between those elements. The functional form of this representation is elegant and leads to a natural treatment of the parent CV in three parts: (1) the weighted average of the child CVs, (2) an adjustment to account for summing in quadrature, and (3) an adjustment to account for correlation. Rules of thumb are given to facilitate "back of the envelope" calculations, and a graphical display of more precise results is suggested for briefing to decision makers and other stakeholders.

## INTRODUCTION

There is an old adage in the field of cost analysis that the only thing one can know about a point estimate is that it is always wrong. This tired truism is often used in the context of basic courses or discussions on cost risk and uncertainty to introduce the concept of an estimate as a range of possible values with an associated probability distribution, as opposed to a hard, fast number (which, indeed, will invariably be wrong). Probabilistic estimates are created by assigning uncertainty to model inputs, typically by defining the shape and spread of common probability density functions (e.g. normal, triangle, etc.) around the input point estimate, which generally serves as an anchor or measure of central tendency of the distribution. Random draws are taken from these input distributions and manipulated via the same math as the point estimate, yielding a simulated result for the total value of the estimate. The results of many, many consecutive iterations (usually on the order of thousands) produce a range of values for the total estimate with an associated probability density function or distribution.

Analysts and stakeholders alike should be interested in not only the central or most likely values of the distribution, but also its spread. A common measure of spread is the coefficient of variation (CV), defined as the standard deviation divided by the mean. The CV is a useful metric because it provides a relative (and thereby more intuitive) measure of how widely values in a probability distribution vary. (By its definition, the CV describes what percentage of the mean value is represented by a single standard deviation of the data set.) Because of its relativity, the CV can be easily compared among similar distributions or to industry standards or rules-of-thumb, such as those given in the Air Force Cost Risk and Uncertainty Handbook (AFCAA 2007, 26-27) to ascertain whether an appropriate amount of uncertainty has been captured by the cost model.

When the CV of an estimate is far from the value one would expect, given a program's size, complexity, maturity, etc., diagnostics are warranted to answer two basic questions. First, why is the CV so low or so high? Second, what, if anything, can (or should) be done to fix it? (These are the types of questions that a senior analyst or stakeholder may use to ambush junior analysts during an estimate briefing!) Variance analyses and sensitivity tests can be used to better understand what drives the CV, but these tests can be computationally quite time consuming. Certainly, they are not an option in a briefing setting, or any other time a quick diagnostic is needed.

There is the temptation to attempt to deconstruct the top-level CV by treating it as a simple weighted average of the next lower level elements' coefficients of variation, possibly using the point estimates as the weights. This approach is intuitive and lends itself to fairly easy mental math. Unfortunately, it is also fundamentally flawed, not only misrepresenting the mathematics of summing distributions, but also completely ignoring the effects of correlation between the

elements. However, though mathematically unsound as an approximation of the top-level CV, an intuitive weighted average coefficient of variation ($CV_{w.ave}$) with judiciously chosen weighting factors can be used as a logical starting point for understanding what drives the true CV, provided that certain adjustments are made.



If the weighting factors are chosen properly, the true, top-level CV can be expressed in terms of the lower-level weighted CVs. This expression separates the effects of correlation from the effects of the lower-level CVs alone and shows how to properly sum the weighted lower-level coefficients. In this way, there emerges a logical mental path toward what drives the true CV and how to address the questions of why a CV may be unexpectedly low (or high) and what may be done to bring it to a more reasonable level.

## THE WEIGHTED AVERAGE CV

Consider a simple, generic Work Breakdown Structure (WBS), consisting of a top-level parent element with *n* children. By definition, the value of the parent is the sum of the children's values. Because of this relationship, it is natural to assume that the CV of the parent should fall somewhere around the average value of its children's CVs. (Remember, CV is a relative measure of spread.) Therefore, if you know the spread of the children (or at least of the "big-ticket" items), you should be able to estimate the CV of the parent with little effort. However, this simply isn't the case for two fundamental reasons.

First, the CV of a distribution is, by definition, directly dependent on the distribution's standard deviation. Unlike the mean, the standard deviation (and by extension the CV) does not add linearly. That is, the standard deviation of the sum is *not* equal to the sum of the standard deviations. Rather, it is the *variance* (or the square of the standard deviation) that sums linearly (for independent variables). Consequently, the standard deviation sums *quadratically*, meaning the standard deviation of the sum is the root-sum-square[1] of the addend standard deviations. The root-sum-square is less than or equal to the simple sum (Salas, Hille, and Etgen 1999, 18), which means that adding multiple independent distributions has a narrowing effect on the sum distribution.

Second, a simple weighted average does not take into account the effect of correlation between the child WBS elements. Even if one deals with the square of the CV to avoid the narrowing effect discussed previously, the variances only add linearly in the absence of correlation. Correlation tends to have a widening effect of the sum distribution. These two effects, the

[1] The root-sum-square of a set of variables is the square root of the sum of the square of each variable.

narrowing from adding the standard deviation in quadrature and the widening from correlation, can mask each other to some degree, obscuring the true nature of the top-level CV and how it is driven by lower-level spread.



Though not a good approximation of the parent CV, the weighted average of the child CVs is not without value. As mentioned, it is simple, intuitive, and fairly easy to calculate inside one's head. As such, $CV_{w.ave}$ is a convenient place to begin when trying to understand the true CV. In addition, the true CV can be written in terms of the weighted child CVs—the addends that sum together to obtain $CV_{w.ave}$.

## THE PARTIAL COEFFICIENT OF VARIATION

The terms which, when summed, equal $CV_{w.ave}$ may be called the *partial coefficients of variation* (pCV) of each child element. The pCV is a weighted coefficient of variation and has little meaning as a stand-alone entity. However, the collective[2] pCV are the key to describing the parent CV in terms of its children.

It is important to be clear on what weighting factors are used to calculate the pCV. Because the mean is the only point on a distribution that sums linearly, it makes sense to use the relative means (not the point estimates) as the weights. Specifically, define a weight $p_j$ corresponding to a child WBS element, denoted by $x_j$ to be the mean of $x_j$ divided by the mean of the parent, denoted by capital $X$:

Let $X$ be the parent of $n$ children, each denoted by $x_j$, in a typical Work Breakdown Structure (WBS). In general, the value of the parent is equal to the sum of the children, so that

$$p_j \equiv \frac{\overline{x_j}}{\overline{X}} \qquad \text{(Eq. 1)}$$

The partial CV of an element $x_j$ (denoted by $pCV_j$) is defined as the product of that element's CV ($CV_j$) and its weighting factor ($p_j$):

$$pCV_j \equiv p_j \cdot CV_j \qquad \text{(Eq. 2)}$$

The weights all add to unity, so the weighted average CV is given as simply the sum of the partial CV:

[2] Note that pCV can be used to denote the singular or plural, as in partial coefficients of variation.

$$CV_{w.ave} = \sum_{j=1}^{n} (pCV_j) \qquad \text{(Eq. 3)}$$



Again, it must be understood that, except in the most special of circumstances, this is *not* equal to the true CV of the parent distribution. However, as stated previously, the true CV can be expressed in terms of the pCV.

## THE pCV REPRESENTATION OF THE PARENT CV

To arrive at the partial CV representation of the coefficient of variation for a parent WBS element ($X$), start by expressing the variance ($\sigma^2$) of $X$ as a sum of each pair of child element standard deviations ($\sigma_j$, $\sigma_k$) multiplied by each other and by the correlation between the two elements ($r_{j,k}$). (Be sure to include each element paired with itself!)

$$\sigma_X^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} r_{j,k} \sigma_j \sigma_k \qquad \text{(Eq. 4)}$$

(Eq. 4) is given in equivalent form by Alfred Smith (Smith 2011, 22) and is derived independently in this paper's appendix.

From the definition of the CV, the standard deviation $\sigma$ is equal to the product of the CV and the mean. Plugging this product into (Eq. 4) for every instance of $\sigma$, and dividing out by the square of the total mean $(\bar{X}^2)$ yields:

$$CV_X^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} r_{j,k} \left(\frac{\bar{x}_j}{\bar{X}}\right) CV_j \left(\frac{\bar{x}_k}{\bar{X}}\right) CV_k \qquad \text{(Eq. 5)}$$

Applying the definition of the weights in (Eq. 1) and the definition of the pCV from (Eq. 2) and taking the square root of both sides, the following expression for the parent CV emerges:

$$CV_X = \sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} r_{j,k} (pCV_j)(pCV_k)} \qquad \text{(Eq. 6)}$$

This is the pCV representation of the coefficient of variation in compact form.

The true math enthusiast may notice that the double sum under the radical is the equal to the inner product of the $n$-dimensional vector, whose entries are given by the set of pCV, with itself,

after being acted upon by a Hermitian operator, $\hat{R}$ , defined by the $n \times n$ fully populated correlation matrix.



$$CV_X = \sqrt{(pCV_T \ \hat{R} \ pCV)} \quad \text{(Eq. 7)}$$

In addition to the compact, elegant notation, this vector form of the pCV representation can come in handy when working with spreadsheet software, such as Excel®, and is particularly useful in at least one special case.

The utility of the pCV representation in (Eq. 6) is made particularly clear by separating the terms for which $j = k$ from the cross terms. These are the terms which correspond to the diagonal entries of the correlation matrix, which are each equal to one. Furthermore, the inherent symmetry in the off-diagonal entries of the correlation matrix (i.e. $r_{j,k} = r_{k,j}$) causes each cross term to repeat twice. Thus, the terms under the radical in (Eq. 6) can be separated into a sum over the diagonal entries and a double sum (by row and column) over the off-diagonal entries in the upper-right of the correlation matrix, multiplied by the corresponding pCV:

$$CV_X = \sqrt{\sum_{j=1}^{n} pCV_j^2 + 2\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} r_{j,k}(pCV_j)(pCV_k)} \quad \text{(Eq. 8)}$$

Again, the factor of 2 in the cross terms is a result of the symmetry in the correlation matrix.

## THE ZERO-CORRELATION CV

In the absence of correlation, the off-diagonal entries of the correlation matrix disappear, eliminating the cross-terms in (Eq. 8) and reducing the CV to root-sum-square of the pCV. Recall from earlier discussions that this is proper way to average the child-level coefficients of variation, as it accounts for the narrowing effect of adding the standard deviation in quadrature. Because it ignores the effects of correlation, call this expression the zero-correlation coefficient of variation ($CV_{zero\ corr}$).

$$CV_{zero\ corr} = \sqrt{\sum_{j=1}^{n} pCV_j^2} \quad \text{(Eq. 9)}$$

Both $CV_{w.ave}$ and $CV_{zero\ corr}$ are easy to calculate with spreadsheet software or even a standard scientific calculator. However, $CV_{w.ave}$ is considerably easier to deal with mentally. Not only is the arithmetic linear, but visualizing a simple weighted average is much more natural than a root-

sum-square. Because of these advantages, it is helpful to understand how $CV_{w.ave}$ and $CV_{zero\ corr}$ differ and, more importantly, how to go from one to the other.



Define a factor $f$ as the ratio of the zero-correlation CV to the weighted average CV so that

$$CV_{zero\ corr} = f \cdot CV_{w.ave} \qquad \text{(Eq. 10)}$$

Note that because $CV_{w.ave}$ is a linear combination of the pCV, $f$ can be distributed throughout its terms. In this way, $CV_{zero\ corr}$ can be thought of as either an adjusted $CV_{w.ave}$ or as the weighted average of the adjusted child-level CVs (or equivalently as the sum of the adjusted pCV). In each case, the adjustment is made via multiplication by $f$.

The interpretation of $f$ is that $(1 - f)$ represents the fraction by which the CV is reduced due to summing the standard deviation in quadrature. This factor is not constant and depends on the number of child elements $n$ and how evenly the pCV are distributed. (The minimum value of $f$ is realized when all the pCV are equal. Conversely, $f$ is maximized when one pCV dominates all others.)

Of course, it is simple to back into $f$ by calculating both $CV_{w.ave}$ and $CV_{zero\ corr}$ and taking the ratio. In fact, if time and resources permit, this is a fast, easy way to picture the reduction in CV that occurs from summing independent distributions. However, if a quick, "back of the envelope" analysis is needed, it is easiest to start with $CV_{w.ave}$ and rely of rules of thumb to approximate $f$.

Because $f$ depends on how the pCV are distributed, it is helpful to normalize the pCV so that they sum to one. Each normalized pCV will represent the percent of the total $CV_{w.ave}$ represented by each raw pCV. By way of notation, define a variable $q$ to be the ratio of a single pCV to the sum of all pCV:

$$q \equiv \frac{pCV}{\sum pCV} = \frac{pCV}{CV_{w.ave}} \qquad \text{(Eq. 11)}$$

and define $q_1$ to be the largest value of $q$ (i.e. the leading normalized pCV) for a given set of WBS elements.

Figure 1 illustrates the behavior of $f$ versus the leading normalized pCV ($q_1$) for several values of $n$. (The data in the graph consist of randomly distributed $q$ for a number of arbitrary sets of child elements of various sizes.) The data for each value of $n$ follow a distinct curve that is "sharper" for smaller values of $n$. Notice that for smaller values of $q_1$, not only do the curves fan out, but

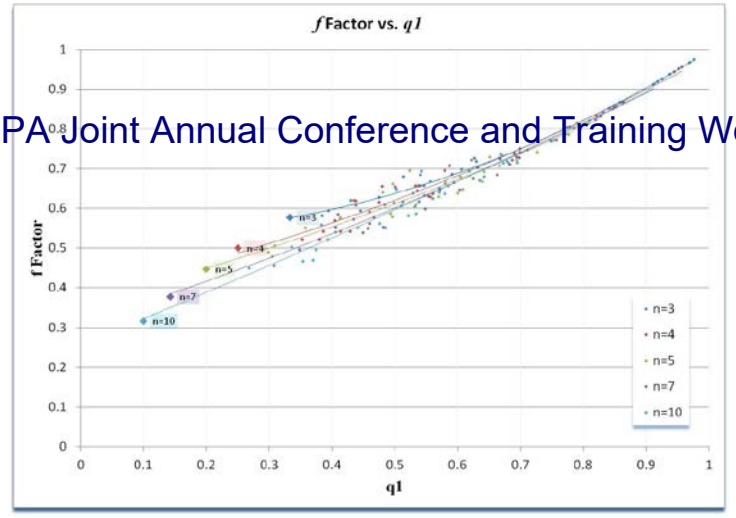


Figure 1. The behavior of *f* depends on the number and distribution of the pCV.

the data are more loosely scattered about their respective curves. As the leading *q* dominates less, the relative influence of *n* increases and the variability in the distribution of the other-than-leading *q*'s becomes more pronounced. Contrariwise, for $q_1$ closer to 1, the data converge, following a nearly linear trend, as $q_1$ dominates all other factors. The slope of the *f*-curve for high $q_1$ is roughly equal to one.

The graph of *f* vs. $q_1$ can be more or less divided into three regions, based on the value of $q_1$, for the purposes of developing rules of thumb for approximating *f*. The first region is that of high $q_1$, greater than about 0.70. In this region, $f \sim q_1$. The data in the middle region ($q_1$ from about 0.4 to 0.7) are more scattered, so the approximate values for *f* are less precise. A rough rule of thumb is that $f \sim$ ½ for low values of $q_1$ with high *n* and $f \sim$ ¾ for high values of $q_1$ with low *n*. For the middling range, *f* can be approximated by ~⅔. (The approximations may overstate *f* a bit, but they are easy numbers to work with and should be sufficiently accurate for a rough mental calculation.) Finally, in the low $q_1$ realm (where $q_1 \sim n^{-1}$), the normalized pCV tend to be more or less equal, and the value of *f* can be approximated by its minimum value, $1/\sqrt{n}$. These rules of thumb are summarized in Table 1.

| | $q_1$ | < 0.40 | 0.40 – 0.70 | > 0.70 |
|---|---|---|---|---|
| Low *n* | High $q_1$ | $f \sim 1/\sqrt{n}$ | $f \sim$ ¾ | $f \sim q_1$ |
| Med *n* | Med $q_1$ | | $f \sim$ ⅔ | |
| High *n* | Low $q_1$ | | $f \sim$ ½ | |

Table 1. Rules-of-thumb for approximating *f*.

## CORRELATION EFFECTS

Of course, the zero-correlation case is theoretical only. All real estimates must contain some degree of correlation between WBS elements, and the effects of the presence of correlation (which tends to widen the parent distribution) must be added to $CV_{zero\ corr}$ in order to obtain the true CV of the parent element.

The double sum in (Eq. 8) may be called the "correlation effect term," since it arises from the correlation between the elements and clearly depends on $r_{j,k}$. However, a dose of caution should be applied to such nomenclature. The terms in (Eq. 8) are added under a radical, so one cannot simply add the zero-correlation and correlation effect terms to obtain the true CV. Furthermore, it is convenient to refer to the additional CV that results from the presence of correlation as the "correlation effect," but this delta is not equal to the so-called correlation effect *term* shown in (Eq. 8). To avoid confusion, this paper will refer to the increase in CV due to correlation as the *correlation delta-CV* or $\Delta CV_{corr}$.

The correlation effect is quite cumbersome and very difficult to calculate mentally. The double sum contains many terms, and the correlation matrix is not, generally speaking, simple. (Sometimes, the correlation matrix is not even readily available!) And, there are no easy rules-of-thumb to help approximate the correlation term.

Fortunately, it is not necessary to find the value of the correlation term, because the true coefficient of variation is known *a priori*. What is really of interest is by how much the zero-correlation CV is changed (by the presence of correlation) to reach the true CV. This is the correlation delta-CV and is, by inspection, equal to the difference between the true CV and $CV_{zero\ corr}$:

$$\Delta CV_{corr} = CV_X - CV_{zero\ corr} \qquad \text{(Eq. 12)}$$

There is an interesting result for the perfect correlation case (i.e. every entry in the correlation matrix is equal to one) that is made clear by the vector form of the pCV representation of $CV_X$ shown in (Eq. 7). If all entries in the matrix operator $\widehat{R}$ are set equal to 1, and the right-hand-side of the equation is multiplied out explicitly, it turns out that the resultant $CV_X$ is equal to the sum of the pCV vector entries, which is, of course, the weighted average CV! In other words, the perfect correlation case is that one special case for which $CV_{w.ave}$ is equal to the true coefficient of variation.

Of course, the perfect correlation case is just as theoretical as the zero correlation case. Still, it is useful to note, because it puts an upper bound on the additional CV that can be gained for the parent element by increasing the correlation between the next-level child elements alone. In other words, $CV_{w.ave}$ is the highest CV that can be obtained, given a constant set of pCV.

Again, a word of caution is warranted here. Typically, correlation is applied to the lowest level inputs of a model, not directly to the Level 2 WBS elements. Though the addition of correlation to independent input distributions should not change their spreads, any higher level distributions that depend on those inputs will be altered. So, if correlation is applied or changed at levels below the Level 2 child elements, the set of pCV will, of course, *not* be constant, and $CV_{w.ave}$ (and the new upper limit for the CV) will be something of a moving target. Nevertheless, for diagnostic purposes, the first-order approximation of holding the pCV constant is useful.

## A MENTAL PATH AND GRAPHICAL REPRESENTATION

Up to this point, the discussions on the pCV representation have blazed a trail toward understanding the true CV of a parent WBS element, starting with the intuitive (but faulty) $CV_{w.ave}$ and working through the two different parts of the pCV representation, which expresses the true CV as a function of the partial CV (the addends of $CV_{w.ave}$). The resulting mental path is powerfully simple, consisting of three basic steps, each of which can be done fairly easily inside one's head: (1) start with $CV_{w.ave}$, (2) multiply by *f* to arrive at $CV_{zero\ corr}$, and (3) add the effects of correlation.

To illustrate how this path might be followed in a briefing setting, without the aid of a computer or calculator, consider a hypothetical sample program with a total point estimate of one billion dollars, spread across seven child WBS elements, as shown in Figure 2.

| **Costs in BY2011 $M** | | | |
|---|---|---|---|
| WBS | Point Estimate | Mean | CV |
| | | | |
| **SAMPLE PROGRAM** | **$ 1,000.0** | **$ 1,072.2** | **0.1242** |
| Prime Mission Product | $ 500.0 | $ 546.2 | 0.1977 |
| System Engineering | $ 150.0 | $ 150.8 | 0.1009 |
| Program Management | $ 150.0 | $ 150.8 | 0.1009 |
| System Test & Evaluation | $ 100.0 | $ 120.4 | 0.2161 |
| Training | $ 9.0 | $ 7.2 | 0.3270 |
| Data | $ 16.0 | $ 20.0 | 0.3618 |
| Support Equipment | $ 75.0 | $ 76.9 | 0.3489 |

Figure 2. Statistics for the sample program

When the senior analyst is briefed on this estimate, he expresses concern over a CV of only 0.12, stating that the size, complexity, and risk involved would suggest a much wider distribution. So, he asks the question, "Why is your CV so low? What can we do to fix it?" To answer the question, follow the steps outline above.

*Step 1: Start with $CV_{w.ave}$.*

The "big ticket" item in this sample program is clearly the Prime Mission Product (PMP), representing a little over half the total mean of the estimate. With a CV of about 0.2, the PMP partial CV is roughly 0.10. Other contributors include System Engineering (SE) and Program Management (PM), together comprising a bit less than 30% of the total mean, and with a CV of around 0.1 each (resulting in a total pCV of about 0.03 for the two elements). The elements that make up the remaining 20% of the estimate have varying CVs, but for a rough mental exercise, they can be treated as having an average CV of about 0.3 each, leading to a total pCV (for the remaining elements) of 0.06. Adding the estimated pCV yields an estimate for $CV_{w.ave}$ of 0.10 + 0.03 + 0.06 = 0.19, as might be expected based on the influence of PMP on the weighted average. (These mental calculations are summarized in Figure 3.)

| Costs in BY2011 $M | | | | | |
|---|---|---|---|---|---|
| WBS | Point Estimate | Mean | CV | Approx. Weight (p) | Approx. pCV |
| | | | | p = Mean/Total Mean | pCV = p*CV |
| **SAMPLE PROGRAM** | **$ 1,000.0** | **$1,072.2** | **0.1242** | **100%** | **N/A** |
| Prime Mission Product | $ 500.0 | $ 546.2 | 0.1977 | 50% | 0.10 |
| System Engineering | $ 150.0 | $ 150.8 | 0.1009 | 30% | 0.03 |
| Program Management | $ 150.0 | $ 150.8 | 0.1009 | | |
| System Test & Evaluation | $ 100.0 | $ 120.4 | 0.2161 | 20% | 0.06 |
| Training | $ 9.0 | $ 7.2 | 0.3270 | | |
| Data | $ 16.0 | $ 20.0 | 0.3618 | | |
| Support Equipment | $ 75.0 | $ 76.9 | 0.3489 | | |
| | | | | Σ(pCV) = | 0.19 |

Figure 3. Summary of mental calculations to arrive at $CV_{w.ave}$

*Step 2: Multiply by f to arrive at $CV_{zero\ corr}$:*

The weighted average CV is, of course, just a starting point. Each of the child-level CVs must be effectively adjusted by *f* to arrive at the zero-correlation CV, the next stop on the path. The Prime Mission Product pCV accounts for a little more than half of $CV_{w.ave}$, so the leading $q_1$ is about 0.5—a fairly medium range. Seven is a fairly average number of child elements as well, so it is justified to approximate *f* by ⅔ (as shown in Table 1) to go from $CV_{w.ave}$ to $CV_{zero\ corr}$. To make the mental math easier, treat $CV_{w.ave}$ as close to 0.18, two-thirds of which is 0.12.

*Step 3: Add the effects of correlation to reach the true CV:*

To find the effects of correlation, simply take the difference between the true CV (which includes correlation) and $CV_{zero\ corr}$ (which does not). In this case, the true CV is very close to the estimated $CV_{zero\ corr}$—certainly within the margin of error. So, a reasonable conclusion is that

very little correlation exists between the child WBS elements. This can be verified by examining the correlation matrix, shown in Figure 4.



| | PMP | SE | PM | STE | Training | Data | Spt Eq |
|---|---|---|---|---|---|---|---|
| PMP | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.6 |
| SE | | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| PM | | | 1 | 0.0 | 0.0 | 0.0 | 0.0 |
| STE | | | | 1 | 0.0 | 0.1 | 0.0 |
| Training | | | | | 1 | 0.0 | 0.0 |
| Data | | | | | | 1 | 0.3 |
| Spt Eq | | | | | | | 1 |

Figure 4. Correlation matrix for the sample program.

Having followed the mental path from $CV_{w.ave}$ to the true CV, the findings can be summarized by saying that the child elements have a (weighted) average CV of about 0.19, but that because of the number of child elements, that value must be reduced by a third to about 0.12, which is quite close to the true parent CV, indicating that insufficient correlation has been applied to this estimate. If additional correlation is applied, the true CV may be able to go as high as 0.19 (likely not even that high, because perfect correlation is absurd), but to go any higher, the distributions on the inputs may need to be adjusted.

The rough approximations made above can be validated after the fact by working through the math on a spreadsheet, as shown in Figure 5. Notice that the estimate for $CV_{w.ave}$ (or the sum of

| **Costs in BY2011 $M** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| WBS | Point Estimate | Mean | CV | Weight (p) | pCV | Sq. pCV | Adj. CV | Adj. pCV |
| | | | | p = Mean/Total Mean | pCV = p*CV | pCV^2 | Adj CV = f*CV | Adj pCV = f*pCV |
| **SAMPLE PROGRAM** | **$ 1,000.0** | **$1,072.2** | **0.1242** | **100.0%** | **N/A** | **N/A** | **N/A** | **N/A** |
| Prime Mission Product | $ 500.0 | $ 546.2 | 0.1977 | 50.9% | 0.1007 | 1.01E-02 | 0.1147 | 0.0584 |
| System Engineering | $ 150.0 | $ 150.8 | 0.1009 | 14.1% | 0.0142 | 2.01E-04 | 0.0585 | 0.0082 |
| Program Management | $ 150.0 | $ 150.8 | 0.1009 | 14.1% | 0.0142 | 2.01E-04 | 0.0585 | 0.0082 |
| System Test & Evaluation | $ 100.0 | $ 120.4 | 0.2161 | 11.2% | 0.0243 | 5.89E-04 | 0.1254 | 0.0141 |
| Training | $ 9.0 | $ 7.2 | 0.3270 | 0.7% | 0.0022 | 4.77E-06 | 0.1897 | 0.0013 |
| Data | $ 16.0 | $ 20.0 | 0.3618 | 1.9% | 0.0068 | 4.57E-05 | 0.2099 | 0.0039 |
| Support Equipment | $ 75.0 | $ 76.9 | 0.3489 | 7.2% | 0.0250 | 6.26E-04 | 0.2024 | 0.0145 |

| | | |
|---|---|---|
| B | Σ(pCV) = | 0.1873 |
| A | RSS (pCV) = | 0.1087 |
| | f = (A/B) = | 0.58 |

| | |
|---|---|
| Σ(Adj. pCV) = | 0.1087 |
| Adj. $CV_{w.ave}$ = | 0.1087 |

| | |
|---|---|
| $CV_{w.ave}$ = | 0.1873 |
| $CV_{zero\ corr}$ = | 0.1087 |
| True CV = | 0.1242 |
| Corr. Effect = | 0.0155 |

Figure 5. Working through the above analysis with spreadsheet software can validate the mental approximations.

the pCV) was fairly close, but the zero-correlation CV was off by about 0.01. This is due to the rough approximation for *f*, which was actually closer to three-fifths than two-thirds. (Recall that the rules-of-thumb have a tendency to overstate *f*.) However, the conclusion that insufficient correlation was applied is the same, which should answer the immediate question of why the CV was so low and what can be done to bring it up.



If the full analysis is done ahead of time, these results can be displayed graphically, as shown in Figure 6. The parent CV is shown by the large bar in the background, divided into the zero-correlation CV and the correlation effect, or (more accurately) the correlation delta-CV. The smaller bars in the foreground represent the CV, before and after adjustment, of each child element, and the yellow stripe on each represents the relative weight *p*. Finally, a dotted line in drawn for $CV_{w.ave}$ to give an indication of how much additional CV can be gained by adjusting the correlation between the child WBS elements.

This analysis can be repeated for lower level elements, especially influential ones (such as PMP) that may warrant their own CV analysis. If the calculations are performed dynamically, and the graph is linked to the cells in which the calculations are performed, the analysis can be done automatically by simply pasting the relevant statistics (see Figure 2) into the appropriate cells, overriding the previous values.

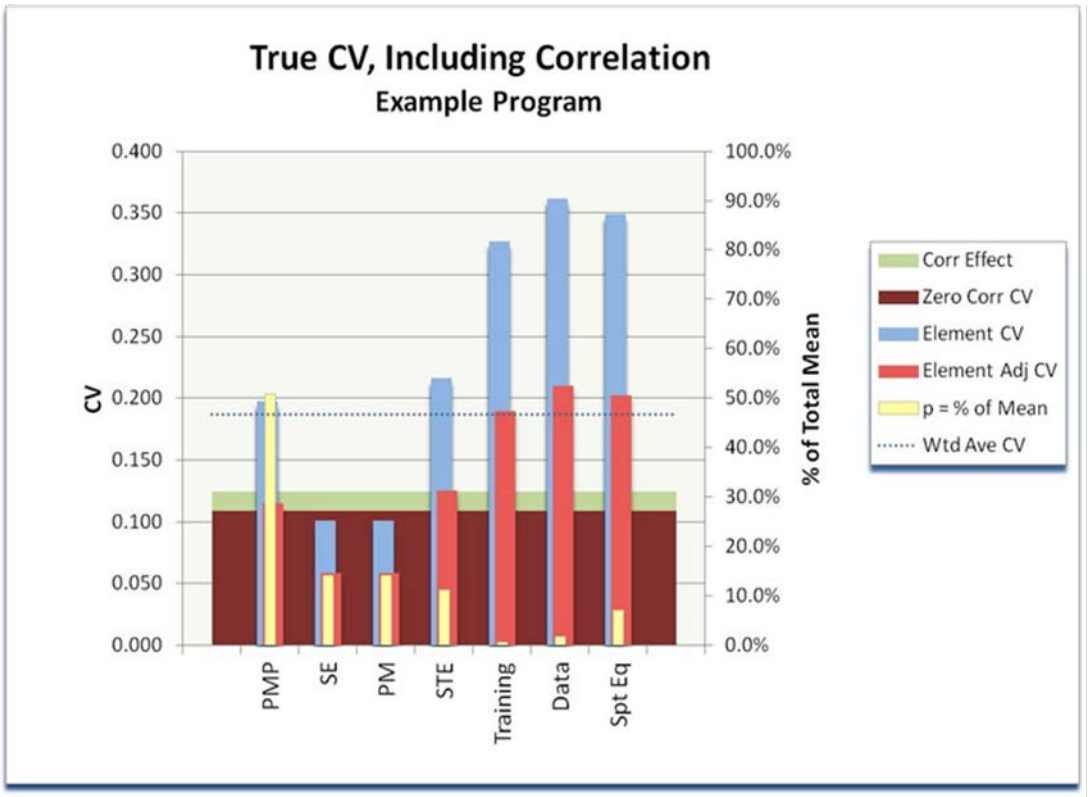


Figure 6. Graphical representation of the above CV analysis

## CONCLUSIONS

The coefficient of variation is a basic, fundamental statistical measure of spread, but what drives the CV is not always clearly understood. This paper has presented an alternative approach to diagnosing the top-level CV of an estimate by expressing it in terms of the next-level children. This approach can be drilled down to the lowest level for which the child element is a simple sum of still lower level children. In addition, the analysis is reasonable for a quick, mental diagnosis in a briefing situation, but also robust enough to lend significant insight if done thoroughly and can be automated for maximum efficiency.



## REFERENCES

Air Force Cost Analysis Agency (AFCAA). 2007. Air Force Cost Risk and Uncertainty Handbook.

Salas, Hille, and Etgen. 1999. *Calculus: One and Several Variables*, 8th ed. New York, NY: John Wiley & Sons, Inc.

Smith, A. 2011. Effective use of cost reports. Proceedings of The International Society of Parametric Analysts and The Society of Cost Estimating and Analysis 2011 Joint Annual Conference & Training Workshop.

## APPENDIX: DERIVATION OF (EQ. 4)



The variance $\sigma^2$ of a parent Work Breakdown Structure (WBS) element can be written as a sum of each pair of child element standard deviations multiplied by each other and by the correlation between the two elements. This formula is the basis for expressing the coefficient of variation (CV) of the parent in terms of its children. The following is a derivation of that expression for $\sigma^2$.

Let $X$ be the parent of $n$ children, each denoted by $x_i$, in a typical WBS. In general, the value of the parent is equal to the sum of the children, so that

$$X = \sum_{j=1}^{n} x_j = x + y + z + \cdots \qquad \text{(Eq. A1)}$$

where on the right-hand side of (Eq. A1), each child element is represented by a unique variable (which will aid notational simplicity in the following analysis).

To account for uncertainty in the estimate, each quantity can be represented as a probability distribution about some point estimate (PE). While in theory these distributions are thought of as continuous, consisting of an infinite set of possible values and corresponding probabilities, in practice they are almost always dealt with as large sets of discrete simulated data. In such a case, the subscript $i$ can be used to denote particular data points within that set. By this notation $x_i$, $y_i$, represent the values for the children elements $x$ and $y$ in the $i^{th}$ iteration of the simulation, and by extension,

$$X_i = \sum_{j=1}^{n} x_{j,i} = x_i + y_i + z_i + \cdots \qquad \text{(Eq. A2)}$$

Note that in (Eq. A2), $x_{j,i}$ is the $i^{th}$ iteration of the variable $x_j$. For notational simplicity, this paper adopts the convention on the right-hand side of (Eq. A1) and (Eq. A2) and refers to each child element by a unique (but general) variable ($x$, $y$, or $z$), unless otherwise specified.

Finally, recall the definitions of mean ($\bar{x}$), variance ($\sigma^2$), and coefficient of variation (CV) from basic statistics, as shown below:

$$\bar{x} = \frac{\Sigma_i^N x_i}{N} \qquad \text{(Eq. A3)}$$

$$\sigma^2 = \frac{\sum_i^N (x_i - \bar{x})^2}{N-1} \qquad \text{(Eq. A4)}$$



$$CV = \frac{\sigma}{\bar{x}} \qquad \text{(Eq. A5)}$$

where $N >> 1$ is the number of iterations used for the Monte Carlo simulation. (Typically, $N$ is on the order of several thousand.)

Now, begin with the definition of CV from (Eq. A5), as applied to $X$, and square both sides to avoid working with radicals.

$$(CV_X)^2 = \left(\frac{\sigma_X}{\bar{X}}\right)^2 = \frac{1}{\bar{X}^2}\left[\frac{\sum_i^N (X_i - \bar{X})^2}{N-1}\right] \qquad \text{(Eq. A6)}$$

It is well known and easily proved that the mean of a sum is equal to the sum of the means of the individual terms. Making note of this fact and leveraging (Eq. A1), the factor in square brackets above becomes

$$\left[\frac{\sum_i^N (X_i - \bar{X})^2}{N-1}\right] = \left[\frac{\sum_i^N \left([\![x]\!]_i + y_i + z_i) + \cdots - (\bar{x} + \bar{y} + \bar{z} + \cdots)\right)^2}{N-1}\right]$$

$$= \left[\frac{\sum_i^N \left([\![x]\!]_i - \bar{x}) + (y_i - \bar{y}) + [\![z]\!]_i - \bar{z}) + \cdots\right)^2}{N-1}\right]$$

$$= \left[\frac{\sum_i^N ([\![x]\!]_i - \bar{x})^2 + (y_i - \bar{y})^2 + [\![z]\!]_i - \bar{z})^2 + \cdots + \sum cross\ terms)}{N-1}\right] \qquad \text{(Eq. A7}$$

Comparison with (Eq. A4) shows that this is simply the sum of the variances of $x$, $y$, and $z$, plus cross terms. Thus, (Eq. A7) can be re-written:

$$(CV_X)^2 = \left(\frac{\sigma_X}{\bar{X}}\right)^2 = \left[\frac{\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + \cdots + \sum cross\ terms}{\bar{X}^2}\right] \qquad \text{(Eq. A8)}$$

It is expected that cross terms involving two variables (say, $x$ and $y$) should be related to the correlation between the two, as defined below:

$$r_{xy} = \frac{N\sum_i^N (x_i y_i) - \left(\sum_i^N x_i\right)\left(\sum_i^N y_i\right)}{\sqrt{\left[N\sum_i^N x_i^2 - \left(\sum_i^n x_i\right)^2\right]\left[N\sum_i^N y_i^2 - \left(\sum_i^n y_i\right)^2\right]}}$$

$$= \frac{N\sum_i^N (x_i y_i) - (\sum_i^N x_i)(\sum_i^N y_i)}{D(x,y)} \qquad \text{(Eq. A9)}$$

and indeed this is the case, as will be shown shortly. First, however, it is helpful to examine a single, generalized term. Going through the algebra, it is easy to show that each cross term in (Eq. A7) is of the form

$$2\frac{\sum_i^N [(x_i - \bar{x})(y_i - \bar{y})]}{N-1} \qquad \text{(Eq. A10)}$$

where *x* and *y* are generic variables, and either may be substituted with *z* or any other general variable. By multiplying the numerator out and taking advantage of the assumption (asserted earlier) that *N* is much greater than 1, this generalized term becomes (temporarily ignoring the factor of 2):

$$\frac{\sum_i^N [(x_i - \bar{x})(y_i - \bar{y})]}{N-1} = \frac{\sum_i^N (x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{N-1}$$

$$= \frac{\sum_i^N x_i y_i - \bar{y}\sum_i^N x_i - \bar{x}\sum_i^N y_i + N\bar{x}\bar{y}}{N-1}$$

$$\approx \frac{\sum_i^N x_i y_i}{N-1} - \frac{\bar{y}\left(\sum_i^N x_i\right)}{N} - \frac{\bar{x}\left(\sum_i^N y_i\right)}{N} + \bar{x}\bar{y}$$

$$= \frac{\sum_i^N x_i y_i}{N} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{\sum_i^N x_i y_i}{N-1} - \bar{y}\bar{x} \qquad \text{(Eq. A11)}$$

From the definition of the mean in (Eq. A3), (Eq. A11) becomes

$$\frac{\Sigma_i^N x_i y_i}{N-1} - \overline{yx} = \frac{\Sigma_i^N x_i y_i}{N-1} - \frac{\Sigma_i^N x_i}{N}\frac{\Sigma_i^N y_i}{N}$$



$$= \frac{\Sigma_i^N x_i y_i}{N-1} - \frac{\Sigma_i^N x_i \Sigma_i^N y_i}{N^2}$$

$$= \frac{N\Sigma_i^N x_i y_i - \frac{1}{N}(N-1)\Sigma_i^N x_i \Sigma_i^N y_i}{N(N-1)}$$

$$\approx \frac{N\Sigma_i^N (x_i y_i) - \left(\Sigma_i^N [x_i]\left(\Sigma_i^N y_i\right)\right]}{N(N-1)} \qquad \text{(Eq. A12)}$$

where we again make use of the assumption that $N >> 1$ (and consequently, $\frac{1}{N}(N-1) \approx 1$ ).

Note the form of the numerator and compare it with the definition of correlation in (Eq. A9). This result allows the cross term (Eq. A10) to be written as

$$2\frac{\Sigma_i^N[(x_i-\bar{x})(y_i-\bar{y})]}{N-1} \approx 2\left[\frac{\Sigma_i^N x_i y_i}{N-1} - \overline{yx}\right]$$

$$\approx 2\frac{r_{x,y}D(x,y)}{N(N-1)} \qquad \text{(Eq. A13)}$$

where $D(x,y) = \sqrt{\left[N\sum_i^N x_i^2 - \left(\sum_i^n x_i\right)^2\right]\left[N\sum_i^N y_i^2 - \left(\sum_i^n y_i\right)^2\right]}$ is the denominator in the expression for the correlation between variables *x* and *y* shown in (Eq. A9). Although cumbersome in this form, a little algebra demonstrates that *D*(*x*, *y*) can be written in terms of the standard deviations of the *x* and *y* distributions.

Start with the definition of the variance of *x*:

$$\sigma_x^2 = \frac{\Sigma_i^N (x_i - \bar{x})^2}{N-1}$$

$$= \frac{\Sigma_i^N (x_i^2 + \bar{x}^2 - 2x_i\bar{x})}{N-1}$$

$$= \frac{\Sigma_i^N [x]_i^2) + \Sigma_i^N [(\bar{x}]^2) - 2\Sigma_i^N (x_i\bar{x})}{N-1}$$

$$= \frac{\sum_i^N [\![(x]\!]_i^2) + N\bar{x}^2 - 2\bar{x}\sum_i^N [\![(x_i]\!])}{N-1} \quad \text{(Eq. A14)}$$



Note that from the definition of the mean, $\sum_i^N [\![(x_i]\!]) = N\bar{x}$, which, when plugged into (Eq. A14) yields

$$\sigma_x^2 = \frac{\sum_i^N [\![(x]\!]_i^2) + N\bar{x}^2 - 2N\bar{x}^2}{N-1} = \frac{\sum_i^N [\![(x]\!]_i^2) - N\bar{x}^2}{N-1} \quad \text{(Eq. A15)}$$

and consequently,

$$\sum_i^N [\![(x]\!]_i^2) = (N-1)\sigma_x^2 + N\bar{x}^2 \quad \text{(Eq. A16)}$$

Plugging the result of (Eq. A16), as well as the identical result for *y*, into $D(x, y)$ yields

$$D(x,y) = \sqrt{\left[N\left((N-1)\sigma_x^2 + N\bar{x}^2\right) - \left(\sum_i^n x_i\right)^2\right]\left[N\left((N-1)\sigma_y^2 + N\bar{y}^2\right) - \left(\sum_i^n y_i\right)^2\right]}$$

$$= \sqrt{[N(N-1)\sigma_x^2 + N^2\bar{x}^2 - (N\bar{x})^2][N(N-1)\sigma_y^2 + N^2\bar{y}^2 - (N\bar{y})^2]}$$

$$= \sqrt{[N(N-1)\sigma_x^2][N(N-1)\sigma_y^2]} = \sqrt{N^2(N-1)^2\sigma_x^2\sigma_y^2}$$

$$= N(N-1)\sigma_x\sigma_y \quad \text{(Eq. A17)}$$

Inserting this result into (Eq. A13), the cross terms can be written as

$$2\frac{r_{x,y}D(x,y)}{N(N-1)} = 2r_{x,y}\sigma_x\sigma_y \quad \text{(Eq. A18)}$$

Finally, taking this result all the way back to (Eq. A7), the variance of the parent element *X* can be written as the sum of each child's variance, plus a cross term of the form in (Eq. A18) for each variable pair, as follows:

$$\sigma_X^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 + \cdots + 2r_{x,y}\sigma_x\sigma_y + 2r_{x,z}\sigma_x\sigma_z + 2r_{y,z}\sigma_y\sigma_z + \cdots \quad \text{(Eq. A19)}$$

Recognizing that, by definition, the correlation of any variable to itself is unity, this can be further simplified as



$$\sigma_X^2 = \sum_{j=1}^{n}\sum_{k=1}^{n} r_{j,k}\sigma_j\,\sigma_k \quad \text{(Eq. A20)}$$

where here subscripts are used to denote independent variables, not simulation iterations. (Logically, $r_{j,k}$ is the correlation between variables $x_j$ and $x_k$, with similar notation for the standard deviations of those variables.)