PRT-207

IR&D 831.106/831.099

The Zone System of Uncertainty Analysis

Presented at the
International Cost Estimating and Analysis Association
(ICEAA)
Professional and Training Workshop
Atlanta Georgia
7-10 June 2016

Jeff A. McDowell, CCEA
jmcdowell@tecolote.com

ABSTRACT

This paper presents a methodology for selecting cost uncertainty distributions and their bounds via a sliding complexity scale. The methodology is inspired by the Zone System for photography and is characterized by five gradations of uncertainty. The distribution shapes and their bounds are drawn from interesting macro patterns present in the AFCAA Cost Risk and Uncertainly Analysis Metrics Manual (CRUAMM) body of work.

A ZONE METHOD FOR PHOTOGRAPHY

Utilizing gradations of uncertainty for cost analysis parallels the use of light gradations in photography. Pioneering photographers such Ansel Adams developed this technique to attain predicable and repeatable results in image capture and printing. Figure 1 presents key concepts in the Zone System which centers on evaluating a scene in terms of five shades of gray and then setting exposure by sliding camera settings darker or lighter. This is exemplified in the top of Figure 2 by evaluating five gray tones in a scene. This is the photographer's judgment performed by "stepping back' from the detail and making an overall assessment of amount of light in the scene. The bottom of the figure illustrates how the camera settings may be used to affect the final image captured.

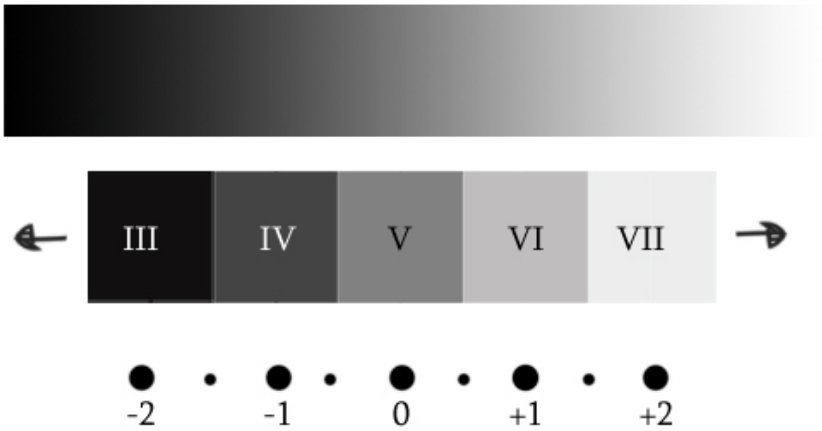
The Zone System's Key Concepts

"The zone system divides a scene into 10 zones on the tonal scale (though there are variations of 9 and 11 zones). Every tonal range is assigned a zone. Every zone differs from the one before it by 1 stop, and from the one following it by 1 stop. So every zone change equals 1 stop difference. Zones are identified by roman numbers, with the middle tone (with 18% reflectance) being a zone V which is zone 5.

For us digital photographers, we are only concerned with zones III through VII (zones 3 through 7). The darkest part of a scene would fall into zone III, while the brightest part of a scene would fall into zone VII. Anything darker than zone III would render as pure black with no detail (under-exposed), while anything brighter than zone VII would render as pure white with no detail (over-exposed).

If you point your camera at an area with average reflectance and obtain the correct meter readings (a zero on the light meter), that area would be rendered as average. If you open up your lens or slow down your shutter speed by one stop, that area will become over-exposed by one stop. If you close down your lens or increase your shutter speed by one stop, that area will become under-exposed by one stop.

Now, we've agreed that an average tone is naturally placed into zone V. If you over-expose it by one stop, you'll be placing it in zone VI (zone 6), causing it to render brighter than it actually is. If you under-expose it by one stop, you'll be placing it in zone IV (zone 4) causing it to render darker than it actually is."



Source: Understanding & Using Ansel Adam's Zone System, Diana Eftaiha, 20 Mar 2013

Figure 1: Photography Zone System

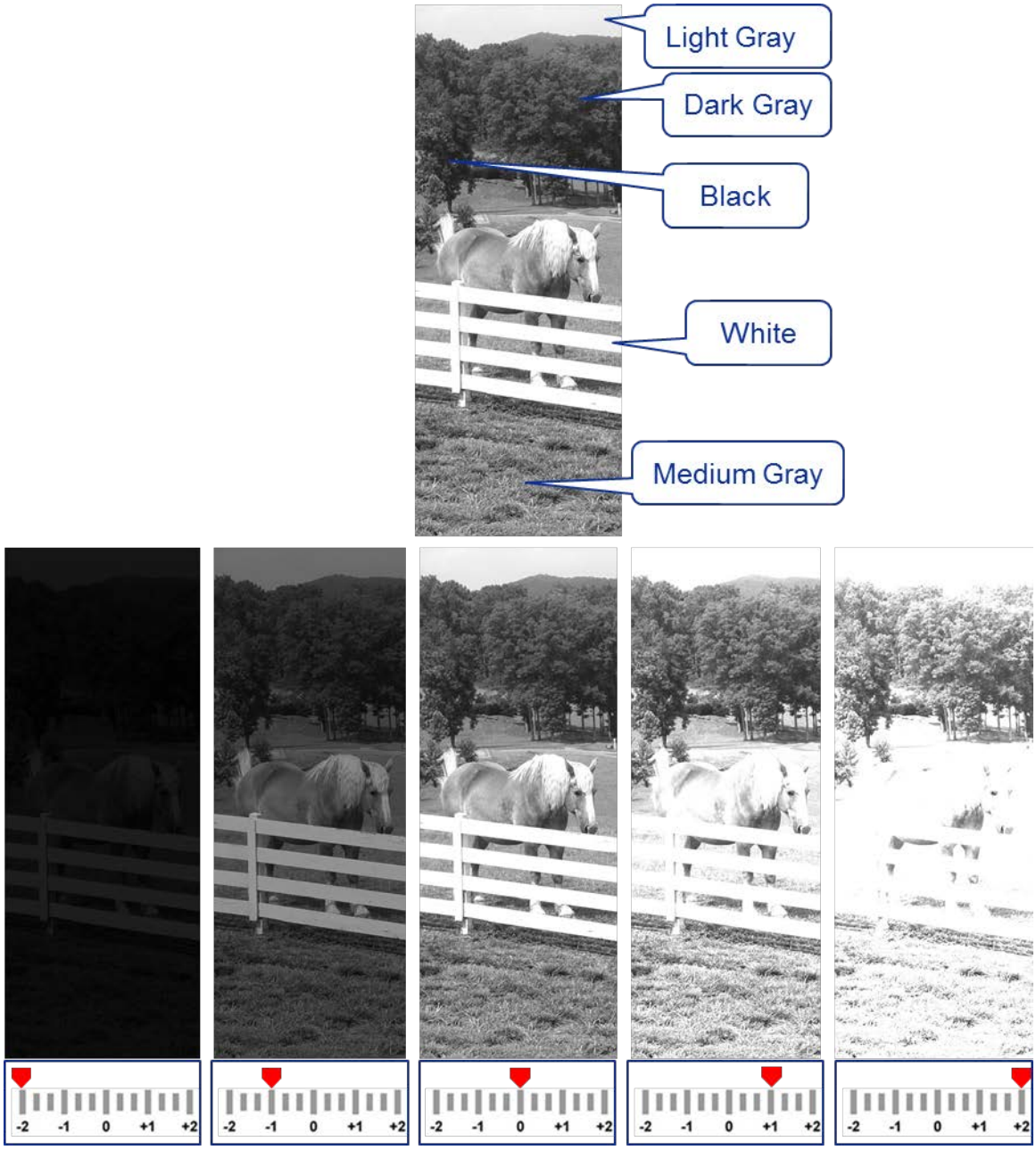


Figure 2: Photography Zone System Example

A METHODOLOGY FRAMEWORK

The preceding discussed utilizing five gradations of light in a system for photography. The topic of this paper is utilizing five gradations of uncertainty for use in cost uncertainty modeling. An analyst familiar with a program and its life-cycle cost model could "step back" from their detailed knowledge of the program and the cost model's inner workings and consider from a bird's eye view the relative uncertainty of each model element as depicted in Figure 3 and summarized in steps:

1. A few elements could be deemed much more uncertain than the other elements.
2. Likewise relative judgment made on a few elements could deem them much more certain than the other elements.
3. An even greater number of elements could be deemed as having medium uncertainty.
4. Inevitably some elements would be deemed to have a degree-of-uncertainty that is not medium nor at either extreme: medium-high or low-medium.

Figure 3: Relative Uncertainty in a Cost Model

Thus five gradations are possible for each cost element as depicted in Figure 4 as five shades of gray. Conceptually these five, reading left-to-right, would have increasingly wide dispersions.

Low Uncertainty	Low-Medium Uncertainty	Medium Uncertainty	Medium-High Uncertainty	High Uncertainty

Figure 4: Five Gradations of Uncertainty

The dispersion metric used throughout this paper is the Coefficient of Variation (CV). CV is a normalized relative measure of overall dispersion expressed as standard deviation divided by the mean. A low CV is associated with a narrow distribution (low uncertainty) and a high CV is associated with a wide distribution (high uncertainty) as depicted in Figure 5.

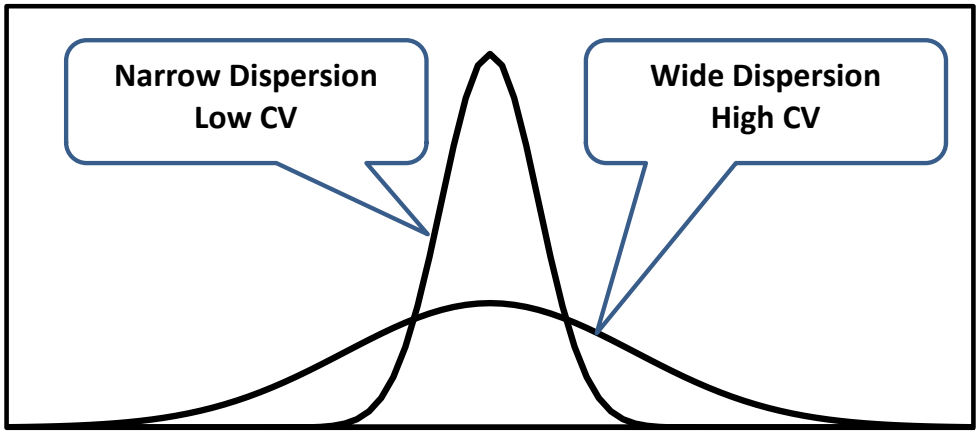


Figure 5: Low and High CVs

For the sake of illustrating the framework shown in Figure 4 further, let's assign a CV of 0.5 to medium uncertainty and increment or decrement the adjacent gradation's CV with 0.1 so the five shades of gray have CVs ranging from 0.3 to 0.7. With this construct, shown in Figure 6, the analyst could choose their preferred distribution, determine the parameters for each of the five CVs and readily assign them to each model element based on their bird's eye assessment, and rapidly assemble a functioning Monte Carlo simulation.

	Low Uncertainty	Low-Medium Uncertainty	Medium Uncertainty	Medium-High Uncertainty	High Uncertainty
Dispersion (Notional CV)	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**

Figure 6: Five Gradations with Notional CVs

Now consider these values as representative of a typical, ordinary program and how these values may differ depending on program complexity. For example, the analyst applying this technique to a simple, less complex program that:

- Is very routine,
- has a higher technology readiness level (TRL),
- is not pushing the state-of-the-art,
- has a narrow mission,
- has a narrow user base, and
- is executing a simple acquisition strategy.

The analyst could again readily identify five gradations of uncertainty in their model. However, given the simplicity of the program, the analyst may judge that the dispersion of medium uncertainty in this simple program is equivalent to the dispersion of low uncertainty on the typical program (CV=0.3). And the most uncertain elements of the simple program are no more uncertain than the medium ones in the typical program (CV=0.5).

Alternatively, consider the analyst applying the same techniques to a complex challenging program. This program:

- has low TRLs,
- is utilizing exotic materials,
- is pursuing an ambitious schedule,
- has a nebulous mission,
- has a multi-Service user base, and
- is engaged in a complex acquisition approach.

Again five gradations of uncertainty could be identified, but the medium uncertain items on this complex program may be judged to be as widely dispersed as the most highly dispersed items in the typical program (CV=0.7).

To illustrate this expanded framework to accommodate varying program complexities, visualize the five core gradations shifting to the right or to the left as depicted in Figure 7.

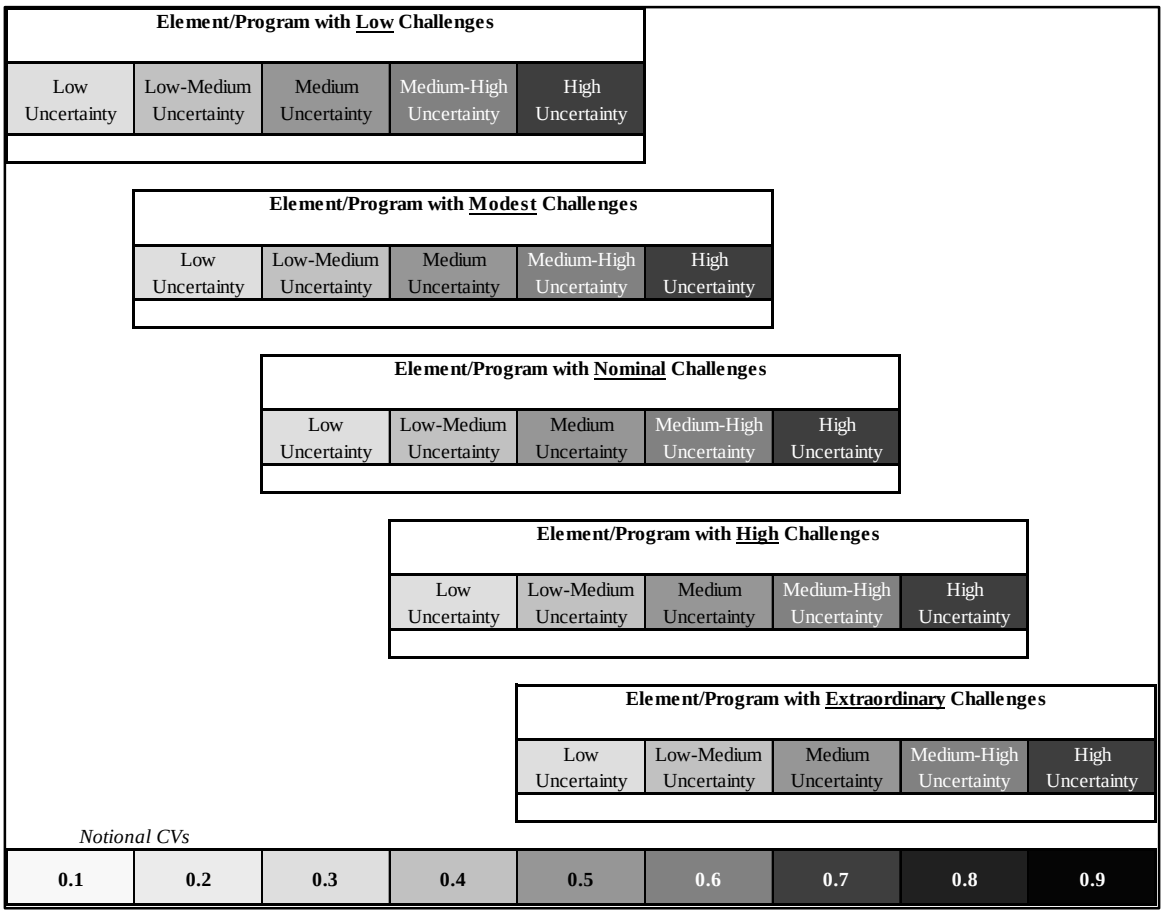


Figure 7: Offsetting Gradations for Program Complexity Yields Nine Gradations

With five gradations of uncertainty shifted across five gradations of program challenges, our framework consists of a total of nine gradations. If we continue our assignment of notional CV values anchored with 0.5 in the center and separating each increment with 0.1, our resulting range of notional CVs would be 0.1 – 0.9 as shown at the bottom of Figure 7.

To summarize our framework thus far: We have reasoned ourselves into a framework of nine bins for which the using analyst would make three judgment calls and employ a set of default values. To apply this to an estimate, the user would follow four steps.

1. First, determine (via discernment) the degree of program challenge.
2. Second, determine (via informed opinion) the relative uncertainty of each cost model element.
3. Third, select a distribution for each of the model's uncertain elements.
4. And finally, use the set of default CVs to determine the parameters for each distribution.

While the prior section may be insightful and interesting, it is not in the form of a readily usable methodology. This four-step process is straightforward and orderly, but thus far we have offered only default CVs fabricated for illustration with no supporting data or analysis. As such, each of the four steps need additional rigor to advance this framework into a viable methodology. Fortunately the Air Force Cost Analysis Agency (AFCAA) Cost Risk and Uncertainly Analysis Metrics Manual (CRUAMM)[1][2][3][4] body of analysis can advance the last two steps by anchoring them in data. The remainder of this paper exploits the macro-patterns that emerged from the CRUAMM descriptive analysis for distribution shape selection and CV selection.

[1] AFCAA Cost Risk and Uncertainly Analysis Metrics Manual, ICEAA Workshop Presentation, Wilson Rosa AFCAA, Alfred Smith Tecolote Research, Dr Lew Fichter Tecolote Research, Jeff McDowell Tecolote Research, 08 June 2010.
[2] Real Data, Real Uncertainty, 45th Annual Department of Defense Cost Analysis Symposium on Cost Analysis and the Downturn February 14-17, 2012, Alfred Smith, Jeff McDowell, Dr. Lew Fichter Tecolote Research.
[3] Real Data, Real Uncertainty, Dr. Wilson Rosa AFCAA, Alfred Smith Tecolote Research, Jeff McDowell Tecolote Research, Dr. Lew Fichter Tecolote Research, 27 June 2012.
[4] CR-1501/1 Cost Risk and Uncertainty Analysis Metrics Manual (CRUAMM), Alf Smith, Jeff McDowell, Dr. Lew Fichter, Bryan Blevins, Nick DeTore. Tecolote Research Prepared for Air Force Cost Analysis Agency (AFCAA), November 2011.

CRUAMM BACKGROUND

The purpose of CRUAMM was to provide the analyst with a source for locating, by weapon system commodity, distributions to use in modeling low-level cost element uncertainty. The CRUAMM document was completed in November 2011 and consists of a main volume and nine appendices for a total of 1737 pages. This comprehensive effort involved fitting over 1400 separate distributions to Cost Estimating Relationship (CER) residuals and other data for use as cost uncertainty distributions. To convey the depth of the CRUAMM analysis, three excerpts of the CRUAMM final report are provided in three figures. Figure 8 summarizes a few key points from the main CRUAMM document as it relates to this paper. The process used in CRUAMM to develop the fitted distributions is summarized in Figure 9. The main volume consists of tables whose usage is illustrated in Figure 10. While successful as a wide-ranging catalog, its scale can be a daunting resource for a time-pressed analyst to use.

Though the specific activities of the CRUAMM task were micro-focused on the individual WBS elements of specific commodities, it became apparent as the task progressed that many interesting macro patterns were also present. This paper further explores those macro patterns for the purposes of populating the Zone System of Uncertainty Analysis with usable distributions based on data.

The remainder of the paper presents an examination and categorization of the CRUAMM results. First, the analysis process is described in which estimating methodology residuals were fit to distributions. Second, the principles of Orders of Dispersion are described. Third, cluster analysis is performed to divide the results into nine bins. Finally, it assembles the analysis into the Zone System for Uncertainty Analysis.

First, CRUAMM provides guidance on commonly used distribution shapes. Often an analyst may have a low and a high value for a given element but no guidance on choosing a distribution shape. For the analyst who can establish bounds but has no basis for selecting a shape, CRUAMM provides recommended shapes by commodity and cost element. By locating the item that most closely represents their element in the tables, an analyst can locate a recommended distribution shape.

Second, CRUAMM is a source of not only the distribution shapes but their parameters as well. When the analyst has no basis for determining shape or bounds, this document provides guidance on both for specific WBS elements and their cost drivers by commodity. By locating the element that most closely resembles their own, the analyst can obtain a recommend distribution shape and bounds to represent uncertainty.

Third, CRUAMM's distributions are all unitized. Unitized means the parameters have been normalized and are designed to be modeled as multipliers of point estimates. So, given a point estimate where:

$$\textbf{Cost Element}_{\textbf{Point Estimate}} = \textbf{Your Methodology}$$

its uncertainty can then be modeled as follows (causing the uncertainty to scale with the point estimate):

$$\textbf{Cost Element}_{\textbf{Uncertainty}} = \textbf{Your Methodology} * \textbf{Unitized Distribution}$$

Normalized cost data (recurring, non-recurring, T1, etc.), which was collected by commodity and by WBS element, were stratified consistent with typical cost estimate types (for instance new vs. modified, development vs. acquisition). Additionally, key technical parameters commonly used to develop cost estimating relationships were also collected by WBS element. Once collected, organized, normalized and stratified (new, modified, purpose, etc.), the data was then subjected to a distribution-fitting process. Results of this process yielded both descriptive statistics and the distribution fit results.

A distribution fitting utility was developed to fit a lognormal, normal, triangular and beta distribution to the selected data. The process is depicted in Figure 9.

Source: CR-1501/1 Cost Risk and Uncertainty Analysis Metrics Manual (CRUAMM), Alf Smith, Jeff McDowell, Dr. Lew Fichter, Bryan Blevins, Nick DeTore. Tecolote Research Prepared for Air Force Cost Analysis Agency (AFCAA), November 2011.

Figure 8: CRUAMM Key Points

ANALYTICAL PROCESS

Source: Real Data, Real Uncertainty, 45th Annual Department of Defense Cost Analysis Symposium on Cost Analysis and the Downturn February 14-17, 2012, Alfred Smith, Jeff McDowell, Dr. Lew Fichter Tecolote Research.

- **Sort sample data in ascending order**
- **Assign a cumulative percentile using the NIST[1] formula (different than Excel) but apply a "correction for continuity"[2]**

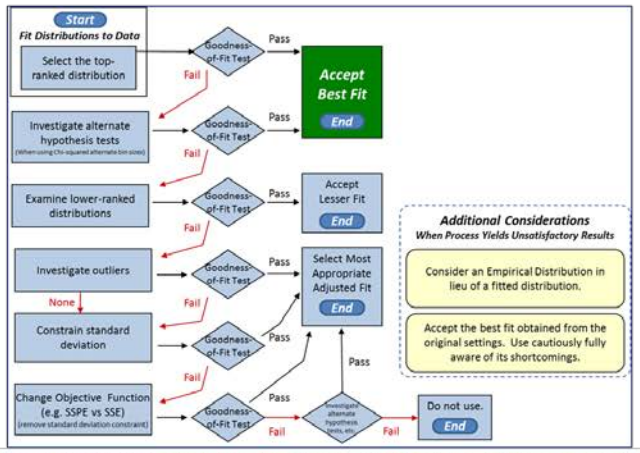
  (0.5*ObsFreq+NumObsBelow)/ObsCount
- **Use the sample descriptive statistics to provide a starting point for fit parameters**
- **Assess the difference between the sample and fit using either:**
  - Sum squared error
  - Sum squared percent error
    - n = number of data points
    - y = a sample data point
    - $\hat{y}$ = a fitted point

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSPE = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{\hat{y}_i} \right)^2$$

- **Use *Excel Solver to find the fit parameters that minimize SSE or SSPE**
  - Set optional constraints such as: Low>0, High>HighestSamplePoint
  - Select SSE or SSPE as error to be minimized
    - SSE is highly influenced by very large sample points (as compared to mean)
    - SSPE is highly influenced by fitted points close to zero (divide by zero)
- **Rank the fits using standard error of the estimate (SEE) or standard percent error (SPE)**
  - Where k = number of parameters in the fit
    - Normal, lognormal k = 2
    - Triangular k = 3
    - Beta k = 4
  - This is a preferred method to rank the fits (rather than SSE or SSPE directly) because it accounts for the degrees of freedom
- **Use a Goodness-of-Fit test (Chi^2) to determine significance of the fit**
  - Minimum SSE or SSPE alone does not necessarily mean the fit is meaningful
  - Only Chi^2 provides a p-value for any distribution

1. NIST= National Institute of Standards and Technology
2. From "Reliability and Information Functions for Percentile Ranks" Kim May and W. Alan Nicewander, Journal of Educational Measurement, Vol. 31, No. 4 (Winter, 1994), pp. 313-325



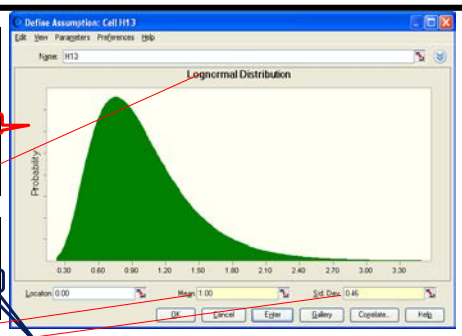
Process for evaluating distribution fits. Source: Joint Agency Cost Schedule Risk and Uncertainty Handbook, 12 March 2014.

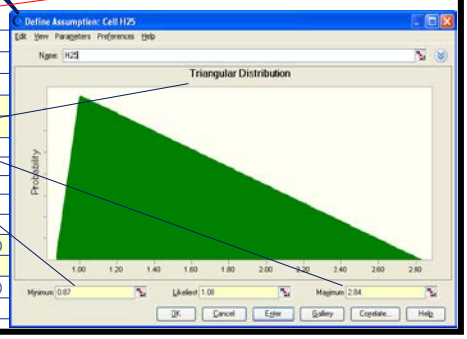
Figure 9: CRUAMM Analytical Process

WBS Description	Unique ID	CER	Point Estimate	Forecast (PE * Uncert)	Distribution	Uncertainty	Min	Max or Std Dev
Satellite			$298,647	$298,647				
Payload	RDPayload		$168,462	$168,462				
IR Sensor		356851 * PayloadDiam^0.562	$168,462	$168,462	Triangular	1	0.071	**1.9640**
Spacecraft Bus	RDSpBus		$15,096	$15,096				
Structure		157 * StrWgt^0.83	$2,787	$2,787	Triangular	1	0	1.9528
Thermal		394 * ThermWgt^0.635	$1,331	$1,331	Triangular	1	0	1.9528
Electrical Power System (EPS)		82.7 * EPSWgt	$2,865	$2,865	**Lognormal**	1		**0.4512**
Telemetry Tracking & Command (TT&C) and Command & Data Handling (C&DH)		545 * TTCWgt^0.761	$2,344	$2,344	Lognormal	1		0.2941
Attitude Determination & Control Sys		464 * ADCSWgt^0.867	$5,769	$5,769	Lognormal	1		0.4231
Int & Assy		989 + 0.215 * (RDPayload$ + RDSpBus)	$40,454	$40,454	Normal	1		**0.3544**
Program Level Costs		1.963 * (RDPayload$ + RDSpBus) ^0.841	$52,451	$52,451	Lognormal	1		**0.4609**
Ground Support Equip (GSE)		9.262 * (RDPayload$ + RDSpBus) ^0.642	$22,184	$22,184	Triangular	1	0	**3.1204**

Input Variable	Unique ID	CER	Point Estimate	Forecast		Mean		Std Dev
Payload Sensor diameter (m)	PayloadDiam		0.263	0.263	Lognormal	1		0.3000
Structure weight (kg)	StrWgt		32,000	32,000	Lognormal	1		0.5019
Thermal weight (kg)	ThermWgt		6,800	6,800	Triangular	1	0.261	2.7028
EPS weight (kg)	EPSWgt		45,700	45,700	**Triangular**	1	**0.8742**	**2.8357**
TT&C/DH weight (kg)	TTCWgt		6,800	6,800	Normal	1		0.4623
ADCS weight (kg)	ADCSWgt		18,300	18,300	Normal	1		0.3100
Spacecraft + Payload Weight (kg)	SpPyWgt		140,000	140,000	Lognormal	1		0.2307

Parameter	Mean	Median	Mode
T1 CER [1]	Lognormal (1.0000, 0.4512)	Lognormal (1.0971, 0.4950)	Lognormal (1.3204, 0.5958)
NR CER [1]			
T1 CER [3]	Triangular (0.1552, 0.7265, 2.1183)	Triangular (0.1635, 0.7651, 2.2308)	Triangular (0.2137, 1.0000, 2.9158)
NR CER [3]	Triangular (0.5206, 0.7584, 1.7210)	Triangular (0.5418, 0.7892, 1.7911)	Triangular (0.6865, 1.0000, 2.2694)
Weight Growth [2]	Triangular (0.5568, 0.6370, 1.8062)	Triangular (0.5850, 0.6692, 1.8975)	Triangular (0.8742, 1.0000, 2.8357)
	Triangular (0.5853, 0.7180, 1.6967)	Triangular (0.6092, 0.7474, 1.7662)	Triangular (0.8151, 1.0000, 2.3631)
	Triangular (0.5043, 0.6168, 1.8789)	Triangular (0.5316, 0.6502, 1.9805)	Triangular (0.8176, 1.0000, 3.0461)
T1 / Weight [3]	Triangular (0.3442, 0.6256, 2.0302)	Triangular (0.3649, 0.6631, 2.1520)	Triangular (0.5503, 1.0000, 3.2454)
	Lognormal (1.0000, 0.7077)	Lognormal (1.2251, 0.8670)	Lognormal (1.8387, 1.3013)
NRE / Weight [3]	Triangular (0.0283, 0.1925, 2.7792)	Triangular (0.0316, 0.2155, 3.1103)	Triangular (0.1468, 1.0000, 14.4348)
	Lognormal (1.0000, 0.9282)	Lognormal (1.3644, 1.2663)	Lognormal (2.5397, 2.3573)
NRE / T1 [3]	Triangular (0.0000, 0.1905, 2.8095)	Triangular (0.0000, 0.2136, 3.1495)	Triangular (0.0000, 1.0000, 14.7445)
	Lognormal (1.0000, 1.3844)	Lognormal (1.7078, 2.3642)	Lognormal (4.9806, 6.8950)





Source: Real Data, Real Uncertainty, Dr. Wilson Rosa AFCAA, Alfred Smith Tecolote Research, Jeff McDowell Tecolote Research, Dr. Lew Fichter Tecolote Research, 27 June 2012.

Figure 10: Use of CRUAMM Tables

CRUAMM DATA ANALYSIS

CRUAMM data was organized into many categories representing differ estimating methodologies and type of WBS elements. One category (All CERs) is used here to exemplify the data analysis process.

The four distribution types considered in the CRUAMM study were lognormal, beta, triangular, and normal. Throughout this paper, the terms “tidy” and “untidy” are used to simply divide these four distributions into two categories. The terms do not convey desirability or suitability. Lognormal and beta are called untidy because lognormal distributions are skewed with an infinite right tail and beta distributions can adopt virtually any shape. In contrast, normal and triangular distributions are called tidy because normal distributions are symmetric and triangular distributions are limited to a linear reduction in probability from the mode to the bound. In many of the charts to follow, the tidy distributions are depicted in the color blue and the untidy distributions are depicted in the color red.

Figure 11, Fit Results, describes the category in terms of number of distributions and range of fitted CVs. The Fit Results panel also describes the fit results by distribution type. The number of distributions is shown as a pie chart as well as in tabular format. In this example 153 fitted distributions comprise this category. The fitted CVs range from a low of 0.048 to a high of 1.149. The triangular distribution was the most common being selected 67 times and representing 44% of the fits. The average CV over those 67 triangular fits is 0.341.

Fit Results

Number of Fitted Distributions	153
Range of Fitted CVs	0.048 - 1.149

	Selection Count	Average Fitted CV
Triangular	67	0.341
Normal	19	0.270
Beta	6	0.687
Lognormal	61	0.488

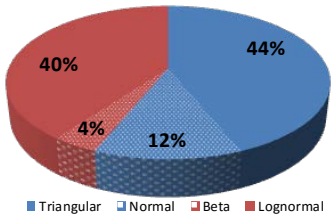
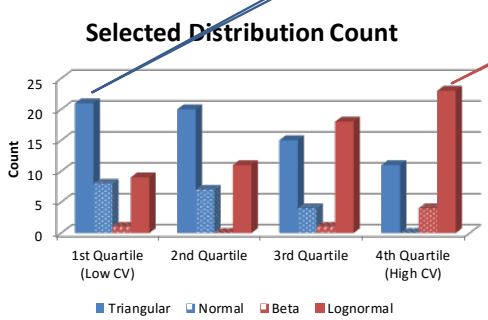


Figure 11: Fit Results Example

Figure 12, Quartile Results, describes the results by CV quartile where the first quartile is the lowest dispersion and the fourth quartile is the highest dispersion. This example shows that in the lowest quartile (where CVs ranged from 0.048 – 0.160) the triangular was the most common distribution being selected 21 times. In the highest quartile (where CVs ranged from 0.516 to 1.149) lognormal was the most common, being selected 23 times. The results are shown in tabular format and as a column chart, where height is the number of times selected. The final chart in this panel presents the same data on an area chart depicting the relative occurrences of each distribution by quartile as proportions. The fill colors have been selected to visually amplify the relative occurrence of tidy distributions to untidy ones.

Quartile Results

	Selection Count			
	1st Quartile (Low CV)	2nd Quartile	3rd Quartile	4th Quartile (High CV)
CV Min	0.048	0.163	0.390	0.516
CV Max	0.160	0.390	0.512	1.149
Triangular	21	20	15	11
Normal	8	7	4	0
Beta	1	0	1	4
Lognormal	9	11	18	23



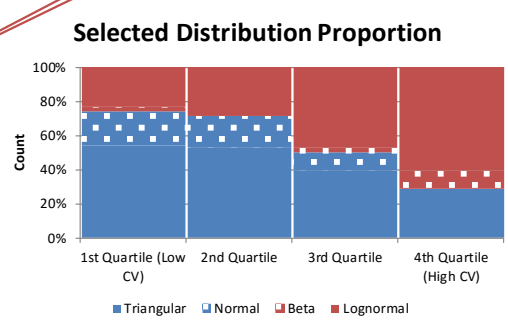


Figure 12: Quartile Results Example

ORDERS OF DISPERSION

A taxonomy for organizing the collection of CRUAMM residuals and their fitted distributions is based on the estimating methodology each fitted distribution represents. In the case of a CER, cost has been regressed against one or more cost drivers resulting in a fitted line as shown in Figure 13. Each data point in the regression data set is an actual observed value; each data point has an associated predicted value[5] which is the value of the CER at that point's cost driver value. The difference between the two is the error or residual value. (In fact the regression line was determined by some minimizing process on some form of those errors.) The residual may be expressed as multiplicative errors by dividing the actual by predicted so that actuals below the regression line have residual values less than one and actuals above the regression line have residual values greater than one. The full set of residuals will have a mean and will have a standard deviation. The standard deviation divided by the mean is a measure of the residuals' dispersion called the CV. When the residuals are binned for the purposes of distribution fitting the resulting histogram exhibits a mode near one. In CRUAMM, the residuals were placed in ascending order and CDF determined. Candidate distributions were then fitted to the residuals and the distribution with the lowest Sum of Squared Error (SSE) is the number one ranked distribution. In summary, the CER has three characteristics: It is based on a collection of like items (homogeneity), it has a cost driver (independent variable), and its form has been determined via some error minimization process (fit). It follows then the fitted distribution of the CER's residuals is also a function of those same three characteristics.

[5] Regression was performed using ordinary least squares yielding the familiar predictive statistics.

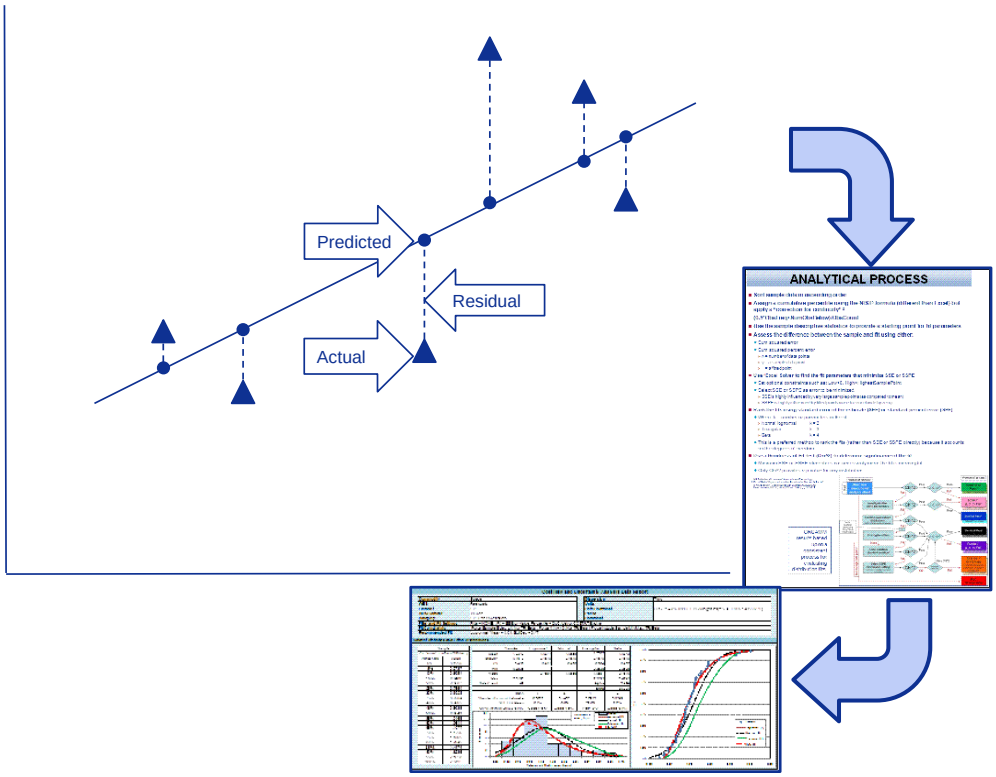


Figure 13: Fitting Distributions to CERs

Now consider the case of fitting distributions to factors. In the case of a factor relationship, each data point's cost is divided by its cost driver. The average (or median) of these become the methodology. The relationship can be drawn on a scatterplot, as was done for the CER, as depicted in Figure 14. And, as was done with the CER, each datapoint is an actual and has an associated predicted value and an associated residual. From here the rest of the process proceeds as before. In summary, the factor has two characteristics: It is based on a collection of like items (homogeneity) and it has a cost driver (independent variable). However its form is not determined via some error minimization process (it lacks the characteristic of fit). It follows that the fitted distribution of the factor's residuals is also a function of those same two characteristics.

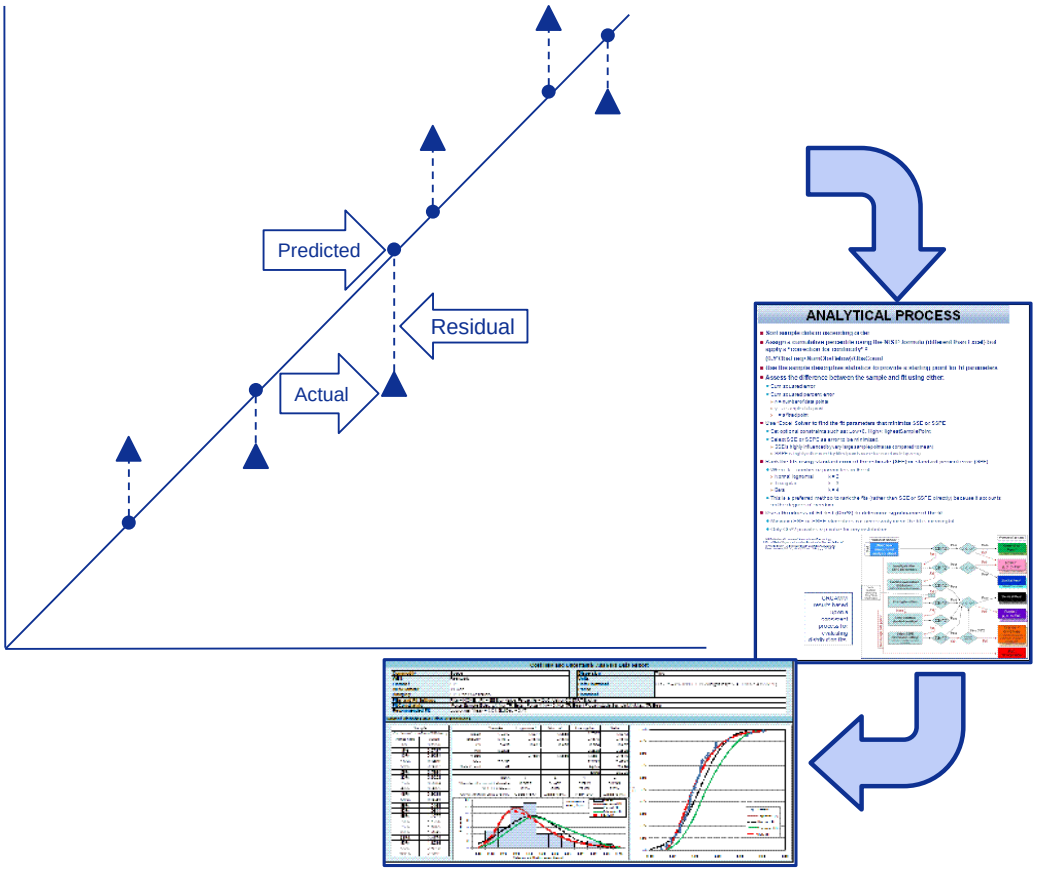


Figure 14: Fitting Distributions to Factors

Now consider the case of fitting distributions to unstructured data. In the case of a simple value relationship such as average price of an item, the central tendency of the data (the average or the median) becomes the methodology. This relationship, like the two preceding types of relationships, may be depicted on a scatterplot as depicted in Figure 15. Once again, each data point is an actual with an associated predicted value and a residual value. As before, the residuals were processed into fitted distributions. In summary, the simple value methodology has one characteristic: It is based on a collection of like items (homogeneity). It lacks a cost driver (independent variable). Its form was not determined via some error minimization process (it lacks the characteristic of fit). It follows then the fitted distribution of the simple value methodology's residuals is also a function of this single characteristic.

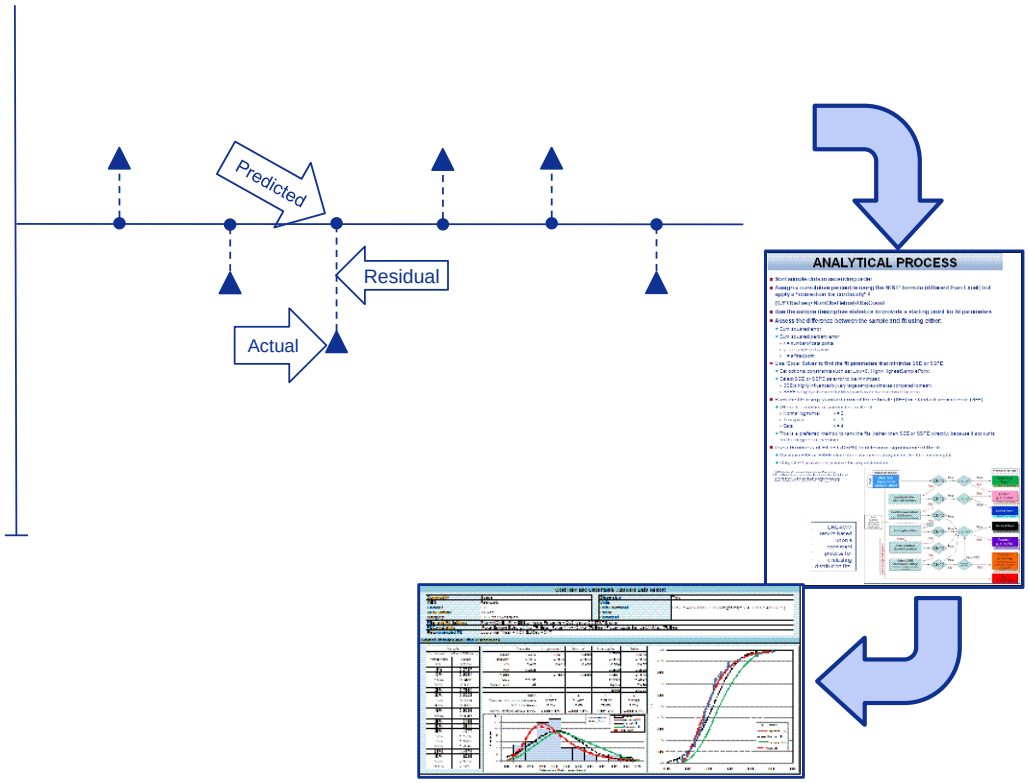


Figure 15: Fitting Distributions to Data

In summary:

CER Distributions = *f*(data homogeneity, use of a cost driver, error minimization)
Factor Distributions = *f*(data homogeneity, use of a cost driver)
Data-only Distributions = *f*(data homogeneity)

Given the relative amount of information in each of these three settings, it is reasonable to expect tighter dispersions would emanate from the settings utilizing the greatest information. Therefore, one would expect.

CER Dispersion < Factor Dispersion < Data-only Dispersion

The lexicon for describing these three settings is:

First Order Dispersion (Data-only)
Second Order Dispersion (Factor)
Third Order Dispersion (CER)

CLUSTER ANALYSIS ON FITTED DISTRIBUTIONS

This section presents the cluster analysis results. The analysis first addresses the CV to be assigned to each of the nine clusters and second addresses the distribution for each cluster. To determine a recommended CV for each of the nine gradations, cluster analysis is utilized to form nine representative CV values. To determine a distribution for each of the nine gradations, the most commonly selected distributions within each cluster will be recommended.

Clustering is a process of slicing a set of data into a set of meaningful subclasses called clusters. There are numerous clustering techniques which fall into two primary categories: Hierarchical or Partitional. Heirarchical clustering algorithms determine not only the cluster content but determine the number of final bins. Since the problem at hand has already been structured as a predetermined number of bins (nine), a hierarchical technique is not necessary. A Partional technique, *k-means*, was selected for use on this study. K-means is a centroid-based method defined as an optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Each of the orders of dispersion categories was separately examined. The first subset (All CERs) is used as an example to describe the analysis process. Some of these figures resemble the results by quartile from Section 2 but this time will be in nine partitions rather than four.

Figure 16, Cluster Results, presents the cluster results for the All CERs category with the nine large markers labeled with the centroid value. The smaller markers represent the fitted distributions' CVs in the dataset belonging to each cluster. Each partition is separated with vertical distance for readability; the height of each partition has no meaning other than providing visual separation.

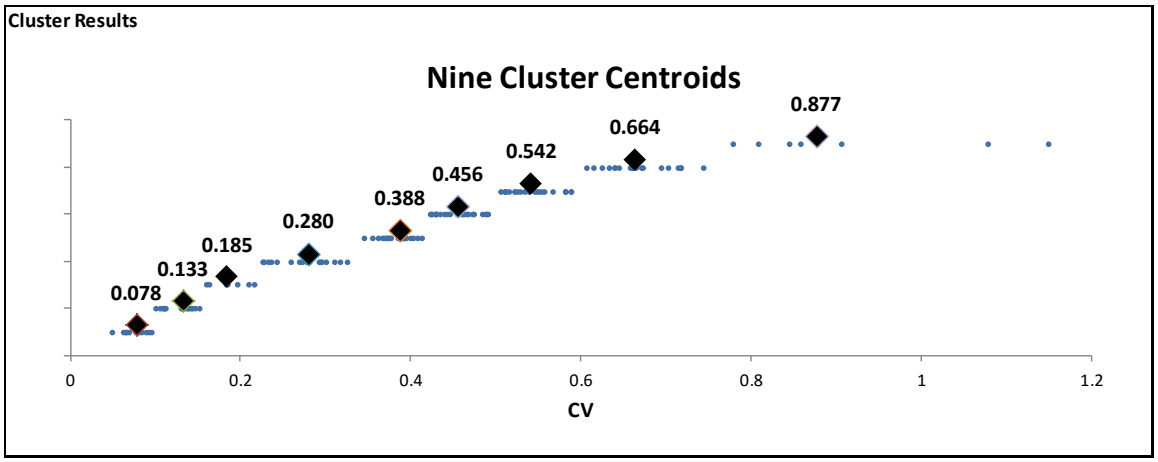
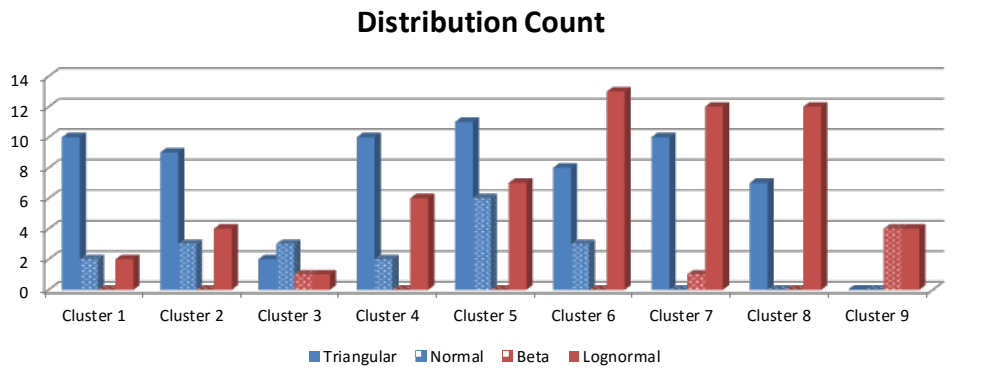


Figure 16: Cluster Results Example

Figure 17, Distribution Results by Cluster, presents the count of each selected distribution by cluster. The information is shown first as a table, then a column chart by count, and then by proportion in an area chart. The bottom of the tabular chart also shows the modal distribution for each cluster. For the highest count of tidy distributions in each cluster triangular is listed; for the highest count of untidy distributions in each cluster lognormal is listed. As was the case in the earlier quartile charts, the color scheme was chosen to amplify that tidy distributions are common to low CVs and untidy distributions are common to high CVs. Cluster 2 for example had nine triangular distributions and two lognormal distributions. In contrast, cluster 8 has twelve lognormal and seven triangular distributions.

Distribution Results by Cluster

	Number of Times Distribution was Selected								
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Triangular	10	9	2	10	11	8	10	7	0
Normal	2	3	3	2	6	3	0	0	0
Beta	0	0	1	0	0	0	1	0	4
Lognormal	2	4	1	6	7	13	12	12	4
Typical Distribution	Triangular	Triangular	Triangular	Triangular	Triangular	Lognormal	Lognormal	Lognormal	Lognormal



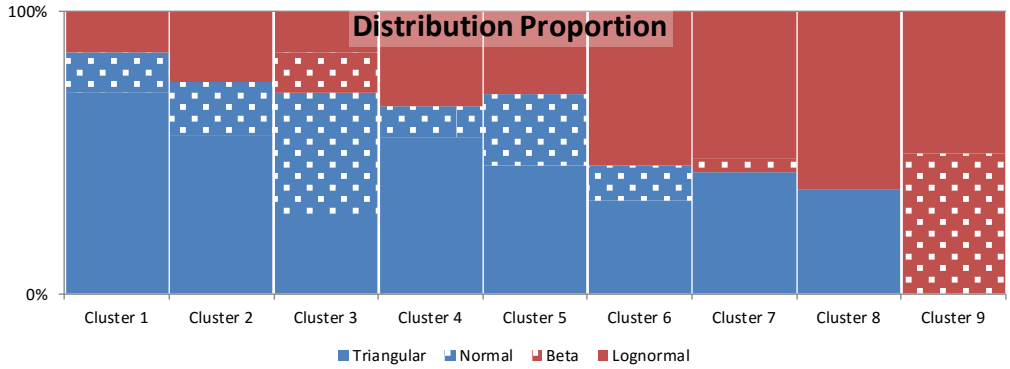


Figure 17: Distribution Results by Cluster Example

Figure 18, Final Results, presents a tabular summary of the parameters that are useful in the emerging zone methodology. The first two rows repeat the skew and CV values from the previous panels. The third row presents a selected distribution which, in most cases, is the most common distribution within that cluster. Given the infrequency of normal and beta as the modal distribution and the desirability of a final system as simple as reasonable, only lognormal and triangular were selected. In the few cases where it occurred, triangular was always substituted for normal and lognormal was always substituted for beta. The content of the remaining rows on this table vary depending on the selected distribution. In the case of triangular, the minimum and maximum values are shown. These values are the average minimum and average maximum of the triangular fits in that cluster. The

methodology shown for cluster 2 for example, is a triangular bounded by minimum/maximum values of 0.654 and 1.264. With these two values, one can model a triangular distribution in their cost model where the point estimate is assumed to be the mode. In the case of lognormal, the mean and standard deviation are shown. For example, the methodology parameters for cluster 8 are a lognormal mean of 1.210 with a standard deviation of 0.803. The mean shown on the table is the average of the means of the lognormal fits in that cluster. The standard deviation is the cluster centroid CV multiplied by that mean. With the mean and standard deviation, one can model a lognormal distribution in their cost model.

Final Results

	Methodology Parameters of Interest								
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**	**Cluster 9**
CV	0.08	0.13	0.18	0.28	0.39	0.46	0.54	0.66	0.88
Skew[1]	0.46	0.45	0.44	0.25	0.27	0.17	0.10	0.07	0.33
Select Distribution	Triangular	Triangular	Triangular	Triangular	Triangular	Lognormal	Lognormal	Lognormal	Lognormal
Triangular Min	0.813	0.654	0.517	0.518	0.329				
Triangular Max	1.160	1.264	1.465	1.768	2.092				
Lognormal Mean						1.101	1.140	1.210	1.399
Lognormal StdDev						0.502	0.618	0.803	1.227

1 - Skew expressed as CDF at the mode

Figure 18: Final Results by Cluster Example

These techniques were applied to each of three Orders of Dispersion. Figure 19 presents the cluster analysis CVs for each. Note the relative spread of dispersion by order. Figure 20 presents the selected distributions in each cluster. Note the relative occurrences of tidy and untidy distributions by order.

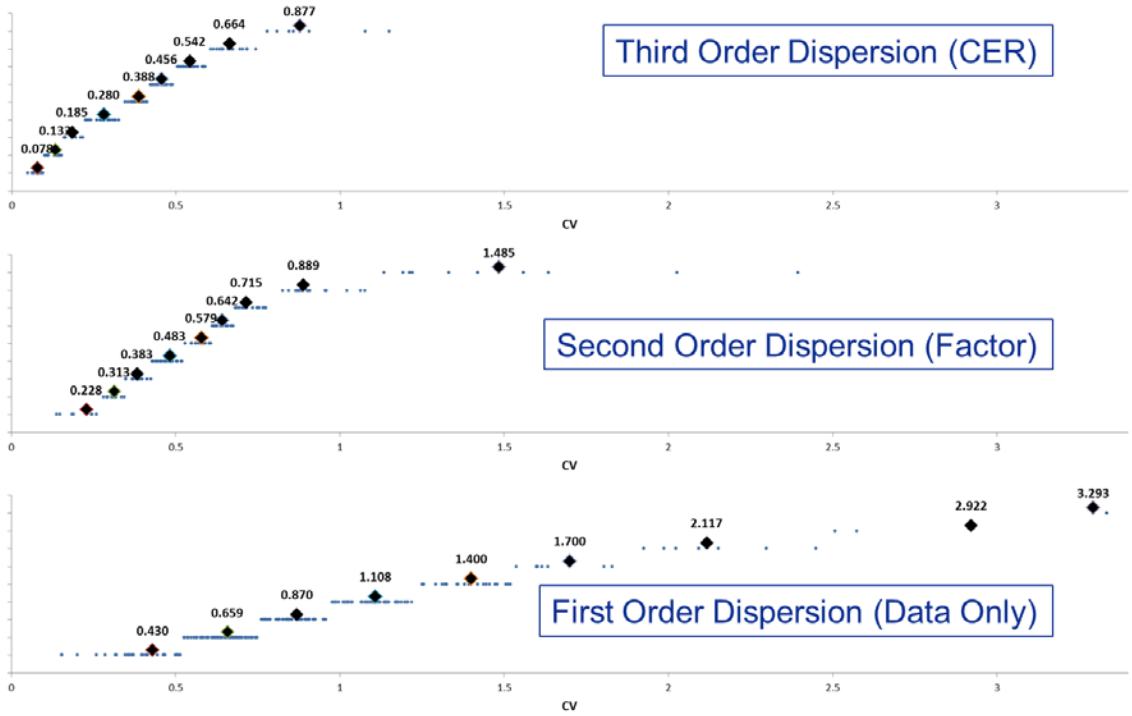


Figure 19: CV Cluster Results by Order of Dispersion

Figure 20: Distribution Shape Cluster Results by Order of Dispersion

Figure 21 presents a summary of all the parameters of interest resulting from this process. The figure presents 36 distributions (nine each for three categories plus all in an aggregate category).

	Methodology Parameters of Interest								
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**	**Cluster 9**
All CERs	TRI(0.813, 1, 1.160)	TRI(0.654, 1, 1.264)	TRI(0.517, 1, 1.465)	TRI(0.518, 1, 1.768)	TRI(0.329, 1, 2.092)	LN(1.101, 0.502)	LN(1.140, 0.618)	LN(1.210, 0.803)	LN(1.399, 1.227)
Factors, All	LN(1.124, 0.572)	LN(1.179, 0.763)	LN(1.265, 0.981)	LN(1.351, 1.231)	LN(1.405, 1.390)	LN(1.509, 1.687)	LN(1.655, 2.166)	LN(1.836, 2.752)	LN(2.298, 4.857)
All First Order Data	TRI(0.181, 1, 2.138)	TRI(0.049, 1, 3.161)	LN(1.315, 1.144)	LN(1.480, 1.640)	LN(1.714, 2.399)	LN(1.932, 3.285)	LN(2.356, 4.989)	LN(2.731, 7.978)	LN(3.550, 11.692)
Everything	TRI(0.635, 1, 1.745)	TRI(0.286, 1, 3.171)	LN(1.139, 0.625)	LN(1.214, 0.822)	LN(1.332, 1.184)	LN(1.489, 1.644)	LN(1.704, 2.352)	LN(2.070, 3.752)	LN(2.849, 7.320)

Legend
TRI(a, b, c) = Triangular(minimum, most-likely, maximum)
LN(a, b) = Lognormal(mean, standard-deviation)

Figure 21: Final Results Summary

RECOMMENDED METHODOLOGY

Figure 22 presents the recommended Zone Method of Cost Uncertainty Analysis. Each row represents a category. Each column represents one of the nine clusters. Within each cell is a distribution unitized for use on a most-likely point estimate. An analyst would use this table to select a distribution shape and distribution parameters via these four steps:

1. First, determine the degree of program challenge.
    a. Use judgment.
    b. Use a commodity-specific scoring matrix or complexity calculator.
2. Second, identify the category row that most closely matches each element's estimating methodology.
    a. Use the Zone Method table.
3. Third, determine the relative uncertainty of this cost model element compared to the rest of the cost model.
    a. Use judgment.
4. Fourth, use the distribution shape and parameters shown in the cell intersected by the first three choices.
    a. Use the Zone Method table.

			Program with Nominal Challenges						
	Shift Left for Less Challenging Programs		Low Uncertainty	Low-Medium Uncertainty	Medium Uncertainty	Medium-High Uncertainty	High Uncertainty	Shift Right for More Challenging Programs	
CER Methods	TRIANGULAR (0.813, 1, 1.160)	TRIANGULAR (0.654, 1, 1.264)	TRIANGULAR (0.517, 1, 1.465)	TRIANGULAR (0.518, 1, 1.768)	TRIANGULAR (0.329, 1, 2.092)	LOGNORMAL (1.101, 0.502)	LOGNORMAL (1.140, 0.618)	LOGNORMAL (1.210, 0.803)	LOGNORMAL (1.399, 1.227)
Factor Methods	TRIANGULAR (0.701, 1, 1.517)	TRIANGULAR (0.718, 1, 1.981)	TRIANGULAR (0.435, 1, 2.131)	TRIANGULAR (0.273, 1, 2.510)	LOGNORMAL (1.153, 0.667)	LOGNORMAL (1.188, 0.763)	LOGNORMAL (1.238, 0.885)	LOGNORMAL (1.363, 1.212)	LOGNORMAL (1.869, 2.775)
Central Trendency Methods	TRIANGULAR (0.181, 1, 2.138)	TRIANGULAR (0.049, 1, 3.161)	LOGNORMAL (1.315, 1.144)	LOGNORMAL (1.480, 1.640)	LOGNORMAL (1.714, 2.399)	LOGNORMAL (1.932, 3.285)	LOGNORMAL (2.356, 4.989)	LOGNORMAL (2.731, 7.978)	LOGNORMAL (3.550, 11.692)
	Shift Left for Less Challenging Programs		Low Uncertainty	Low-Medium Uncertainty	Medium Uncertainty	Medium-High Uncertainty	High Uncertainty	Shift Right for More Challenging Programs	
			Program with Nominal Challenges						

Legend

TRIANGULAR(a, b, c) = Triangular(minimum, most-likely, maximum)

LOGNORMAL(a, b) = Lognormal(mean, standard-deviation)

Figure 22: The Zone Method of Uncertainty Analysis

SUMMARY

The purpose of this study was to examine the body of work from the CRUAMM task to make top-down observations and offer top-level guidance for use in cost uncertainty analysis. This document provides the analyst a methodology for defining distributions for use in performing cost uncertainty analysis. It is also a useful reference to provide a cross-check when the analyst has developed their own uncertainty distributions.

While the Zone Method guidelines may be useful as a quick-turn method, or for gut-checking another's estimate, or serving as a source for relative uncertainty bounds, it is not intended to supplant the current best practices of detailed uncertainty analysis using program-specific data.