PRT-152

Fit, Rather Than Assume, a CER Error Distribution

Dr. Shu-Ping Hu

2013 ICEAA Annual Conference and Training Workshop
New Orleans, LA
18 to 21 June 2013

TECOLOTE RESEARCH, INC.
420 S. Fairview Avenue, Suite 201
Goleta, CA 93117-3626
(805) 571-6366

TECOLOTE RESEARCH, INC.
5266 Hollister Ave., Suite 301
Santa Barbara, CA 93111-2089
(805) 964-6963

Fit, Rather Than Assume, a CER Error Distribution

Dr. Shu-Ping Hu

ABSTRACT

Analysts usually **assume** a distribution (e.g., normal, log-normal, or triangular) to model the errors of a cost estimating relationship (CER) for cost uncertainly analysis. However, this hypothetical assumption may not be suitable to model the underlying distribution of CER errors. A distribution fitting tool is often used to hypothesize an appropriate distribution for a given set of data. It can also be applied to fit a distribution to

1. the CER residuals (i.e., Actual - Predicted = $y_i - \hat{y}_i$) for additive error models and
2. the CER "percent" errors in the form of ratios (i.e., Actual/Predicted = $y_i / \hat{y}_i$) for multiplicative error models.

This way, the CER error distribution is derived based upon the residuals (or percent errors) specific to the analysis, rather than a generic assumption applied to any analysis.

If we use a distribution fitting tool to analyze the $y_i/\hat{y}_i$ ratios for a multiplicative error CER, we cannot apply the fitted distribution directly for cost uncertainty analysis. This is because it does not account for (1) a distance assessment between the estimating point and the centroid of the data set, (2) the sample size, or (3) the degrees of freedom of the respective CER. We must make adjustments when using the fitted distribution to perform uncertainty analysis in a simulation tool. This paper proposes an objective method to account for the above elements when modeling CER uncertainty with a fitted distribution; namely, it develops a prediction interval for cost uncertainty analysis using a distribution fitting tool.

Furthermore, analysts often use a distribution fitting tool to analyze the residuals (or percentage errors) from various CERs all together. This paper discusses issues associated with this approach and explains why it is not appropriate to do so.

OUTLINE

As noted above, a distribution fitting tool can be used to hypothesize an appropriate distribution for a given set of data points, including the errors of CERs. When analyzing CER errors, the fitted distribution should be adjusted properly to build prediction intervals for cost uncertainty analysis. This specific topic, however, has not yet been discussed in the cost community. Hence, the primary objective of this paper is to develop easy-to-follow guidance for analysts to derive distribution fitting tool results for cost uncertainty analysis.

The topics discussed will be as follows:

- Common questions regarding fitting CER errors
- What should we analyze for log-linear CERs?
- Prediction interval (PI) analysis
- Adjustment factors for cost uncertainty analysis
- Easy-to-follow implementation steps

- Concerns about analyzing different CER errors all together
- Analysis of the Unmanned Space Vehicle Cost Model, Ninth Edition (USCM9) subsystem-level CERs
- Conclusions
- Recommendations and future study items

COMMON QUESTIONS REGARDING FITTING CER ERRORS

Listed below are a few common questions that occur when fitting the residuals or percent errors using a distribution finding tool:

- Should we analyze the residuals or standardized residuals for ordinary least squares (OLS) models?
- Should we analyze the CER percent errors in the form of ratios ($y_i/\hat{y}_i$) or the standardized residuals (i.e., normalized residuals) for linear Minimum-Unbiased-Percentage-Error (MUPE) equations?
- Should we analyze the CER percent errors as ratios ($y_i/\hat{y}_i$) or should we analyze the percent errors by the error definition (i.e., $(y_i - \hat{y}_i)/\hat{y}_i = y_i/\hat{y}_i - 1$) for nonlinear MUPE CERs, e.g., $y = ax^b$?

Before answering these questions, we will first define additive and multiplicative error models.

Additive Error Model. An additive error model is generally stated as follows:

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i = f_i + \varepsilon_i \quad \text{for i = 1, …, n} \tag{1}$$

where:

Y_i = observed cost of the i^{th} data point, i = 1 to n
$f(\mathbf{x}_i, \boldsymbol{\beta}) = f_i$ = the value of the hypothesized equation at the i^{th} data point
$\mathbf{x}_i$ = vector of the cost driver variables at the i^{th} data point
$\boldsymbol{\beta}$ = vector of coefficients to be estimated by the regression equation
ε_i = error term (assumed to be independent of the cost drivers)
n = sample size

Multiplicative Error Model. Similarly, a multiplicative error model is specified by

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) * \varepsilon_i = f_i * \varepsilon_i \quad \text{for i = 1, …, n} \tag{2}$$

The definitions of Y_i, $f(\mathbf{x}_i, \boldsymbol{\beta})$, and ε_i are the same as given in Equation 1. Unlike the additive error model (Equation 1), the standard deviation of the dependent variable in Equation 2 is proportional to the level of the hypothetical equation rather than some fixed amount across the entire data range. Both the MUPE and Minimum-Percentage Error Regression under Zero-Percentage Bias (ZMPE) methods are commonly used to model multiplicative error models when the error term ε is assumed to have a mean of zero and variance, σ^2. The MUPE method is an Iteratively Reweighted Least Squares (IRLS) regression technique (see References 2, 3, and 5 for details). For a detailed explanation of the ZMPE method, see Reference 4.

Log-Error Model. If the multiplicative error term (ε_i) in Equation 2 is further assumed to follow a log-normal distribution, then the error can be measured by the following:

$$e_i = \ln(\varepsilon_i) = \ln(Y_i) - \ln(f(\mathbf{x}_i, \boldsymbol{\beta})) \tag{3}$$

where "ln" stands for the natural logarithm function. The objective is then to minimize the sum of squared e_is (i.e., $(\Sigma(\ln(\varepsilon_i))^2$). If the transformed function is linear in log space, then OLS can be applied in log space to derive a solution for $\boldsymbol{\beta}$. If not, we need to apply a non-linear regression technique to derive a solution.

Although the theoretical errors (ε_i's) are assumed to be independently and identically distributed for the entire data range, the residuals of the fitted CER are not independent and they do not have the same variance either. Hence, it is recommended that we examine the standardized residual plots, rather than the residual plots, for any remaining, unexplained variations in the CER.

However, for most of the cases, the residual and standardized residual plots tend to portray the same pattern, and very little information is lost using the residual plot alone. This is also true when analyzing the residuals using the distribution finding tool for OLS models. Using either residuals or standardized residuals does not make any difference, as their respective histograms look very similar. So we conclude that it is adequate to analyze the residuals using a distribution fitting tool for most OLS models.

As for the MUPE models, we should just fit the ratios, not the standardized residuals, because the latter is always centered on zero, which cannot be fitted using a log-normal distribution unless a location parameter is used. In addition, the coefficient of variation (CV) measure does not make sense when the sample mean is close to zero. However, if a normal distribution is chosen to model the residuals or ratios, standardized residuals can be used to further confirm whether normal distribution is a good candidate for modeling CER errors.

WHAT SHOULD WE ANALYZE FOR LOG-LINEAR CERS?

There are two commonly used methods for fitting a log-normal distribution to the CER percent errors in the form of ($y_i/\hat{y}_i$) ratios: the Maximum-Likelihood Estimation (MLE) solution and the distribution fitting tool solution. Note that the former is generated in log space while the latter is fitted in unit space. In this section, we will determine whether we should analyze the percent errors as ratios ($y_i/\hat{y}_i$) in unit space or the residuals in log space for log-linear CERs.

I. Log-Space MLE method. Using the MLE method, the log-space mean and standard deviation of the log-normal distribution are given by

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\left(\ln(y_i) - \ln(\hat{y}_i)\right)}{n} \tag{4}$$

$$\hat{\sigma} = \frac{\sum_{i=1}^{n}\left(\ln(y_i/\hat{y}_i) - 0\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(\ln(y_i) - \ln(\hat{y}_i)\right)^2}{n} \tag{5}$$

Both $\hat{\mu}$ and $\hat{\sigma}$ are evaluated in log space and $\hat{\mu}$ should be zero for log-linear CERs. Note that the degrees of freedom (DF) calculation is different between Crystal Ball (CB) and @Risk. CB

uses (n-1) in the denominator of Equation 5 to adjust for DF, while @Risk uses the sample size n. In fact, we should use (n-p) if the predicted value is based upon a CER with (n-p) DF, where p is the number of estimated parameters in the CER.

II. Unit-Space Method. For purposes of illustration, we use Distribution Finder for this example. The unit-space "Least Square" solution (provided by Distribution Finder) for fitting the log-normal distribution is derived by

$$\text{Minimizing } \sum_{i=1}^{n}\left((y_i/\hat{y}_i) - Loginv((0.5*\text{ObsFreq} + \text{NumObsBelow})/n, \mu, \sigma)\right)^2 \quad (6)$$

where:

ObsFreq = the number of sample points equal to y_i, inclusive
NumObsBelow = the number of observations below the value of y_i

It is obvious that the log-space MLE solution (Equations 4 and 5) and the unit-space Least Square solution (Equation 6) are not the same.

Using the sample data in Appendix A, a weight-based log-linear CER is derived when the regression is done in log space:

$$\text{Cost} = 19.0468 * (\text{Weight})^{0.8391} \text{ (Standard error in log space} = 0.483, \text{ N} = 47) \quad (7)$$

If we fit the percent errors in the form of ratios $(y_i/\hat{y}_i)$ using Distribution Finder, the fitted log-normal distribution has a mean of 1.092 and a standard deviation of 0.6130 (both are unit-space statistics).

Given the unit-space mean (Mean) and standard deviation (Stdev), the log-space mean (μ) and standard deviation (σ) can be derived by the following equations:

$$\sigma \text{ (in log space)} = \sqrt{\ln(1+(Stdev/Mean)^2)} \quad (8)$$

$$\mu \text{ (in log space)} = \ln(\text{Mean}) - \sigma^2/2 \quad (9)$$

Using Equations 8 and 9, the **log**-space mean and standard error (SE) are given by -0.0297 and 0.5163, respectively. Note that this log-space SE (0.5163) is already 7% larger than the SE (0.483) generated by the log-linear CER (Equation 7) without applying any adjustments for the location of the estimating point, sample size, etc.

However, if we fit the log-space residuals using Distribution Finder, the standard error of estimate in log space is 0.4642. With the adjustments for sample size and degrees of freedom (see the details in the section below), the Distribution Finder results will be compatible with the regression statistics associated with Equation 7. Therefore, we conclude that we should fit the log-space residuals instead of percent errors (in the form of $y_i/\hat{y}_i$) for log-error models.

PREDICTION INTERVAL (PI) ANALYSIS

We will first discuss the prediction interval formulas before analyzing the appropriate factors to apply to the residuals or percent errors when using a distribution finding tool.

What is prediction interval? The proper measure of the quality of the estimate is the prediction interval. A PI provides a range of values around the point estimate (PE) at different probability levels to show the degree of confidence in the estimate based upon the sample

evidence. The upper and lower bounds of the intervals form the branches of hyperbolas about the regression equation and illustrate the usefulness of the equation for predicting individual values from the independent variables.

A prediction interval can be thought of as a range defined by the PE plus or minus some number of adjusted standard errors (standard errors adjusted for prediction), depending upon the level of confidence. This adjusted standard error (Adj. SE) is a function of the standard error of the regression, the sample size, and the distance assessment of the estimating point from the center of the database used to generate the CER. We now discuss additive vs. multiplicative PIs for various regression models, beginning with a simple OLS equation.

In the discussion below, the error term ε is assumed to follow a normal distribution, while α indicates the significance level of the test. (The significance level is often bounded between 0.4 and 0.99: $0.4 \le \alpha \le 0.99$.)

Simple OLS. In a simple linear CER with an additive error term, where $Y = \beta_0 + \beta_1 X + \varepsilon$, a $(1-\alpha)100\%$ PI for a future observation Y, when $X = x_o$ is given by

$$\begin{aligned} PI &= \hat{y}_0 \pm t_{(\alpha/2,n-2)} * SE\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} \\ &= \hat{y}_0 \pm t_{(\alpha/2,n-2)} * SE\sqrt{1+\frac{1}{n}+\frac{((x_0-\bar{x})/S_x)^2}{n}} \\ &= \hat{y}_0 \pm t_{(\alpha/2,n-2)} * (Adj.SE) \end{aligned} \quad (10)$$

where:

$\hat{y}_o$ = the estimated value from the CER when $X = x_0$ (also referred to as $f(x_0)$)
x_0 = the value of the independent variable used in calculating the estimate
n = the number of data points
$t_{(\alpha/2,n-2)}$ = the upper $\alpha/2$ cut-off point of Student's t distribution with (n-2) DF
SE = CER's standard error of estimate (also referred to as SEE)
Adj. SE = the adjusted standard error for PI
$\bar{x}$ = $(\sum_{i=1}^{n} x_i)/n$; the mean of the independent variable in the data set
SS_{xx} = $\sum_{i=1}^{n}(x_i-\bar{x})^2$; the sum of squares of the independent variable about its mean
S_x = $\sqrt{SS_{xx}/n}$; the uncorrected sample standard deviation of the independent variable
ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution)

For the simple factor equation $Y = \beta X + \varepsilon$, a $(1-\alpha)100\%$ PI for a future observation Y, when $X = x_o$, is given by

$$PI = \hat{y}_0 \pm t_{(\alpha/2,n-1)} * SE * \sqrt{1+\frac{x_0^2}{\sum_{i=1}^{n} x_i^2}} = bx_0 \pm t_{(\alpha/2,n-1)} * (Adj.SE) \quad (11)$$

where b is the estimated factor, $t_{(\alpha/2,n-1)}$ is the upper $\alpha/2$ cut-off point of Student's t distribution with (n-1) DF, and the rest are defined above.

Multi-Variable OLS. If there are multiple predictors in the CER, namely, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k + \varepsilon$, we will compute the PI using matrix operations. A $(1-\alpha)100\%$ PI for a future observation Y at a given driver vector $\underline{\mathbf{x}}_\mathbf{o}$ is given below:

$$\hat{y}_0 \pm t_{(\alpha/2,\,n-p)} * SE * \sqrt{1 + (\underline{\mathbf{x}}_\mathbf{0})(\mathbf{X}'\mathbf{X})^{-1}(\underline{\mathbf{x}}_\mathbf{0})'} = \hat{y}_0 \pm t_{(\alpha/2,\,n-p)} * (Adj.SE) \tag{12}$$

where:

$\hat{y}_0$ = the estimated value of Y from the CER when $\underline{\mathbf{x}} = \underline{\mathbf{x}}_\mathbf{o}$
$\underline{\mathbf{x}}_\mathbf{o}$ = $(1, x_{1o}, \ldots, x_{ko})$, a row vector of given driver values and 1 is for the intercept
p = the total number of estimated parameters, including the intercept (p = k+1)
n = the number of data points
$t_{(\alpha/2,n-p)}$ = the upper $\alpha/2$ cut-off point of Student's t distribution with (n-p) DF
SE = CER's standard error of estimate
Adj. SE = the adjusted standard error for PI
X = the design matrix of the independent variables
ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution)
(The apostrophe superscript denotes the transpose of a vector or a matrix.)

Simple OLS in Log Space (LOLS). In a simple log-linear CER with an multiplicative error term, where $Y = (\beta_0 X^{\beta 1}) * \varepsilon$, a $(1-\alpha)100\%$ PI for a future observation Y, when $X = x_o$ is given by

$$PI = Exp\left(\hat{y}_0 \pm t_{(\alpha/2,n-2)} * SE \sqrt{1 + \frac{1}{n} + \frac{(\ln(x_0) - \overline{\ln(x)})^2}{SS_{xx}}} \right) \tag{13}$$
$$= Exp\left(\hat{y}_0 \pm t_{(\alpha/2,n-2)} * (Adj.\,SE) \right)$$

where:

$\hat{y}_0$ = the estimated value in **log space** when $X = x_0$,
x_0 = the value of the independent variable used in calculating the estimate
$t_{(\alpha/2,n-2)}$ = the upper $\alpha/2$ cut-off point of Student's t distribution with (n-2) DF
$\overline{\ln(x)}$ = the average value of the independent variable evaluated in **log** space,
SE = the standard error of estimate in **log** space
Adj. SE = the adjusted standard error for PI in log space
SS_{xx} = the sum of squares of the independent variable about its mean (in log space)
"ln" = the natural logarithm function

LOLS with Multiple Drivers. If there are multiple drivers in a log-linear CER, we will compute PI using matrix operations. See the $(1-\alpha)100\%$ PI formula below for a log-linear CER at a given driver vector $\underline{\mathbf{x}}_\mathbf{0}$:

$$PI = Exp\left(\hat{y}_0 \pm t_{(\alpha/2,\,n-p)} * SE * \sqrt{1 + \ln(\underline{\mathbf{x}}_{\mathbf{0}})(\mathbf{X}'\mathbf{X})^{-1}\ln(\underline{\mathbf{x}}_{\mathbf{0}})'}\right)$$
$$= Exp\left(\hat{y}_0 \pm t_{(\alpha/2,\,n-p)} * (Adj.SE)\right) \quad (14)$$

where:

$\hat{y}_0$ = the estimated value from the CER in log space when X = $\underline{\mathbf{x}}_{\mathbf{o}}$
n = the number of data points
p = the total number of estimated parameters, including the intercept
$t_{(\alpha/2,n-p)}$ = the upper $\alpha/2$ cut-off point of Student's t distribution with (n-p) DF
SE = CER's standard error of estimate (evaluated in log space)
Adj. SE = the adjusted standard error for PI
$\ln(\underline{\mathbf{x}}_{\mathbf{o}})$ = $(1, \ln(x_{1o}), ..., \ln(x_{ko}))$, a row vector of given driver values in log space and 1 is for the intercept. (Note: p = k+1)
X = the design matrix of the independent variables in **log** space
ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution)

(The apostrophe superscript denotes the transpose of a vector or a matrix.)

Since the computation of PI for learning curves is done in **log** space, the resultant PI in **unit** space will be asymmetrical.

Simple Linear MUPE. If a MUPE equation is hypothesized as $Y = (\beta_0 + \beta_1 X)*\varepsilon$, then a $(1-\alpha)100\%$ PI for a future observation Y, when X is at x_o is given below:

$$PI = \hat{y}_0(when\ X = x_0) \pm t_{(\alpha/2,\,n-2)} * SE * \sqrt{\frac{1}{w_0} + \frac{1}{\sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{(\sum w_i)S_{wx}^2}}$$
$$= \hat{y}_0 \pm t_{(\alpha/2,\,n-2)} * SE * \sqrt{\hat{y}_0^2 + \frac{1}{\sum w_i} + \frac{((x_0 - \bar{x}_w)/S_{wx})^2}{\sum w_i}} \quad (15)$$
$$= \hat{y}_0\left(1 \pm t_{(\alpha/2,\,n-2)} * SE * \sqrt{1 + \frac{1}{\hat{y}_0^2\sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{\hat{y}_0^2(SS_{wxx})}}\right) = \hat{y}_0\left(1 \pm t_{(\alpha/2,\,n-2)} * (Adj.\,SE)\right)$$

where:

$\hat{y}_0$ = the estimated value from the CER when $X = x_0$ (also denoted by $f(x_o)$)
x_0 = the value of the independent variable used in calculating the estimate
n = the number of data points
$t_{(\alpha/2,n-2)}$ = the upper $\alpha/2$ cut-off point of the t-distribution with (n-2) DF
SE = CER's standard error of estimate
Adj. SE = the adjusted standard error for PI
w_0 = the weighting factor for y when $x = x_o$; $w_0 = 1/(f^2(x_o)) = 1/\hat{y}_0^2$ for MUPE CER
w_i = the weighting factor for the i^{th} data point; $w_i = 1/(f^2(x_i))$ for MUPE CER
$f(x_i)$ = the predicted value of the i^{th} data point
$\bar{x}_w$ = $\sum_{i=1}^{n} w_i(x_i) \Big/ \sum_{i=1}^{n} w_i \ \bar{x}_w$

$$SS_{wxx} = \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2$$

$$S_{wx} = \sqrt{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 / \sum_{i=1}^{n} w_i}$$

ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution)

Note that S_{wx} is the weighted sample standard deviation of the independent variable x. It is the sample standard deviation of the independent variable evaluated in the **fit** space.

Simple Factor MUPE. If a MUPE factor CER is hypothesized as $Y = \beta X^* \varepsilon$, then a (1-$\alpha$)100% PI for a future observation Y, when X is at x_o is given below:

$$PI = f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{SS_{wxx}}} = f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{(1+\frac{1}{n}) b^2 x_0^2}$$
$$= bx_0 \left(1 \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{1+\frac{1}{n}} \right) = bx_0 (1 \pm t_{(\alpha/2,\, n-1)} * Adj.\ SE) \quad (16)$$

where:

x_0 = the value of the independent variable used in calculating the estimate

$f(x_o)$ = the estimated value from the CER when $X = x_0$ (also denoted by $\hat{y}_0$)

n = the number of data points

$t_{(\alpha/2,n-2)}$ = the upper $\alpha/2$ cut-off point of the t-distribution with (n-2) DF

SE = CER's standard error of estimate

Adj. SE = the adjusted standard error for PI

w_0 = the weighting factor for y when $x = x_o$; $w_0 = 1/(f^2(x_o)) = 1/\hat{y}_0^2$ for MUPE CER

w_i = the weighting factor for the i^{th} data point; $w_i = 1/(f^2(x_i))$ for MUPE CER

$$SS_{wxx} = \sum_{i=1}^{n} w_i (x_i^2)$$

ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution)

Based upon Equation 16, we do not need the actual data set to build a PI for a MUPE factor CER because the adjustment is a constant factor. It can be shown that the SE in Equation 16 is equal to the CV of the Y to X ratio, i.e., SE = CV(Y/X). (See Appendix C for a detailed proof.)

Univariate Analysis. The PI listed in Equation 16 is about the same as the PI for the univariate analysis. If $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed (i.i.d.) random variables from a normal distribution with a mean of μ and variance σ^2, i.e., $N(\mu,\sigma^2)$, then the average of Y also follows a normal distribution. If we let the future observation be denoted by Y_0, then $Y_0 \sim N(\mu,\sigma^2)$. The mean and variance of the difference between Y_0 and $\bar{Y}$ are given below, since Y_0 and $\bar{Y}$ are independent:

$$E(Y_0 - \bar{Y}) = 0$$

$$Var(Y_0 - \bar{Y}) = Var(Y_0) + Var(\bar{Y}) - 2Cov(Y_0, \bar{Y}) = (1 + \frac{1}{n})\,\sigma^2$$

Therefore, we can make the following conclusions by the normal distribution theory:

$$(Y_0 - \bar{Y}) / (S_Y * \sqrt{1 + \frac{1}{n}}) \sim t_{n-1} \quad (\text{t } \textit{distribution} \text{ with n} - 1 \text{ degrees of freedom}) \tag{17}$$

where $S_Y = (\sum_{i=1}^{n} (y_i - \bar{y})^2) / (n-1)$ is the sample standard deviation of Y.

Based upon Equation 17, a (1-α)100% PI for a future observation Y_0 is given by

$$\begin{aligned} PI &= \bar{Y} \pm t_{(\alpha/2,\,n-1)} * S_Y * \sqrt{1 + \frac{1}{n}} \\ &= \bar{Y}\,(1 \pm t_{(\alpha/2,\,n-1)} \frac{S_Y}{\bar{Y}} * \sqrt{1 + \frac{1}{n}}) = \bar{Y}\,(1 \pm t_{(\alpha/2,\,n-1)} * (Adj.\,SE)) \end{aligned} \tag{18}$$

Comparing Equation 18 with Equation 16, the PI for univariate analysis is the same as the PI for the MUPE factor CER.

Multi-Variable Linear MUPE. If there are multiple predictors in the above MUPE CER, namely, $Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k) * \varepsilon$, we will compute the PI using matrix operations. A (1-α)100% PI for a future observation Y at a given predictor vector $\underline{\mathbf{x}}_\mathbf{0}$ is given by

$$\begin{aligned} PI &= \hat{y}_0 \pm (t_{\alpha/2,n-p}) * SE * \sqrt{1/w_0 + (\underline{\mathbf{x}}_\mathbf{0})(\mathbf{X}'W\mathbf{X})^{-1}(\underline{\mathbf{x}}_\mathbf{0})'} \\ &= \hat{y}_0 \left(1 \pm (t_{\alpha/2,n-p}) * (Adj.SE)\right) \end{aligned} \tag{19}$$

where:

$\hat{y}_0$	= the estimated value of Y from the CER when $\underline{\mathbf{x}} = \underline{\mathbf{x}}_\mathbf{0}$
$\underline{\mathbf{x}}_\mathbf{0}$	= $(1, x_{1o}, \ldots, x_{ko})$, a row vector of given driver values and 1 is for the intercept
p	= the total number of estimated parameters, including the intercept (p = k+1)
n	= the number of data points
$t_{(\alpha/2,n-p)}$	= the upper α/2 cut-off point of Student's t distribution with (n-p) DF
SE	= CER's standard error of estimate
Adj. SE	= the adjusted standard error for PI
w_0	= the weighting factor for y when $x = x_o$ ($w_0 = 1/(f^2(x_o)) = 1/\hat{y}_0^2$ for MUPE CER)
$\mathbf{X}$	= the design matrix of the independent variables
W	= the weighting matrix, where the i^{th} diagonal element is w_i
ε	= the error term with mean of 1 and variance σ^2 (assumed to follow a normal distribution)

(The apostrophe superscript denotes the transpose of a vector or a matrix.)

ADJUSTMENT FACTORS FOR COST UNCERTAINTY ANALYSIS

Based upon the PI formulas given in the section above, we can deduce **three** adjustment factors that can be multiplied to a fitted distribution to derive appropriate PIs for cost uncertainty analysis.

I. Sample Factor. When analyzing the residuals or percent errors (in the form of ratios $y_i/\hat{y}_i$), the distribution fitting tool does not know whether the data set is an entire population or a random sample. It does not know the degrees of freedom associated with these errors either. Hence, analysts should manually make an adjustment to account for the sampling error or the appropriate degrees of freedom if certain parameters are estimated by the sample:

$$\text{Sample factor} = \sqrt{\frac{n}{df}} \tag{20}$$

where n is the sample size and "df" stands for degrees of freedom.

II. DF Factor. As shown by Equation 12, if the error term more or less follows a normal distribution, we should use the Student's **t** distributions provided in risk analysis tools (such as Crystal Ball, @Risk, or ACEIT) to model the CER uncertainty. We can enter the *Adj. SE* into the scale field and specify the DF in the degrees of freedom field when modeling risk using a Student's t distribution. Alternatively, we can enter the low/high bounds to specify the distribution. Note that the upper cut-off point ($t_{\alpha/2,df}$) is derived from a Student's t distribution that has the same degrees of freedom as the CER. However, if we use the *Adj. SE* to model the CER uncertainty by a different distribution, a DF adjustment factor should be applied to account for small samples. For example, we should multiply the *Adj. SE* measure by the DF factor to account for the broader tails of the t distribution for small samples if we use normal instead of t distribution for cost uncertainty analysis:

$$\text{DF factor} = \sqrt{\frac{df}{df-2}} \tag{21}$$

where "df" stands for the degrees of freedom of the t distribution. In fact, Equation 21 is the standard deviation of the Student's t distribution with a scale parameter one and "df" degrees of freedom.

For LOLS CERs, we should in fact apply the **Log-t** distributions directly to model uncertainties, since the CER errors are commonly assumed to follow the log-normal distribution. (Just as with a Student's t distribution, we can enter the *Adj. SE* into the scale field and specify the DF in the degrees of freedom field when selecting a Log-t distribution. Alternatively, we can enter the low/high bounds to specify the distribution.) If the Log-t distributions are not available in the risk tools, we can specify the Student's t distributions in log space, but we should ensure the PI is transformed back to unit space as given in Equations 13 and 14.

The DF factor can be ignored when the sample size is fairly large (e.g., df > 50) or a Student's t or a Log-t distribution is chosen to model the CER errors.

III. Location Factor (LF). The last term in these PI equations is a "distance" adjustment (i.e., a location factor), which should be applied to account for the location of the estimating point. It assesses the "distance" of the estimating point from the centroid of the predictors. As shown by Equations 10 and 13, the *Adj. SE* (as well as PI) gets larger when the estimating point moves farther away from the center of the database. This is especially true when the CER is used beyond the range of the data used in developing the CER. Hence, using the CER's standard error alone for risk assessment may significantly underestimate the risk associated with the PE unless the PE is very close to the center of the database and the sample size is fairly large. For a

single variable model, the range of PI is the smallest when the estimating point is exactly the mean of the independent variable and the last term is reduced to sqrt(1+1/n) when it happens. This is also the location factor for a MUPE factor CER with one independent variable.

In ACE we define the adjusted SE based upon the distance assessment of the primary independent variable. As shown by Equation 10, the adjusted SE is given by

$$Adj.SE = SE\sqrt{1+\frac{1}{n}+\frac{((x_0-\bar{x})/S_x)^2}{n}} = SE\sqrt{1+\frac{1}{n}+\frac{(Distance/Driver\ Stdev)^2}{n}} \quad (22)$$

where:

"Distance" is the distance assessment between the estimate and the center of the primary independent variable and

"Driver Stdev" (i.e., S_x) is the **uncorrected** sample standard deviation of the primary independent variable.

Note that this distance assessment is only characterized in terms of a number of standard deviations from the center. For example, if the distance is assessed as approximately two sample standard deviations of the driver variable, then the ratio (of "Distance" to "Driver Stdev") is two. For simplicity, ACE provides the default values below to address the assessment of this distance ratio based upon the similarities between the systems:

$$\frac{Distance}{Driver\ Stdev} = \begin{cases} 0.25 & Very\ Similar \\ 0.75 & Similar \\ 1.50 & Somewhat\ Different \\ 3.00 & Very\ Different \end{cases} \quad (23)$$

For example, if the system being estimated is deemed very similar to the database from which the CER was developed, this qualitative assessment may translate into a quantitative assessment of the ratio with a value of 0.25. Similarly, if the system being estimated is deemed very different from the database from which the CER was developed, this qualitative assessment might translate into a quantitative assessment of the ratio with a value of 3.0. The adjusted standard error can then be calculated using these default values. See ACE online help for details.

The LF Table below lists the location factors for a one–independent variable model:

LF Table: Location Factor by Model Type (for one-driver variable)

Model	Location Factor = (Adj. SE)/SE
Additive – Linear	$\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} = \sqrt{1+\frac{1}{n}+\frac{((x_0-\bar{x})/S_x)^2}{n}}$
Additive – Factor	$\sqrt{1+\frac{x_0^2}{\sum x_i^2}}$
Log-Linear	$\sqrt{1+\frac{1}{n}+\frac{(\ln(x_0)-\overline{\ln(x)})^2}{\sum(\ln(x_i)-\overline{\ln(x)})^2}}$

MUPE – Linear	$\sqrt{1+\frac{1}{\hat{y}_0^2\sum w_i}+\frac{(x_0-\bar{x}_w)^2}{\hat{y}_0^2(SS_{wxx})}}$
MUPE – Factor Univariate	$\sqrt{1+\frac{1}{n}}$
Heuristic	$\sqrt{1+\frac{1}{n}+\frac{(Distance/Driver\ Stdev)^2}{n}}$

Note that x_0 is the value of the independent variable used in calculating the estimate and $\hat{y}_0$ is the estimated value from the CER when X = x_0.

The sample, DF, and location adjustments should be considered when constructing PIs. Otherwise, the range of the PI (based upon SE alone) will be smaller than it should be. Additionally, these adjustments should be applied to the residuals or percent errors **prior to** fitting them using a distribution finding tool. This way, the low/high range, mean, mode, and standard deviation of the fitted distribution are adjusted accordingly by the distribution finding tool, so analysts can use the fitted results directly for cost uncertainty analysis. It is much easier and more straightforward to make the adjustments before running the distribution finding tool, than it is to adjust several statistics for the PI after running the tool.

EASY-TO-FOLLOW IMPLEMENTATION STEPS

The primary purpose of this study is to develop easy-to-follow guidance for analysts to derive distribution fitting tool results for cost uncertainty analysis.

Actual Implementation. Based upon the analysis given in the previous section, we can define the net adjustment factor as the product of the three adjustment factors mentioned above:

Net Factor = (Sample Factor)*(DF Factor)*(Location Factor) (24)

Also, a shift should be specified for the MUPE and ZMPE CERs to ensure the fitted distribution is still centered on one:

Shift = Net Factor – 1 (25)

For example, the percent errors from a MUPE or ZMPE CER should be multiplied by the Net Factor and then subtracted by "Shift" before they are analyzed using a distribution finding tool:

$(y_i / \hat{y}_i)$ * (Net Factor) – Shift for MUPE and ZMPE models (26)

As discussed above, the PI for univariate analysis is the same as the PI for the MUPE factor CER. Hence, the above process (Equation 26) should be applied to univariate analysis as well where $\hat{y}_i$ is replaced by $\bar{y}$.

$(y_i / \bar{y})$ * (Net Factor) – Shift for univariate analysis (27)

This way, the sample mean stays the same for univariate analysis.

However, do not include the DF factor in the computation of the Net Factor if a Student's t or a Log-t distribution is chosen to model the CER errors. Also, do not apply Shift to residuals or residuals in log space for additive-error and log-linear models, respectively:

$(y_i - \hat{y}_i)$ * (Net Factor) for additive-error models (28)

$(\ln(y_i) - \ln(\hat{y}_i))$ * (Net Factor) for log-error models, fitted in **log** space (29)

In summary, below is the actual implementation of adjustments for PI by model types:

Adjustment Table: Implementation by Model Type

Models	Adjustments
Additive	$(y_i - \hat{y}_i)$ * (Net Factor)
Log-Error	$(\ln(y_i) - \ln(\hat{y}_i))$ * (Net Factor)
MUPE/ZMPE	$(y_i / \hat{y}_i)$ * (Net Factor) – Shift
Univariate	$(y_i / \bar{y})$ * (Net Factor) – Shift

Make sure all risk inputs are specified properly, especially when using one cell to capture both the PE and error distribution. Suggest using an additional cell for the error term besides the PE.

We now use a data set (in Appendix B) to illustrate how to apply the Net Factor to a MUPE linear CER:

Cost = 220.0895 + 3.8112 * Weight (SE = 28.13%, N = 49) (30)

It follows from Equation 30 that the estimated cost of a "black" box weighing 300 pounds would be $1,363.45.

Given x_0 = 300, $\hat{y}_0$ = 1,363.45, SS_{wxx} = 1.07202, $\bar{x}_w$ = 469.4747, and Σw_i = 8.27953*10^-6, the location factor (see the LF Table) is then given by

$$Location\ Factor = \sqrt{1 + \frac{1}{\hat{y}_0^2 \sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{\hat{y}_0^2 (SS_{wxx})}}$$

$$= \sqrt{1 + \frac{10\text{^}6}{(1346.45)^2 (8.27953)} + \frac{(300 - 469.4747)^2}{(1346.45)^2 (1.07202)}} = 1.038933$$

(If the actual data set is not available, we can use a heuristic approach to approximate the location factor as suggested by Equation 23.)

With 49 data points and two estimated parameters in the CER, the Sample and DF factors are calculated as follows:

Sample Factor = sqrt(49/47) = 1.0211

DF Factor = sqrt(47/45) = 1.022

If we multiply all three factors together (as suggested by Equation 24), the Net Factor is given below:

Net Factor = (1.0211)*(1.022)*(1.038933) = 1.084125

According to Equation 25, a Shift should also be specified because Equation 30 is a MUPE CER:

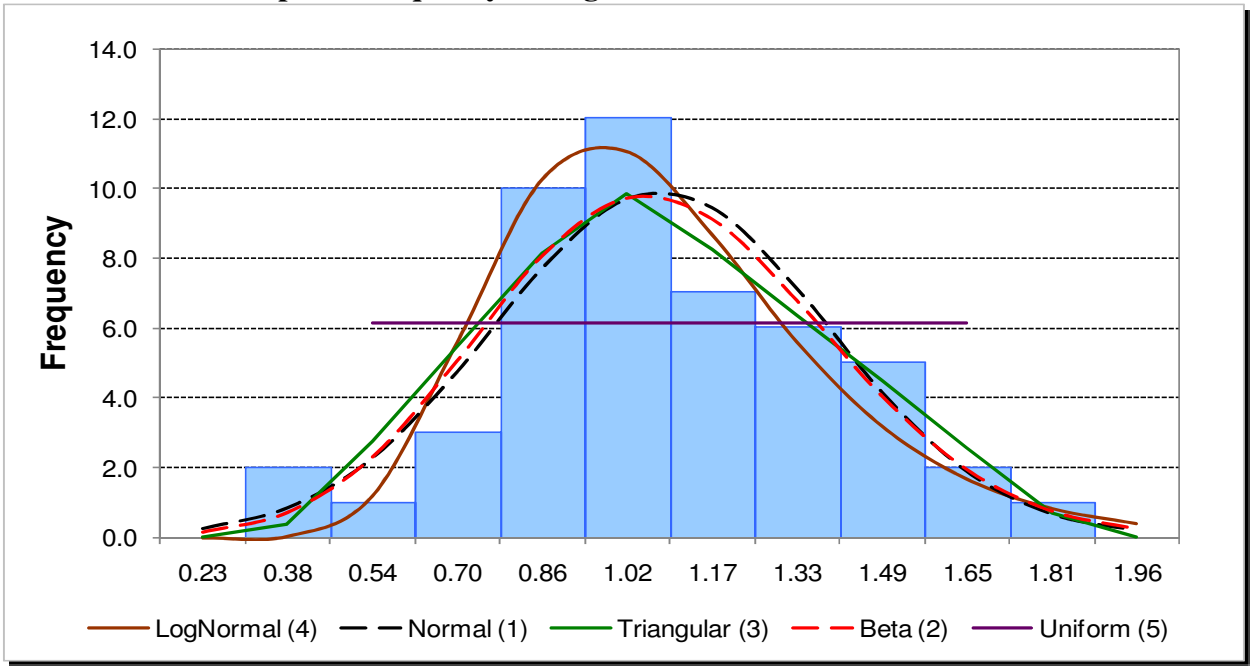
Shift = Net Factor – 1 = 1.084125 – 1 = 0.084125

Based upon Equation 26, if we analyze the percent errors in ratios ($y_i / \hat{y}_i$) with the adjustments of Net Factor and Shift, i.e., ($y_i / \hat{y}_i$) *(1.084125) - 0.084125, the results derived by Distribution Finder for the "adjusted" percent errors are given by the following:

Table 1: Distribution Finder Results – 49 Adjusted % Errors

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0035	1.0000	1.0000	1.0002	1.0000
StdDev	0.3125	0.3009	0.3093	0.3049	0.3078	0.2957
CV	0.3125	0.2998	0.3093	0.3049	0.3078	0.2957
Min	0.2255			0.3005	-0.6144	0.4878
Mode		0.8819	1.0000	0.9130	0.9734	
Max	1.8066			1.7866	3.8257	1.5122
Alpha					17.1439	
Beta					30.0000	
Data Count	49	% < 0 =	0.06%	None	0.01%	None
Standard Error of Estimate		0.0679	0.0504	0.0584	0.0518	0.0926
Rank		4	1	3	2	5
SEE / Fit Mean		6.76%	5.04%	5.84%	5.18%	9.26%
Chi^2 Fit test 9 Bins, Sig 0.05		Good (43%)	Good (32%)	Good (31%)	Good (18%)	Poor (0%)

Graph 1: Frequency Histogram – 49 Normalized % Errors



As shown by Table 1, the normal distribution is ranked #1 with an estimated standard deviation of 0.3093. This standard error is almost the same as the one that is used in CO$TAT to provide the PI report. The rest of the fitted distributions all estimate a smaller standard deviation than 0.3093.

CONCERNS ABOUT ANALYZING DIFFERENT CER ERRORS ALL TOGETHER

A common practice is to pool all the residuals (or percentage errors) from various CERs to analyze them together using a distribution finding tool. However, there are concerns about using this approach to analyze all the residuals (or percent errors) from different CERs. In fact, it is not appropriate to do so; the reasons are given below:

- The CER errors from different CERs might not be identically distributed. Using USCM CERs as an example, the distribution of errors from the Structure CER may not be the same as the distribution of errors from the Electrical Power Subsystem (EPS) CER. The analysis results will be misleading and inaccurate if we combine two or more samples coming from different populations and analyze them all together.
- The CER errors associated with different subsystems might not be independently distributed either. For example, the errors from the Structure CER may be correlated with the errors from the EPS CER at a certain correlation coefficient. The errors from the EPS CER may be correlated with the errors from the Telemetry, Tracking and Command (TT&C) subsystem CER at another correlation coefficient. At a minimum, we should determine whether or not these CER errors are correlated before pooling them together.
- We cannot use this approach to specify PIs for cost uncertainty analysis. For example, how do we define the location factor using the distribution fitting tool results when analyzing the "normalized" errors for all subsystems?

However, several analysts strongly believe that the CER errors from different subsystems may all follow log-normal distributions and they should be analyzed together using a distribution finding tool. We will explain why this is not an appropriate approach even when all CER errors follow log-normal distributions. We will begin with a shifted log-normal distribution.

Shifted Log-Normal Distribution. If X is distributed as a log-normal distribution with a mean of μ and variance σ^2 in log space, i.e., $X \sim LN(\mu,\sigma^2)$, then the following linear transformation on X, i.e., $aX + b$, is said to follow a shifted log-normal distribution:

$$Y = aX + b \sim LN(\mu+\ln(a), \sigma^2, b), \text{ where b is a location parameter.} \quad (31)$$

Using the MUPE and Minimum-Percentage Error under Zero-Percentage Bias (ZMPE) models, the multiplicative error term (ε_i) is assumed to have a mean of one and variance σ_u^2 (in unit space) for all observations. If ε_i is further assumed to follow a log-normal distribution, i.e., $\varepsilon_i \sim LN(\mu, \sigma^2)$, where μ and σ are the mean and standard deviation in log space, then the mean and standard deviation of ε_i are expressed as follows:

$$E(\varepsilon_i) = e^{(\mu+\sigma^2/2)} = 1 \qquad \text{for i = 1, …, n} \quad (32)$$

$$stdev(\varepsilon_i) = e^{(\mu+\sigma^2/2)}\sqrt{e^{\sigma^2}-1} = \sqrt{e^{\sigma^2}-1} = \sigma_u \qquad \text{for i = 1, …, n} \quad (33)$$

Note that the unit-space standard deviation of the error term (i.e., σ_u) can be estimated by the standard percent error (SPE) of a MUPE CER:

$$SPE = \sqrt{\sum_{i=1}^{n}((y_i - \hat{y}_i)/\hat{y}_i)^2/(n-p)} \quad (34)$$

where $\hat{y}_i$ is used to denote the predicted value in unit space for the i[th] data point and p is the total number of estimated coefficients as defined above. The MUPE CER provides consistent estimates of the parameters and has zero proportional error for all points in the data set.

It follows from Equations 32 and 33 that the log-space mean (μ) and standard deviation (σ) can be derived below:

$$\mu = -\sigma^2/2 = -(\ln(1+\sigma_u^2))/2 \quad (35)$$

$$\sigma = \sqrt{\ln(1+\sigma_u^2)} \quad (36)$$

Consequently, the mean and standard deviation of ($\varepsilon_i - 1$) are given as follows, respectively:

$$E(\varepsilon_i - 1) = E(\varepsilon_i) - 1 = 0 \qquad \text{for i = 1, …, n} \quad (37)$$

$$stdev(\varepsilon_i - 1) = stdev(\varepsilon_i) = \sqrt{e^{\sigma^2} - 1} = \sigma_u \qquad \text{for i = 1, …, n} \quad (38)$$

Given ε_i ~LN(μ, σ^2), the "theoretical" percentage error (i.e., $\varepsilon_i - 1$) now has a **shifted** log-normal distribution:

$$\frac{Y_i - f_i}{f_i} = \varepsilon_i - 1 \sim \text{LN}(\mu, \sigma^2, -1), \qquad \text{for i = 1, …, n.}$$

where "-1" is a location parameter. And the normalized percentage error is distributed as

$$\frac{1}{\sigma_u}\left(\frac{Y_i}{f_i} - 1\right) = \frac{\varepsilon_i - 1}{\sigma_u} \sim \text{LN}(\mu - \ln(\sigma_u), \sigma^2, -1/\sigma_u) \qquad \text{for i = 1, …, n} \quad (39)$$

where "$-1/\sigma_u$" is a location parameter.

Plugging Equations 35 and 36 into Equation 39, the normalized percentage error is distributed as

$$\frac{1}{\sigma_u}\left(\frac{Y_i}{f_i} - 1\right) = \frac{\varepsilon_i}{\sigma_u} - \frac{1}{\sigma_u} \sim \text{LN}(-(\ln(1+\sigma_u^2))/2 - \ln(\sigma_u), \ln(1+\sigma_u^2), -1/\sigma_u) \text{ for i = 1, …, n} \quad (40)$$

(Note: σ_u can be estimated by the SPE of a MUPE CER.)

We can easily verify that the mean of the normalized percentage error (Equation 40) is zero and its standard deviation is one.

We now make the following assumptions:

Let ε_{i1} for i = 1,…, n_1 denote the CER errors for subsystem 1

Let ε_{i2} for i = 1,…, n_2 denote the CER errors for subsystem 2

…

Let ε_{ik} for i = 1,…, n_k denote the CER errors for subsystem k

The normalized errors for each subsystem should have the following distribution:

$$\frac{\varepsilon_{i1}-1}{\sigma_{u1}} \sim LN(\mu_1 - \ln(\sigma_{u1}), \sigma_1^2, -1/\sigma_{u1}) \qquad \text{for i = 1, …, } n_1$$

$$\frac{\varepsilon_{i2}-1}{\sigma_{u2}} \sim LN(\mu_2 - \ln(\sigma_{u2}), \sigma_2^2, -1/\sigma_{u2}) \qquad \text{for i = 1, …, } n_2$$

…

$$\frac{\varepsilon_{ik}-1}{\sigma_{uk}} \sim LN(\mu_k - \ln(\sigma_{uk}), \sigma_k^2, -1/\sigma_{uk}) \qquad \text{for i = 1, …, } n_k$$

The above expressions can be summarized as follows:

$$\frac{\varepsilon_{ij}-1}{\sigma_{uj}} \sim LN(\mu_j - \ln(\sigma_{uj}), \sigma_j^2, -1/\sigma_{uj}) \qquad \text{for i = 1, …, } n_j \text{ and j = 1,…, k.} \quad (41)$$

Note that (σ_{uj}) are the unit-space standard deviation for subsystem j (j = 1,…, k), which can be estimated by the CER's SPE.

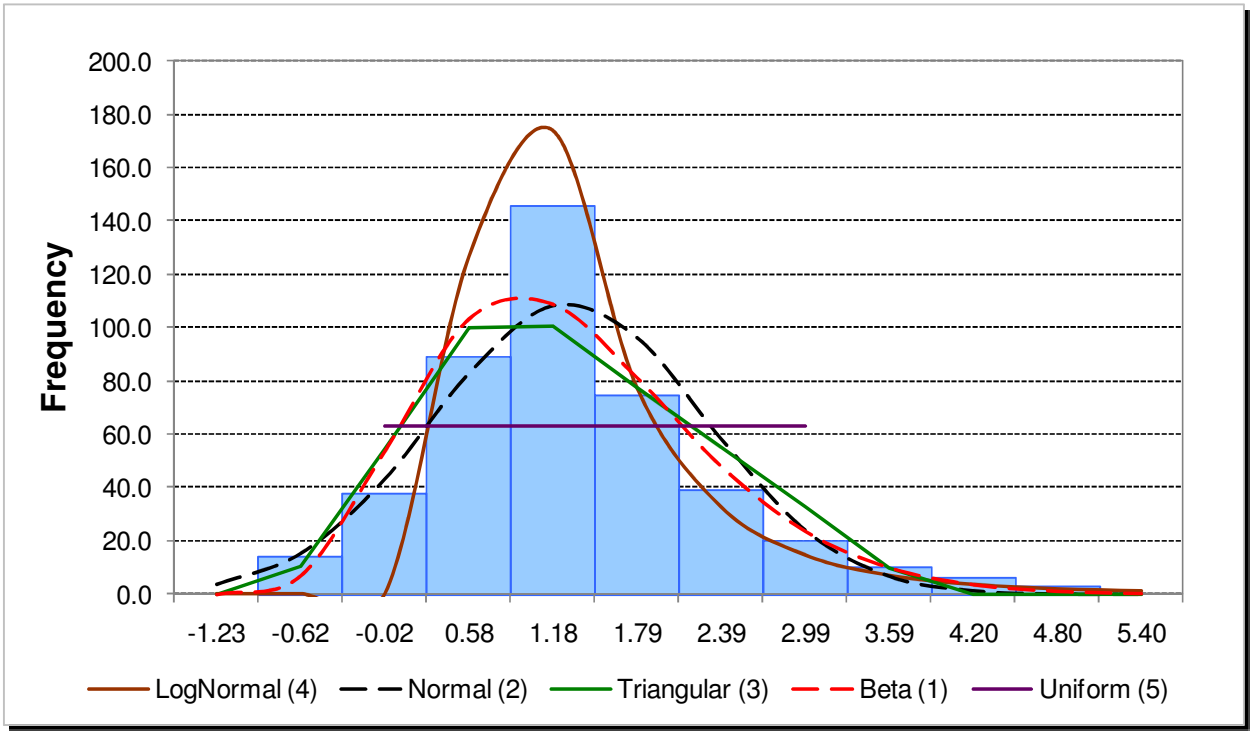
Based upon the above reasoning, we conclude the following: although these normalized percentage errors for different subsystems all have a mean of zero and variance of one in **unit** space, they do not have the same mean and variance in **log** space. Their location parameters are also different for different subsystems. This is why they should not be analyzed together using a distribution finding tool even if they all come from log-normal distributions.

Listed below are the Distribution Finder results using all 440 normalized percent errors for the USCM9 subsystem-level CERs, assuming all these errors are independent and following the same distribution. Here, one is added to the normalized data to avoid centering on zero.

Table 2: Distribution Finder Results – 440 Normalized % Errors + 1

	Sample	LN	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0801	1.0000	1.0000	1.0003	1.0000
StdDev	0.9746	0.8385	0.9565	0.9527	0.9674	0.9127
CV	0.9746	0.7763	0.9565	0.9527	0.9671	0.9127
Min	-1.2272			-1.0200	-1.3128	-0.5808
Mode		0.5323	1.0000	0.4655	0.6295	
Max	4.7993			3.5544	15.3924	2.5808
Alpha					4.7874	
Beta					29.7871	
Data Count	440	**% < 0 =**	14.79%	15.31%	14.45%	18.37%
Std Error of Estimate		0.3231	0.1888	0.2016	0.1206	0.3396
Rank		4	2	3	1	5
SEE / Fit Mean		29.92%	18.88%	20.16%	12.06%	33.96%
Chi^2 Fit test 22 Bins,	**Sig 0.05**	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)

Graph 2: Frequency Histogram – 440 Normalized % Errors + 1



As shown by the Distribution Finder results, the **beta** distribution fits the frequency histogram better than the other four distributions, although none of them pass the Chi-square test.

However, "one" may not be a good location parameter for these normalized percent errors. If Solver is used to fit a shifted log-normal distribution to these normalized data points, the location parameter is estimated to be "-3.8231." Listed below are the Distribution Finder results when subtracting this location parameter from these 440 normalized percent errors. As shown by Table 3, the log-normal distribution is now ranked number one, as it fits the frequency histogram better than the other four distributions. Still, none of them pass the Chi-square test.

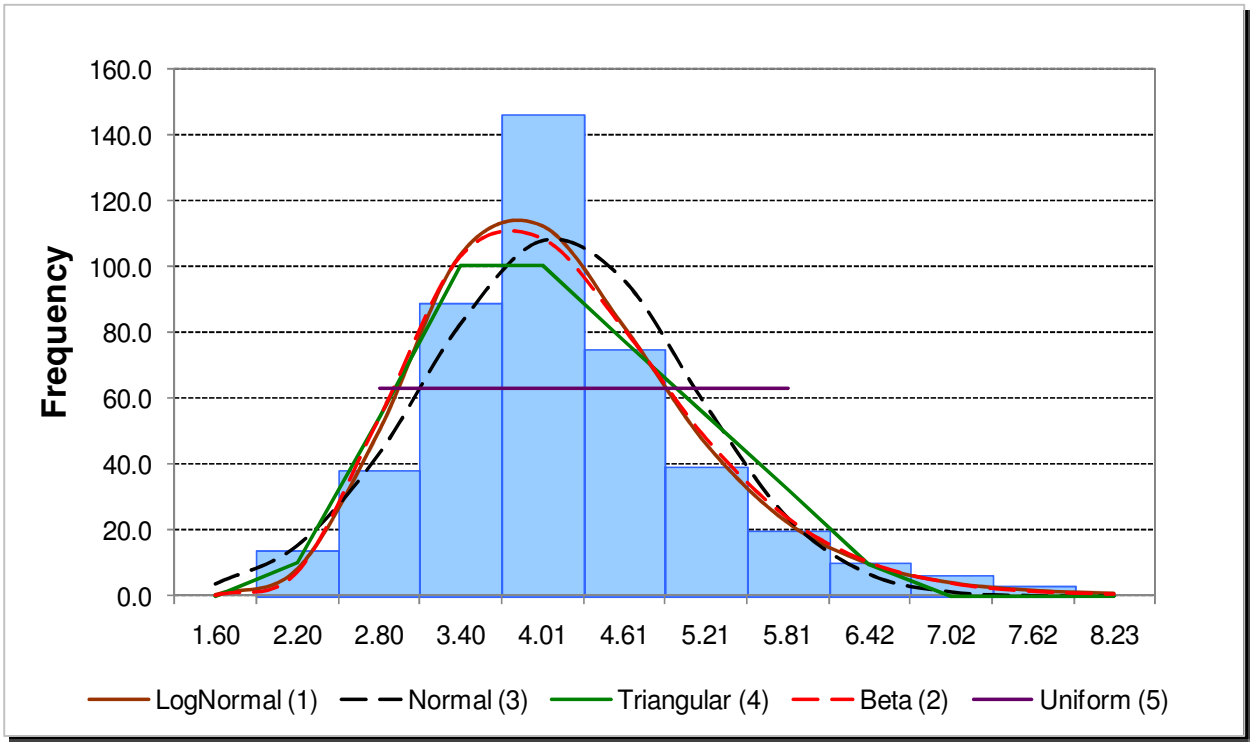
Note that this is just an exercise to demonstrate that a shifted log-normal distribution is more flexible and useful than the ordinary log-normal distribution for modeling. Different shifted log-normal distributions will be derived when fitting the CER errors by individual subsystems. Note also that this particular log-normal distribution has a standard deviation of 0.25 in log space, which is smaller than the smallest SPE of all the eight subsystem CERs under investigation. The fitted results are very doubtful—this example indicates the pitfalls of analyzing all CER errors together.

Table 3: Distribution Finder Results – 440 Normalized % Errors + 3.8231

	Sample	LN	Normal	Triangular	Beta	Uniform
Mean	3.8231	3.8235	3.8231	3.8231	3.8234	3.8231
StdDev	0.9746	0.9709	0.9565	0.9527	0.9674	0.9127
CV	0.2549	0.2539	0.2502	0.2492	0.2530	0.2387
Min	1.5959			1.8031	1.5126	2.2423
Mode		3.4814	3.8231	3.2886	3.4520	

Max	7.6224			6.3775	18.2559	5.4039
Alpha					4.7803	
Beta					29.8555	
Data Count	440	% < 0 =	0.00%	None	None	None
Std Error of Estimate		0.1011	0.1888	0.2016	0.1206	0.3396
Rank		1	3	4	2	5
SEE / Fit Mean		2.64%	4.94%	5.27%	3.15%	8.88%
Chi^2 Fit test 22 Bins,	**Sig 0.05**	Poor (3%)	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)

Graph 3: Frequency Histogram – 440 Normalized % Errors + 3.8231



ANALYSIS OF USCM9 SUBSYSTEM-LEVEL CERS

We used Distribution Finder to model the error distributions for the USCM9 CERs at the subsystem level. Listed below are the Distribution Finder results for analyzing the CER errors in the form of ratios ($y_i / \hat{y}_i$) for the Attitude Control Subsystem (ACS), EPS, Propulsion, Structure, and TT&C subsystem-level CERs. No specific locations are considered in the analysis, as it is a generalized assessment. Due to the large sample size, adjustment factors are not applied to these examples either.

Table 4: Distribution Finder Results – ACS CER % Errors ($y_i / \hat{y}_i$)

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0039	1.0000	1.0001	1.0008	1.0000
StdDev	0.3776	0.3732	0.3722	0.3698	0.3761	0.3562
CV	0.3776	0.3718	0.3722	0.3697	0.3758	0.3562
Min	0.1490			0.2359	0.0684	0.3830
Mode		0.8268	1.0000	0.7637	0.8636	
Max	2.0583			2.0006	6.4772	1.6170
Alpha					5.1081	
Beta					29.9987	
Data Count	56	% < 0 =	0.36%	None	None	None
Standard Error of Estimate		0.0521	0.0696	0.0645	0.0463	0.1171
Rank		2	4	3	1	5
SEE / Fit Mean		5.19%	6.96%	6.45%	4.63%	11.71%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (74%)	Good (17%)	Good (41%)	Good (41%)	Poor (4%)

Graph 4: Frequency Histogram – ACS CER % Errors ($y_i / \hat{y}_i$)

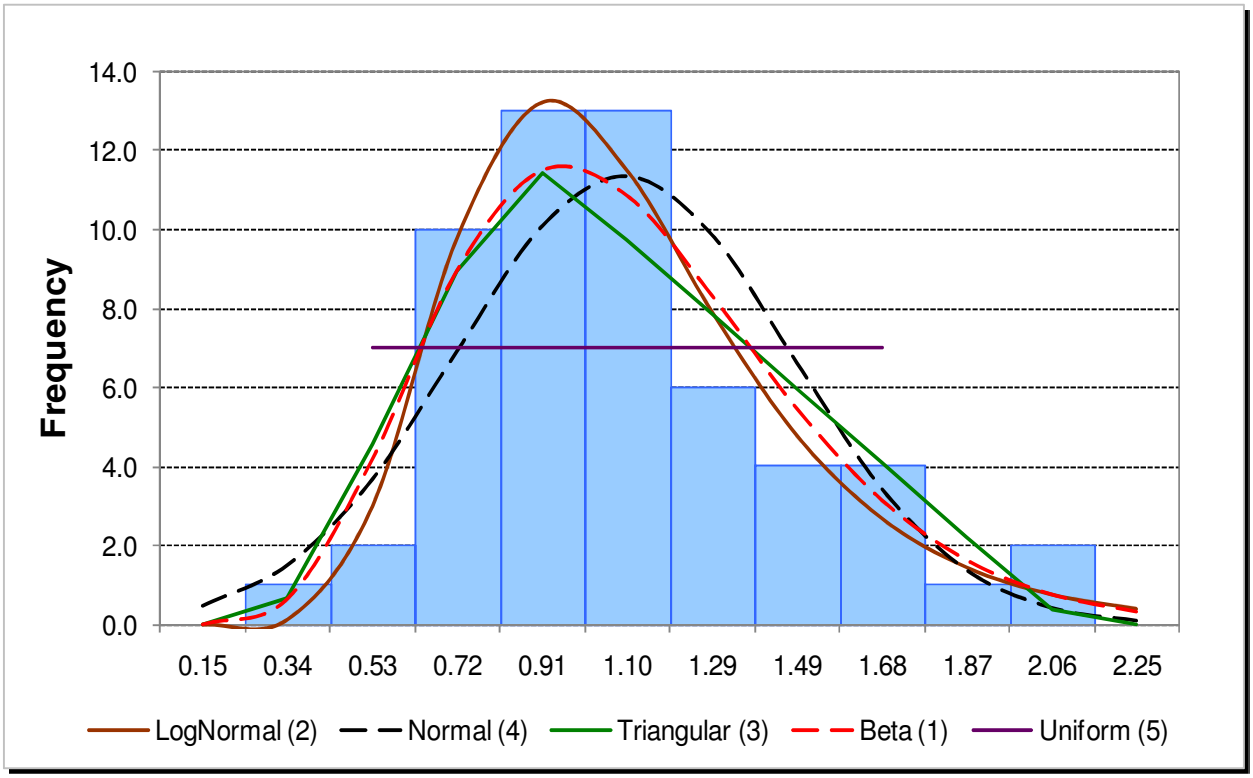


Table 5: Distribution Finder Results – EPS CER % Errors ($y_i / \hat{y}_i$)

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0037	1.0000	1.0001	1.0013	1.0000
StdDev	0.4438	0.4458	0.4308	0.4297	0.4427	0.4097
CV	0.4438	0.4441	0.4308	0.4296	0.4421	0.4097
Min	0.2315			0.1556	0.2236	0.2904
Mode		0.7662	1.0000	0.6654	0.7501	
Max	2.5675			2.1792	9.5042	1.7096
Alpha					2.7440	
Beta					30.0000	
Data Count	62	% < 0 =	1.01%	None	None	None
Standard Error of Estimate		0.0489	0.1111	0.1016	0.0578	0.1638
Rank		1	4	3	2	5
SEE / Fit Mean		4.87%	11.11%	10.16%	5.77%	16.38%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (33%)	Good (17%)	Good (18%)	Good (16%)	Poor (2%)

Graph 5: Frequency Histogram – EPS CER % Errors ($y_i / \hat{y}_i$)

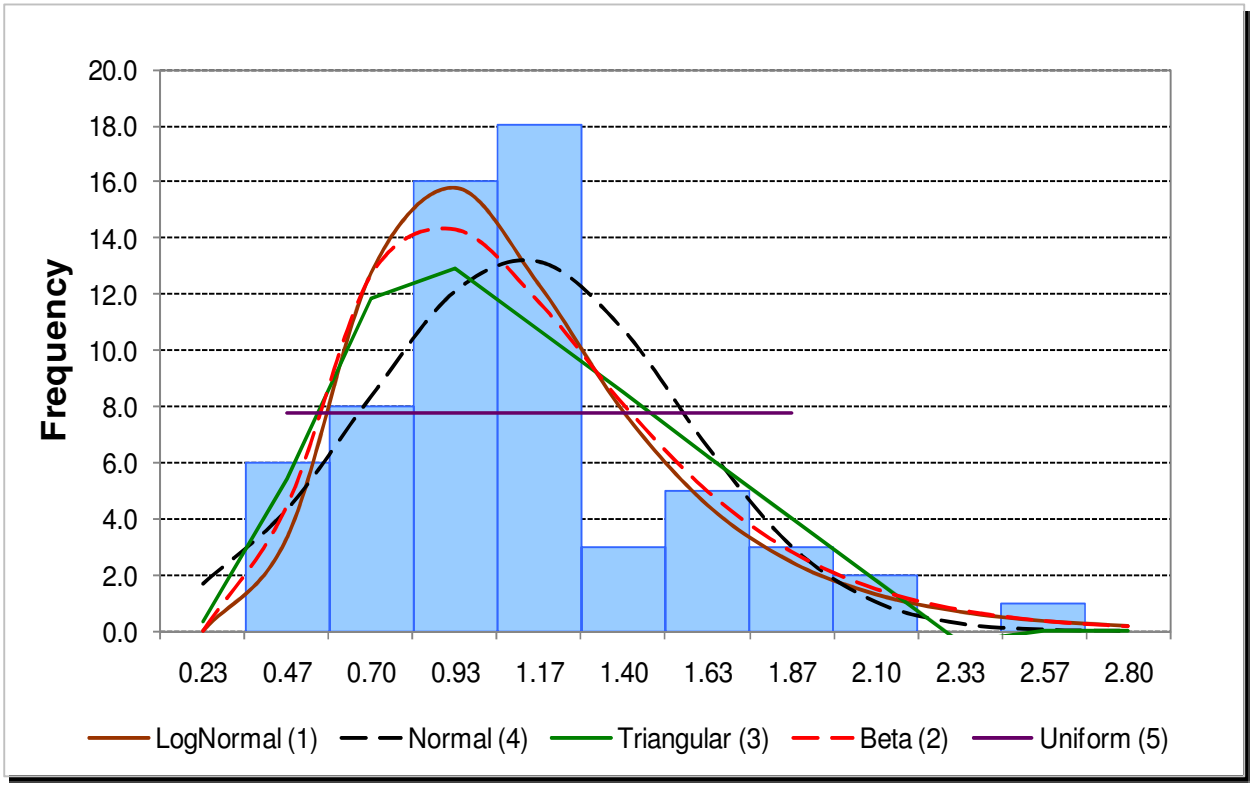


Table 6: Distribution Finder Results – Propulsion CER % Errors ($y_i / \hat{y}_i$)

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0038	1.0000	1.0000	1.0004	1.0000
StdDev	0.3620	0.3550	0.3570	0.3523	0.3578	0.3384
CV	0.3620	0.3536	0.3570	0.3523	0.3576	0.3384
Min	0.2047			0.2185	-0.3405	0.4139
Mode		0.8412	1.0000	0.8556	0.9343	
Max	2.0452			1.9261	4.7226	1.5861
Alpha					10.0616	
Beta					27.9286	
Data Count	54	% < 0 =	0.25%	None	0.01%	None
Standard Error of Estimate		0.0624	0.0657	0.0735	0.0584	0.1212
Rank		2	3	4	1	5
SEE / Fit Mean		6.22%	6.57%	7.35%	5.84%	12.12%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (84%)	Good (28%)	Good (20%)	Good (11%)	Good (9%)

Graph 6: Frequency Histogram – Propulsion CER % Errors ($y_i / \hat{y}_i$)

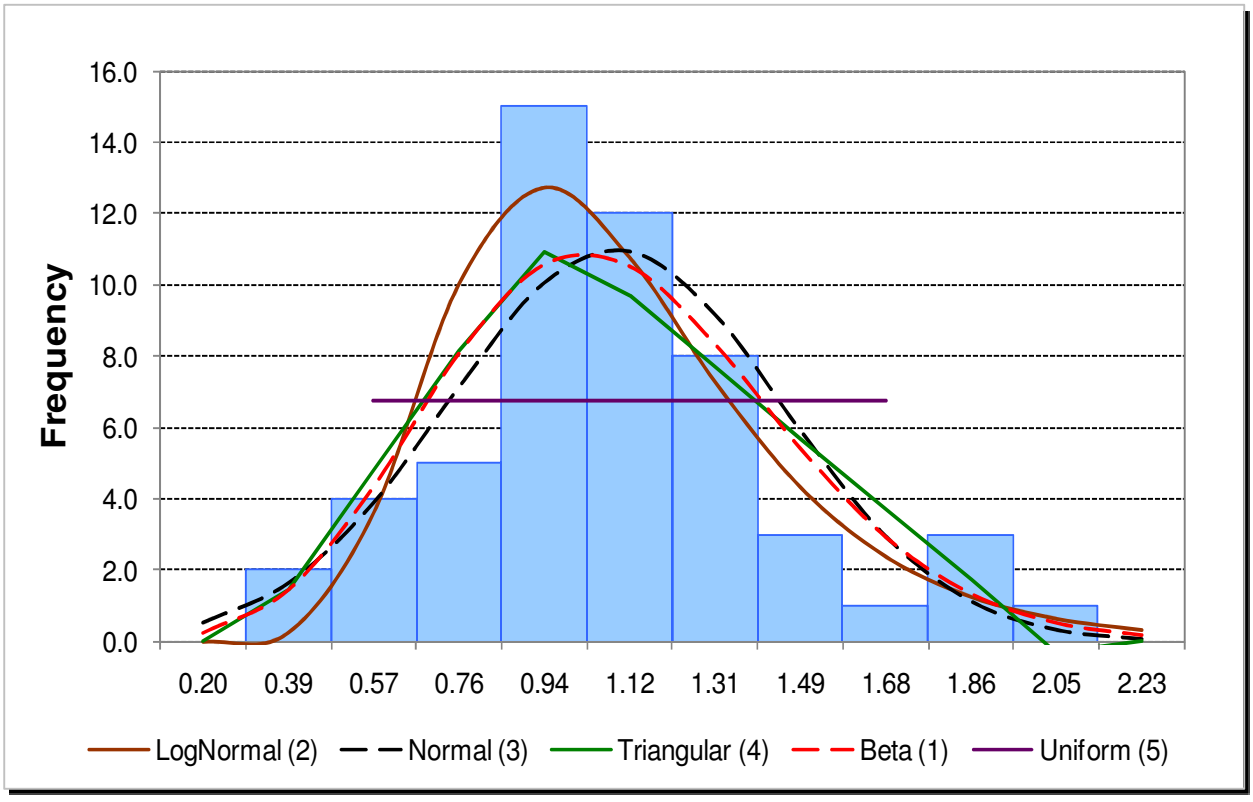


Table 7: Distribution Finder Results – Structure CER % Errors ($y_i / \hat{y}_i$)

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0041	1.0000	1.0000	1.0003	1.0000
StdDev	0.3614	0.3515	0.3561	0.3508	0.3562	0.3378
CV	0.3614	0.3500	0.3561	0.3508	0.3561	0.3378
Min	0.2384			0.1989	-0.6592	0.4148
Mode		0.8443	1.0000	0.8937	0.9577	
Max	2.1318			1.9076	4.4553	1.5852
Alpha					14.3374	
Beta					29.8497	
Data Count	53	% < 0 =	0.25%	None	0.05%	None
Standard Error of Estimate		0.0724	0.0675	0.0778	0.0652	0.1209
Rank		3	2	4	1	5
SEE / Fit Mean		7.21%	6.75%	7.78%	6.51%	12.09%
Chi^2 Fit test 9 Bins, Sig 0.05		Good (60%)	Good (17%)	Good (28%)	Poor (5%)	Poor (2%)

Graph 7: Frequency Histogram – Structure CER % Errors ($y_i / \hat{y}_i$)

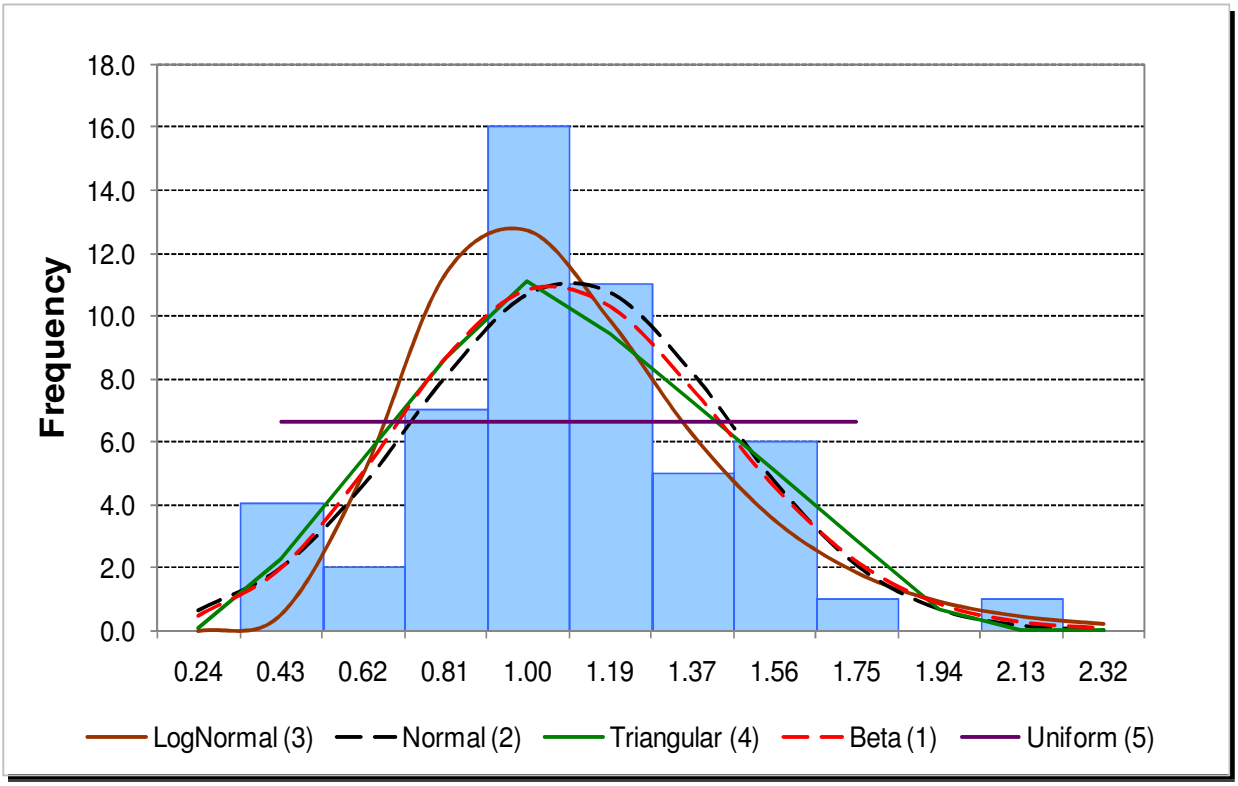
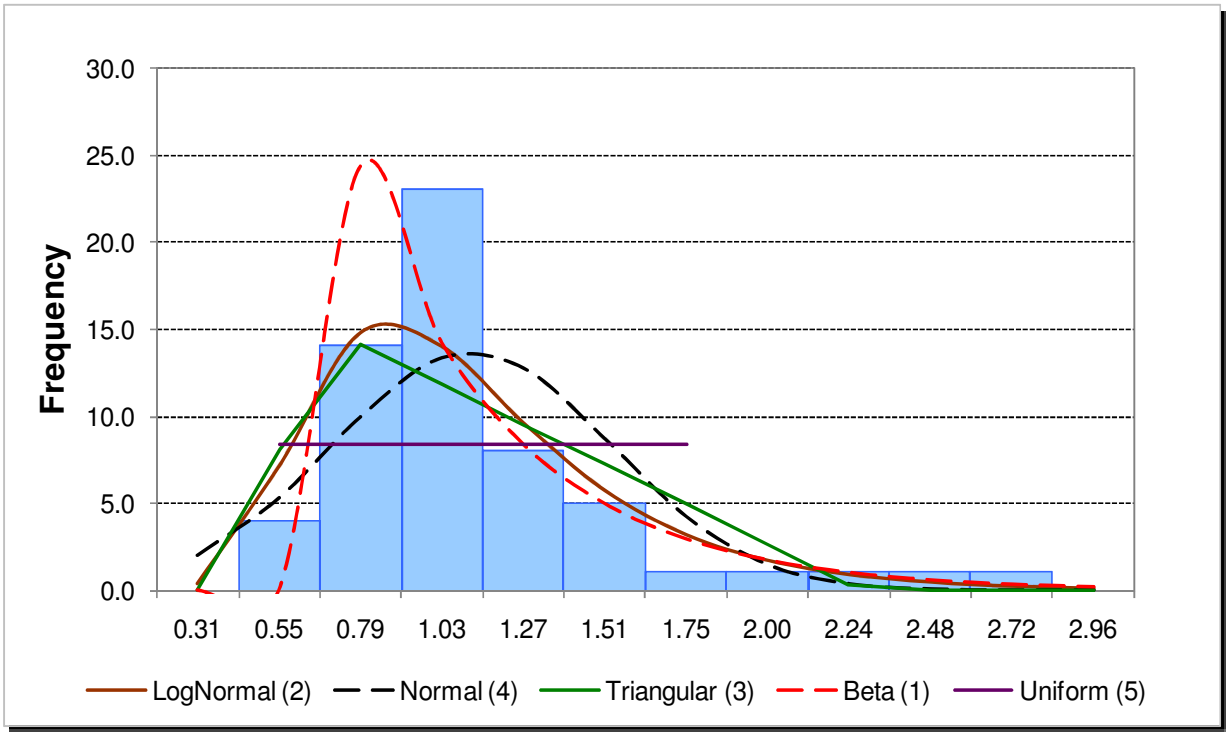


Table 8: Distribution Finder Results – TT&C CER % Errors ($y_i / \hat{y}_i$)

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	0.9964	1.0000	1.0001	1.0023	1.0000
StdDev	0.4446	0.4597	0.4083	0.4116	0.4455	0.3796
CV	0.4446	0.4614	0.4083	0.4115	0.4444	0.3796
Min	0.3086			0.3027	0.5595	0.3425
Mode		0.7460	1.0000	0.5416		
Max	2.7192			2.1561	15.3091	1.6575
Alpha					0.9286	
Beta					30.0000	
Data Count	59	% < 0 =	0.72%	None	None	None
Standard Error of Estimate		0.0856	0.1798	0.1641	0.0833	0.2282
Rank		2	4	3	1	5
SEE / Fit Mean		8.59%	17.98%	16.41%	8.31%	22.82%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (23%)	Poor (4%)	Poor (1%)	Good (10%)	Poor (0%)

Graph 8: Frequency Histogram – TT&C CER % Errors ($y_i / \hat{y}_i$)



In the above examples, we only used the "raw" percent errors (i.e., $y_i / \hat{y}_i$) in Distribution Finder without applying any correction factors. As shown by these tables, the best fitted distribution varies from one subsystem to another and the beta distribution seems to be a popular candidate to model the CER uncertainties. Since the sample and DF corrections are not used in these examples, the estimated standard deviation generated by Distribution Finder is smaller than

the SPE generated by its respective USCM9 CER. We should also use Distribution Finder to analyze the USCM9 suite and component-level CERs to determine whether the beta distribution is still very common for modeling CER errors—this will be an interesting follow-on study item.

CONCLUSIONS

Sample size can be a concern when using a distribution fitting tool. When the sample size is small, say 25 or less, the fitted distribution most likely does not resemble the frequency histogram of the sample data. Since we are usually dealing with small samples in our business, sample size is a concern when using any distribution fitting tool.

To find an appropriate distribution to model CER errors, we should fit (1) residuals for additive error models, (2) percent errors in the form of ratios (i.e., $y_i / \hat{y}_i$) **for MUPE and ZMPE CERs, (3) residuals in log space for log-error models, and (4) ratios of actual to the mean** (i.e., $y_i / \bar{y}$) **for univariate analysis.** If we have no information on how the CER was generated, we can use the following rules to deduce the underlying hypothesis:

- If the sum of the residuals equals zero ($\Sigma (y_i - \hat{y}_i) = 0$), then the CER is probably an OLS.
- If the average of the percent errors equals one ($\Sigma (y_i / \hat{y}_i) / n = 1$), then it may be a MUPE or ZMPE CER.
- If the sum of the residuals in log space equals zero ($\Sigma (\ln(y_i) - \ln(\hat{y}_i)) = 0$), then the CER may be a log-linear model.

However, when an analyst applied the PING factor (or a Smearing Estimate) to the log-linear CER, the average percent errors may be one, but the CER was not generated by the MUPE or ZMPE method.

Consider three adjustment factors when fitting the residuals or percent errors using a distribution fitting tool for cost uncertainty analysis: sample, DF, and location factors. Be sure to apply these adjustment factors to CER errors before fitting them using a distribution finding tool. This way, the low/high range, mean, mode, and standard deviation of the fitted distribution are adjusted accordingly by the distribution finding tool, so analysts can use the fitted results directly for cost uncertainty analysis. It is much easier and more straightforward to make the adjustments before running the distribution finding tool, than it is to adjust several statistics for the PI after running the tool. (Note: Depending upon the CER type, apply the net factor, which is the product of these three factors, to residuals, percent errors, or "residuals in log space" accordingly, prior to fitting them using a distribution fitting tool.)

Do not apply the DF factor when the sample size is fairly large (e.g., DF > 50) or when a Student's t or a Log-t distribution is used to model the CER errors. We should multiply the *Adj. SE* measure by the DF factor to account for the broader tails of Student's t distribution for small samples if we use normal instead of t distribution for cost uncertainty analysis. We should do the same if we use log-normal instead of Log-t distribution for cost uncertainty analysis. Therefore, it is not necessary to apply the DF factor if either Student's t or Log-t distribution is used to model CER errors.

Define a shift factor for MUPE CERs, so the CER errors are centered on one. Do not apply a shift factor for additive or log-error models.

Do not pool all the residuals (or percentage errors) from various CERs to analyze them together using a distribution finding tool. The errors from different CERs might not be

identically distributed and they may also be correlated. Furthermore, we cannot define a meaningful location factor for a shifted log-normal distribution when analyzing pooled errors from different CERs.

RECOMMENDATIONS AND FUTURE STUDY ITEMS

Enrich Distribution Gallery. For small sample analysis, it is very important to apply the DF factor to a CER's standard error to account for the broader tails of Student's t and Log-t distributions when using normal or log-normal distribution to model the CER errors. The shifted log-normal distribution, as well as the Weibull and gamma distributions, can also be found in real life examples. Therefore, a distribution fitting tool should consider including the following distribution in its Distribution Gallery:

> Student's t, Log-t, Weibull, Rayleigh (a special case of Weibull), Shifted Log-Normal, Gamma, Poisson, Extreme Value distribution, and User-Defined Cumulative Distribution function (CDF)

At a minimum, the first six distributions should be considered for inclusion.

Adjust DF for Additional Constraints. It is noted that the lower bound of the "fitted" distribution is sometimes negative, as generated by the curve-fitting process. Also, the upper bound of the fitted distribution can be smaller than the largest data point in the data set, which is neither logical nor desirable. Of course, we can specify certain constraints for the fitted distribution. For example, we can constrain the lower bound of the fitted distribution to be positive or non-negative, but this restriction should be reflected in the DF calculation.

However, an inequality constraint may not be the same as an equality constraint. For example, if the lower bound of the distribution is fixed to be zero, then one less parameter will be estimated by the distribution fitting tool, which translates to a gain of one DF. Another example: we can constrain the mean of the fitted distribution to be the sample mean and then the distribution fitting tool has one less parameter to search, which may be viewed as a gain of one DF. On the other hand, we are using the sample mean to estimate the population mean, which amounts to a loss of one DF. So the DF may stay the same for this case. Furthermore, we should take redundancy into account when counting the DF. The DF adjustment regarding constraints when fitting distributions is a potential topic for future study.

Consider applying User-Defined CDF to model sample data with two or multiple modes. A distribution finding tool is designed to locate a distribution from its distribution gallery to best represent the sample data. However, if a sample data set clearly exhibits more than one mode in its frequency histogram, a distribution finding tool is not likely to find an appropriate distribution to model this data set. An example of this would be if one mode occurs at the lower end and the other mode occurs in the middle of the data set. In this situation, using a User-Defined CDF to model the sample data is suggested.

Additional research for Beta and Log-Normal distributions: can the "world" be described by Beta and Log-Normal? Based upon our empirical studies using the USCM9 subsystem-level and suite-level CERs, the beta and log-normal distributions often fit the frequency histogram reasonably well (i.e., often ranked number 1 or number 2 in the report). Should we go ahead and use these two distributions to model CER error distributions for cost uncertainty analysis?

A beta distribution is commonly defined by four parameters: the lower and upper bounds of the distribution (L and H) and two shape parameters (α and β). This distribution is largely used in management science to model continuous probability distribution over a finite interval. Many analysts prefer using beta distribution in describing cost uncertainty because it is finite, continuous, and allows virtually any degree of skewness and kurtosis. Also, it tends to have more density around the mode than the triangular distribution, which is considered to be a plus. Besides the shape parameters, the low/high bounds of the beta distribution are also estimated by the distribution finding tool. Since the boundaries can have a substantial impact on the beta distribution, we should explore more realistic examples to determine whether beta distribution is more probable than Student's t and Log-t distributions to model CER errors.

Analyze CER error distributions for the USCM9 component-level CERs. This study analyzes the error distribution for several USCM9 subsystem-level CERs. We should also use a distribution fitting tool to analyze the error distribution for the USCM9 component-level CERs. This will be a useful follow-on study because the lower level CERs have many more data points, so we can apply a distribution fitting tool objectively to model the error distributions for USCM9 CERs.

REFERENCES

1. Smith, A., "Build Your Own Distribution Finder," 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010.
2. Nguyen, P., B. Kwok, et al., "Unmanned Spacecraft Cost Model, Ninth Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, August 2010.
3. Hu, S., "The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development," 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001.
4. Book, S. A. and N. Y. Lao, "Minimum-Percentage-Error Regression under Zero-Bias Constraints," Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 Oct 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.
5. Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons, 1989, pages 37, 46, 86-88.

APPENDIX A – DATA SET A

Observations	Cost	Weight
1	4,197.72	378.80
2	344.83	85.00
3	12,019.06	1,442.19
4	11,228.35	1,477.25
5	11,846.05	1,415.99
6	11,279.14	1,354.73
7	10,222.10	1,348.83
8	1,603.85	226.51
9	2,525.09	1,030.78
10	767.68	151.42
11	2,049.64	302.42
12	3,002.22	361.08
13	2,107.59	286.73
14	2,166.00	285.50
15	1,583.90	277.74
16	6,553.17	838.00
17	5,309.96	811.36
18	2,804.87	266.83
19	3,037.80	296.02
20	2,425.72	360.64
21	404.40	156.33
22	1,176.51	151.37
23	1,849.26	285.90
24	3,935.52	371.49
25	1,582.63	286.16
26	4,488.67	427.01
27	2,559.96	448.33
28	839.17	403.61
29	4,244.22	212.58
30	6,268.54	679.30
31	3,028.53	525.88
32	2,567.24	511.42
33	3,622.59	397.73
34	2,861.84	464.84
35	2,566.93	414.43
36	948.44	163.66
37	835.06	134.14
38	4,512.10	293.84
39	699.87	99.32
40	7,348.69	1,101.28
41	4,373.23	608.67
42	3,711.84	428.92
43	4,299.03	806.50
44	4,259.07	451.01
45	4,331.53	393.79
46	6,201.98	976.66
47	747.25	15.13

APPENDIX B – DATA SET B

Observations	Cost	Weight
1	3,793.99	611.38
2	10,676.77	2,327.69
3	10,524.52	2,285.41
4	9,095.45	2,177.02
5	1,511.39	365.58
6	2,322.08	1,663.67
7	775.56	244.39
8	1,903.68	488.10
9	2,741.95	582.79
10	2,006.08	460.80
11	1,493.83	448.28
12	5,866.79	1,352.53
13	4,772.77	1,309.53
14	2,568.28	430.67
15	2,773.27	477.77
16	2,234.63	582.08
17	1,135.33	244.31
18	1,727.35	461.44
19	3,563.26	599.59
20	30,984.36	15,389.49
21	1,492.72	461.86
22	10,073.28	1,457.76
23	4,050.03	689.19
24	5,616.32	1,096.39
25	2,359.17	825.43
26	3,287.88	641.93
27	2,618.42	750.25
28	2,358.90	668.90
29	6,287.16	1,843.79
30	5,510.83	1,250.26
31	3,572.06	1,053.85
32	3,010.41	1,053.85
33	7,165.43	1,780.29
34	6,475.55	1,841.10
35	4,000.64	963.93
36	4,786.95	1,250.26
37	4,693.68	852.71
38	3,027.72	871.23
39	934.63	264.15
40	715.88	160.30
41	6,566.85	1,777.46
42	30,980.07	8,709.75
43	3,948.45	982.39
44	3,366.42	692.27
45	3,883.15	1,301.69
46	1,899.63	1,687.26
47	3,847.98	727.93
48	3,911.74	635.58
49	5,557.74	1,576.33

APPENDIX C – MUPE/ZMPE FACTOR PI = UNIVARIATE PI

Given the following factor equation with a multiplicative error term:

$$Y_i = \beta * X_i * \varepsilon_i \qquad \text{for } i = 1, \ldots, n \tag{42}$$

where:

β = the factor (to be estimated by the regression equation)
n = sample size
Y_i = the i^{th} observation of the dependent variable ($i = 1, \ldots, n$)
X_i = the i^{th} data point of the independent variable ($i = 1, \ldots, n$)
ε_i = the error term (with a mean of 1 and a variance σ^2)

Both the MUPE and ZMPE methods derive the same solution for Equation 42. It can be shown within two iterations that the MUPE solution (denoted by b) for the factor β is a simple average of the Y to X ratios:

$$b = \left(\sum_{i=1}^{n} \frac{Y_i}{X_i} \right) / n = \overline{Z} \tag{43}$$

where n is the sample size and Z denotes the Y to X ratio.

This is also the solution using the ZMPE method and it is even more straightforward to derive the factor β of Equation 42 using the ZMPE method due to the following constraint:

$$\sum_{i=1}^{n} \frac{Y_i - \beta X_i}{\beta X_i} = \left(\sum_{i=1}^{n} \frac{Y_i}{\beta X_i} \right) - n = 0$$

Based upon Equation 16 above, a (1-α)100% PI for a future observation Y, when X is at x_o is given below:

$$PI = bx_0 \pm t_{(\alpha/2, n-1)} * SE * \sqrt{(1 + \frac{1}{n}) b^2 x_0^2}$$

$$= bx_0 \left(1 \pm t_{(\alpha/2, n-1)} * SE * \sqrt{1 + \frac{1}{n}} \right) = bx_0 (1 \pm t_{(\alpha/2, n-1)} * Adj.\ SE) \tag{44}$$

We now prove that the SE in the above equation is equal to the coefficient of variation (CV) of the Y to X ratio:

$$SE = S_z / \overline{Z} = CV(Z) = CV(Y/X) \tag{45}$$

where:

$Z = Y/X$

S_z is the sample standard deviation of Z

$\overline{Z}$ is the sample mean of Z

Equation 45 can be verified by the definition of the standard error of regression equation and Equation 43:

$$
\begin{aligned}
(SE)^2 &= \frac{1}{n-1}\sum_{i=1}^{n} w_i (y_i - bx_i)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \frac{1}{(bx_i)^2}(y_i^2 - 2bx_i y_i + (bx_i)^2) \\
&= \frac{1}{n-1}\left(\frac{1}{b^2}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - \frac{2}{b}\sum_{i=1}^{n}\frac{y_i}{x_i} + n\right) \\
&= \frac{1}{n-1}\left(\frac{1}{b^2}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - \frac{2}{b}(nb) + n\right) \\
&= \frac{1}{b^2(n-1)}\left(\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - nb^2\right) = \frac{1}{b^2}\frac{1}{(n-1)}\left(\sum_{i=1}^{n} z_i^2 - n\bar{z}^2\right) \\
&= \frac{(S_z)^2}{\bar{z}^2} = (CV(Z))^2
\end{aligned}
\tag{46}
$$

where w_i is the weighting factor for the i^{th} data point (i = 1,…, n). Note that for the MUPE and ZMPE CERs, the weighting factor for the i^{th} data point is the square of the reciprocal of its predicted value; namely, $w_i = (1/bx_i)^2$ for i = 1,…, n.

As shown by Equation 46, the SE measure is the same for both the univariate analysis and MUPE/ZMPE factor CER, except that the ratios are used to compute the CV for the MUPE/ZMPE factor CER. Therefore, these two PIs are the same, since the SE, location factor, and t ratio are all the same between them.