What’s the Point?
Discussion on How CER Point Estimates Should Be Interpreted in Lognormal Distributions

Betsy Turnbull
Tom Parkey
Glenn Research Center
August 26, 2015
Agenda

• Overview
• Survey of Current Guidance
• An Alternative Viewpoint
• Summary
Overview

• When deriving CERs for use in cost estimation, a number of techniques are employed, including:
  – Ordinary Least Squares (OLS)
  – Log Transformed OLS (LOLS)
  – Minimum Unbiased Percentage Error (MUPE)
  – Zero Bias Minimum Percent Error (ZMPE)

• After deriving these CERs we apply uncertainty to them, frequently in the form of lognormal distributions, for use in a Monte Carlo simulation (or method of moments)

• The question becomes, where on this uncertainty distribution do we place the CER generated point estimate?
A Significant Issue

While deciding the point on the distribution to use isn’t all that important when error terms are relatively small, it can be critical in a real world application.

Two Estimates of the Same Project

CER result as:
- Mode
- Mean

Vehicle Cost FY15$M

Percentile

$0  $100  $200  $300  $400  $500  $600  $700
Current Guidance

• Various handbooks do briefly address this matter

• We will look at:
  – Expert Opinion
“Depending on the situation, a CER result may represent the mean, median or mode of the CER uncertainty distribution. Therefore, CER results should be anchored to the point in the distribution consistent with how the uncertainty for the CER was defined. In all cases, all uncertainty distributions should be truncated at zero.”

“In the interest of simplifying the cost risk analysis process, the following approach is recommended:

- Regardless of the parametric CER form or regression method used to create it, the uncertainty of the CER may be modeled with a lognormal distribution.
- In the absence of better information, the result of the CER will be treated as the median (50% value).
- The dispersion of the lognormal distribution will be defined by the CER standard error adjusted for sample size and the position the estimate falls within the dataset used to derive the CER”

- CER result can be mean, median, or mode depending on the situation
- When in doubt they recommend lognormal distribution with the point estimate taken as the median
### Table 2-3 Recommended Subjective Uncertainty Distributions

<table>
<thead>
<tr>
<th>Shape</th>
<th>Typical Applications</th>
<th>CER Result</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Linear or non-linear CERs with additive error, mechanical tolerances. All MUPE generated CERS. Univariate methods</td>
<td>Mean, median, mode</td>
<td>Equal probability of overrun or underrun</td>
</tr>
<tr>
<td>Lognormal</td>
<td>Log-linear CERs that transform to linear in log space ( y = a \times e^{b} ) Labor rates, labor rate adjustments, factor methods</td>
<td>Median (^8)</td>
<td>The uncertain variable can increase without limits, but can not fall below zero, is positively skewed, with most of the values near the mode</td>
</tr>
<tr>
<td>Triangular</td>
<td>Engineering data or analogy estimates (throughputs), labor rates, labor rate adjustments, factor methods</td>
<td>Mode</td>
<td>Popular because they are easy to understand and communicate - use when likelihood decreases with distance from PE</td>
</tr>
<tr>
<td>Uniform</td>
<td>Engineering data or analogy estimates (throughputs), Labor rates, labor rate adjustments, factor methods</td>
<td>Unknown</td>
<td>Used when every value across the range of the distribution has an equal likelihood of occurrence</td>
</tr>
<tr>
<td>Beta</td>
<td>Engineering data or analogy estimates (throughputs)</td>
<td>Mode</td>
<td>Complicated to explain and to apply consistently across different tools</td>
</tr>
<tr>
<td>Weibull</td>
<td>Objective relationship to reliability modeling.</td>
<td>Mode</td>
<td>Popular because of the wide variety of shapes that can be defined, including the Raleigh and Exponential distribution</td>
</tr>
</tbody>
</table>

\(^8\) This is recommended as the default point estimate interpretation only because OLS appears to be the most common method used within the community to generate CERs. Those using more sophisticated methods (e.g. MUPE, ZMPE) will recommend the appropriate distribution shape and define how the CER result (the point estimate) is interpreted consistent with their method.
“which is a single estimate, but only one point on a lognormal distribution. What point on the distribution does this represent? Depending on the method used, this may represent a measure at or near the ‘center’ of the distribution, such as the mean or the median”

• CER result said to be some measure of centrality dependent on CER development method
• No description of which methods yield which point estimate locations
“MUPE: The MUPE CER delivers the mean; it has zero proportional error for all points in the CER. Goodness-of-fit measures can be derived to judge the quality of the model if the CER error is assumed to be normal (a common assumption).”

“ZMPE: The ZMPE method also delivers the mean and zero proportional error for all the data points in the CER. Distribution shape is arbitrary; however, some analysts prefer using lognormal.”

“Two critical decisions: Select the uncertainty shape and define where the point estimate falls.”

- Explicitly acknowledges the importance of selecting uncertainty shape and point estimate location
- States that these methods deliver the mean and zero proportional error for all points
- Uncertainty distribution shape is said the arbitrary for ZMPE (preference being lognormal)
# Table 2-2 Recommended Uncertainty Distributions

<table>
<thead>
<tr>
<th>DISTRIBUTION</th>
<th>TYPICAL APPLICATION</th>
<th>KNOWLEDGE OF MODE</th>
<th>NUMBER OF PARAMETERS REQUIRED</th>
<th>RECOMMENDED PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>Default when no better info. Probability skewed right. Replicate another model result. Power OLS CER uncertainty. Mean or median known better than the mode</td>
<td>2</td>
<td>Median, high (some tools have a 3rd parameter: “Location”. By default, it is zero. Used to “shift” the lognormal left or right (even into negative region)</td>
<td></td>
</tr>
<tr>
<td>Log-t</td>
<td>Log-t when &lt; 30 data points</td>
<td>3</td>
<td>Add Degrees of Freedom</td>
<td></td>
</tr>
<tr>
<td>Triangular</td>
<td>Expert opinion. Finite min/max. Probability reduces towards endpoints. Skew possible. Labor rates, labor rate adjustments, factor methods Good idea</td>
<td>3</td>
<td>Low, mode, and high</td>
<td></td>
</tr>
<tr>
<td>BetaPert</td>
<td>Like triangular, but mode is 4 times more important than min or max. Very good idea</td>
<td>3</td>
<td>Low, mode, and high</td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>Like triangular, but min/max region known better than mode. Not sure</td>
<td>4</td>
<td>Min, low, high, and max</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Equal chance low/high. Unbounded in either direction Linear OLS CER uncertainty. Good idea, but unbounded in either direction</td>
<td>2</td>
<td>Mean/Median/Mode and high value</td>
<td></td>
</tr>
<tr>
<td>Student’s-t</td>
<td>t when &lt; 30 data points</td>
<td>3</td>
<td>Add Degrees of Freedom</td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>Equal chance over uncertainty range. Finite min/max. No idea</td>
<td>2</td>
<td>Low and High (some tools require min. and max)</td>
<td></td>
</tr>
<tr>
<td>Empirical Fit</td>
<td>Unable to fit a distribution to the data Not required</td>
<td>N/A</td>
<td>Enter source data and estimated probability for each data point</td>
<td></td>
</tr>
</tbody>
</table>

Note: Low/high are defined with an associated percentile. Min/Max are the absolute lower/upper bound (also known as the 0/100)
Point Estimate Locations in Regard to Skew

Joint Agency Cost Schedule Risk and Uncertainty Handbook

*Figure 2-8* Illustration of Distribution Skew
Expert Opinions

• Dr. Shu-Ping Hu- “You can apply a log-normal distribution to a MUPE or ZMPE CER for cost uncertainty analysis. Distribution assumption is not required when using these two methods. (Just like OLS, the normality assumption is applied for the purpose of statistical inferences when deriving MUPE CERs.) Since there is no sample proportional bias for the MUPE/ZMPE CERs, use “mean” as the PE interpretation.” (from email correspondence)

• Timothy Anderson- “Since you are using ZMPE, then I would state (without proof) that the result of the CER is the MEAN of the distribution. To my knowledge, nobody has proved this, but my logic tells me, since we construct the ZMPE CER in a way that the BIAS is zero, that the result is the MEAN. Why? Because the sample mean can be shown to be an unbiased estimator of the population mean for any distribution. Therefore, since we force the ZMPE CER to produce an unbiased estimate, then it follows that the estimate must be the mean.” (from email correspondence)
All That Being Said...

- Although somewhat daunted, especially by the intellectual weight of the two expert opinions, we would like to make an argument for using an alternative measure of central tendency: the mode.

We will now attempt to defend this seemingly tenuous position with some basic observations.
Effect of SPE on Confidence Level

- If the CER result is assumed to be the mean of a lognormal, the confidence level of that result INCREASES when the error increases.
- The opposite occurs if the mode is assumed.

Mean = 100:

<table>
<thead>
<tr>
<th>Percent Error</th>
<th>Percentile of Mean</th>
<th>Mode</th>
<th>Percentile of Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>52.0%</td>
<td>98.5</td>
<td>46.0%</td>
</tr>
<tr>
<td>20%</td>
<td>53.9%</td>
<td>94.3</td>
<td>42.2%</td>
</tr>
<tr>
<td>30%</td>
<td>55.8%</td>
<td>87.9</td>
<td>38.5%</td>
</tr>
<tr>
<td>40%</td>
<td>60.9%</td>
<td>63.1</td>
<td>29.0%</td>
</tr>
<tr>
<td>50%</td>
<td>59.3%</td>
<td>71.6</td>
<td>31.8%</td>
</tr>
<tr>
<td>60%</td>
<td>60.9%</td>
<td>63.1</td>
<td>29.0%</td>
</tr>
<tr>
<td>70%</td>
<td>62.4%</td>
<td>55.0</td>
<td>26.4%</td>
</tr>
<tr>
<td>80%</td>
<td>63.7%</td>
<td>47.6</td>
<td>24.1%</td>
</tr>
<tr>
<td>90%</td>
<td>65.0%</td>
<td>41.1</td>
<td>22.1%</td>
</tr>
<tr>
<td>200%</td>
<td>73.7%</td>
<td>8.9</td>
<td>10.2%</td>
</tr>
<tr>
<td>300%</td>
<td>77.6%</td>
<td>3.2</td>
<td>6.5%</td>
</tr>
<tr>
<td>500%</td>
<td>81.7%</td>
<td>0.8</td>
<td>3.6%</td>
</tr>
</tbody>
</table>

“I can see your error bars using google earth”

Taken from VADLO.com
When the CER Result is Modeled as the Mode

- For most Monte Carlo software applications, the arithmetic mean along with the standard deviation are the parameters used to define the lognormal distribution function.
- The mean must then be calculated based on the CER result and the CER’s SPE

If $SPE = 0.1$, then $\mu = 1.015 \times \text{mode}$

If $SPE = 0.3$, then $\mu = 1.111 \times \text{mode}$

If $SPE = 0.5$, then $\mu = 1.250 \times \text{mode}$
Okay, so what’s the point?

- The arguments for using the point estimate as the mean of a lognormal distribution center around ZMPE/ MUPE CER creation, namely that there is no sample proportional bias for CERs created with these techniques.

- However, there are no underlying distributional assumptions in the ZMPE/MUPE processes (i.e.; the analyst can choose any reasonable probability distribution to encapsulate uncertainty); we carry over only a point estimate and error term.

- We posit, therefore, that this point estimate is not inherently tied to a specific measure of central tendency in the assigned distribution.

- Assigning your CER result to the mode allows the error term to directly affect the estimate.
Summary

• Location of the point estimate is a critical issue especially when error terms are significant (i.e., when developing parametric cost estimates)

• Assigning the point estimate to the mode allows the error terms to realistically affect the estimate

• More discussion and research is warranted with the objective of developing clear and consistent guidance

• For more discussion, contact:
  – thomas.j.parkey@nasa.gov
  – elizabeth.r.turnbull@nasa.gov
Backup Slides
Calculating Lognormal Mean From Mode and Standard Deviation

For a lognormal distribution, the mode = $\mu^4/(\sigma^2+\mu^2)^{1.5}$

Using Matlab and making 2 substitutions, the following solution for $\mu$ is obtained:

Let $a = \text{Mode}^2$

Let $b = \sigma^2$

$$\mu = (1/4*a+1/12*3^{(1/2)}*3*a^2+24*a*b+2*(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)*a*b^3)
+1/12*6^{(1/2)}*3*a^2+24*a*b-(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)*a*b^3+36/(3*a^2+24*a*b+2*(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)-96/(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)*a*b^3)
+1/2*a^2*3^{(1/2)}+72/(3*a^2+24*a*b+2*(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)-96/(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)*a*b^3)
+1/2)*b^2+3/(3*a^2+24*a*b+2*(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)-96/(108*b^4*a^2+12*(768*a^3*b^9+81*b^8*a^4)^{(1/2)}(1/3)*a*b^3)\)^0.5
The Distribution of Choice!

Lognormal Distribution

- Used widely in cost estimation
- Costs tend to overrun, rather than underrun
- Has beneficial properties that reflect cost actuals
  - Skewed to the right
  - Does not allow values less than 0
Various Other Distributions

- Normal
  - Allows for negative costs
  - Symmetric (unrealistic in cost estimation)

- Triangular
  - Can be symmetric
  - Has a definite upper bound

- Truncated Normal
  - Skewed to the left

- Paranormal
  - Of course there are non-standard distributions…

Taken from: A visual comparison of normal and paranormal distributions Matthew Freeman J Epidemiol Community Health 2006;60:6.
Effect of Assumed Distribution
Lognormal Distributions with Low Error
Lognormal Distributions with More Typical Error