



2015 NASA Cost Analysis Symposium

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



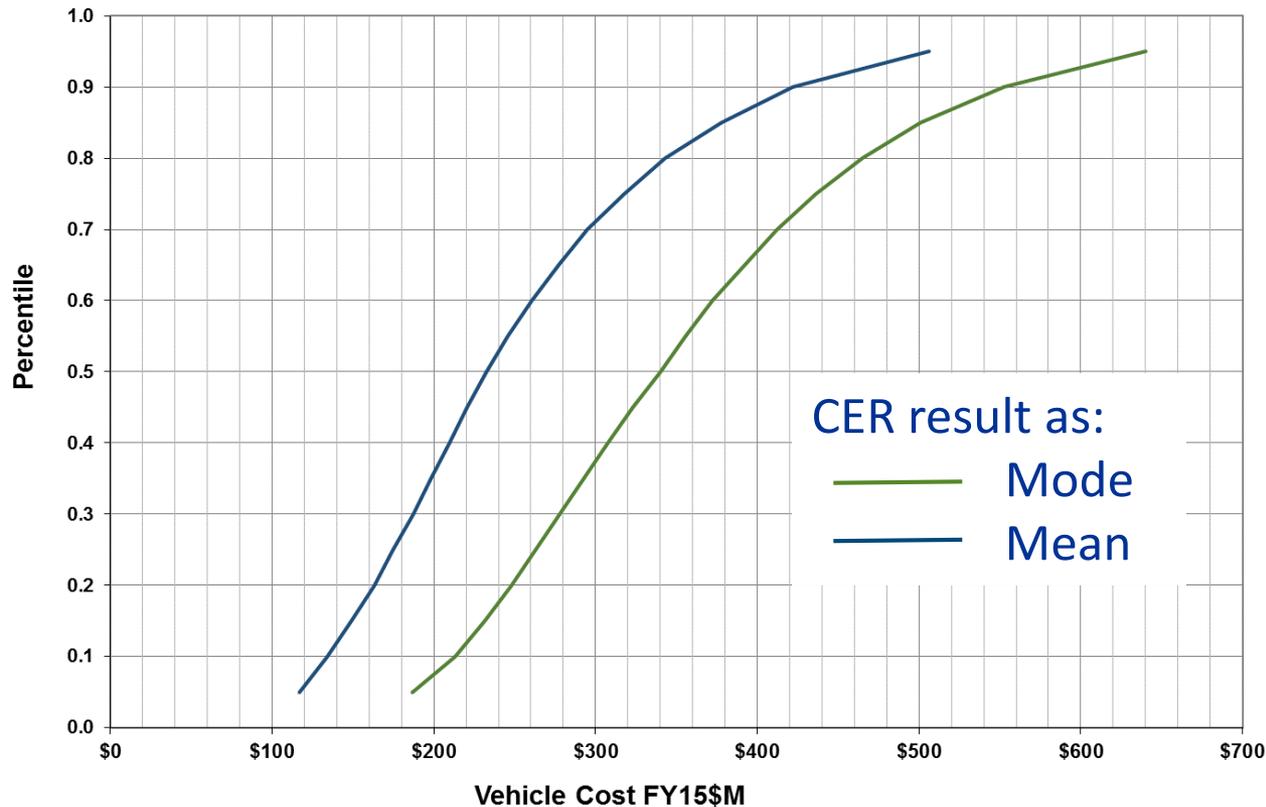
Overview

- 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

Cost and Economic Analysis Office



Current Guidance

- Various handbooks do briefly address this matter
- We will look at:
 - 2007 U.S. Air Force Cost Risk and Uncertainty Analysis Handbook (Air Force CRUAH)
 - 2012 Missile Defense Agency Cost Estimating and Analysis Handbook (MDA CEAH)
 - 2014 Joint Agency Cost Schedule Risk and Uncertainty Handbook (Joint Agency CSRUH)
 - Expert Opinion



2007 Air Force CRUAH

“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



2007 Air Force CRUAH

Table 2-3 Recommended Subjective Uncertainty Distributions

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

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.



2012 MDA CEAH

“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



2014 Joint Agency CSRUH

“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)



2014 Joint CSRUH

Table 2-2 Recommended Uncertainty Distributions

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

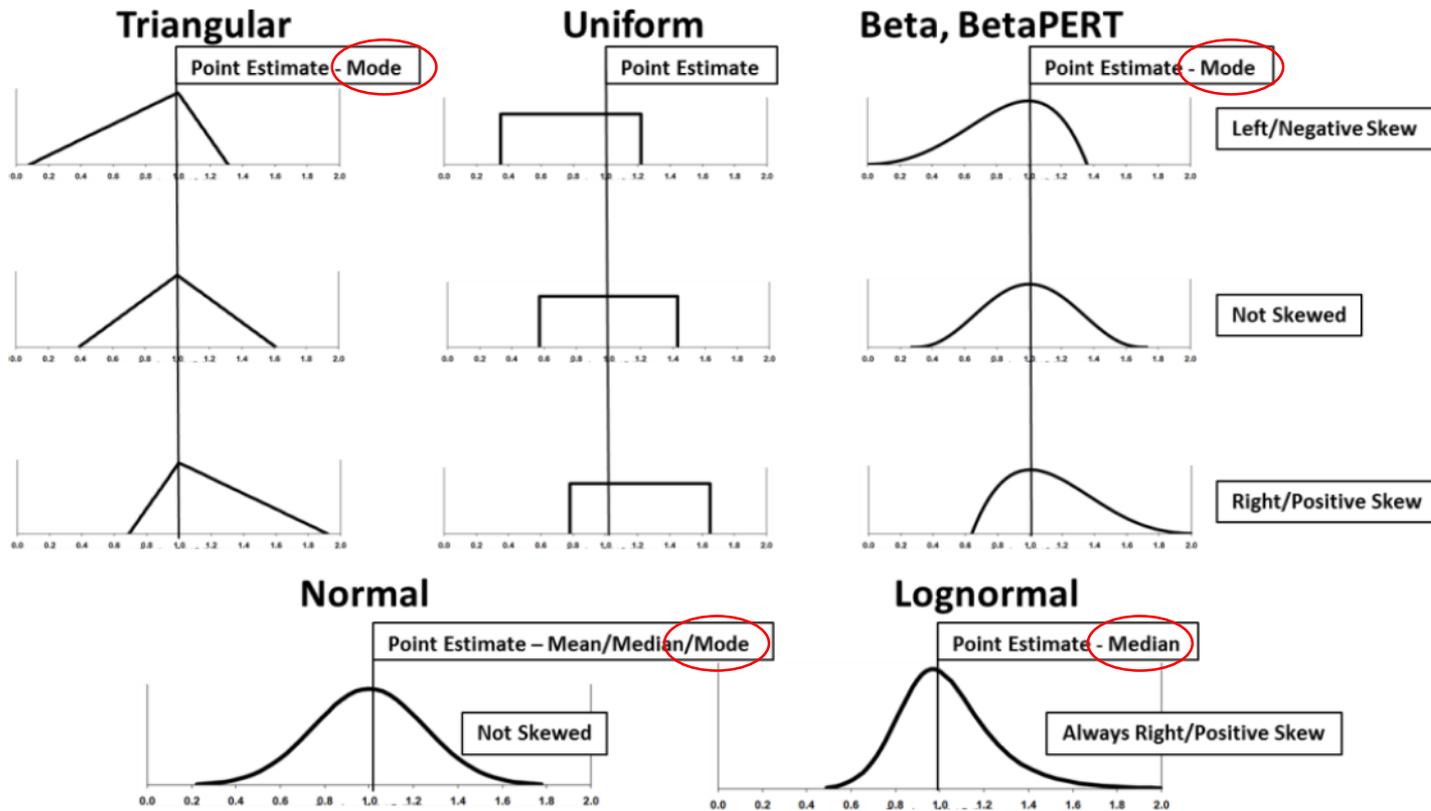


Figure 2-8 Illustration of Distribution Skew



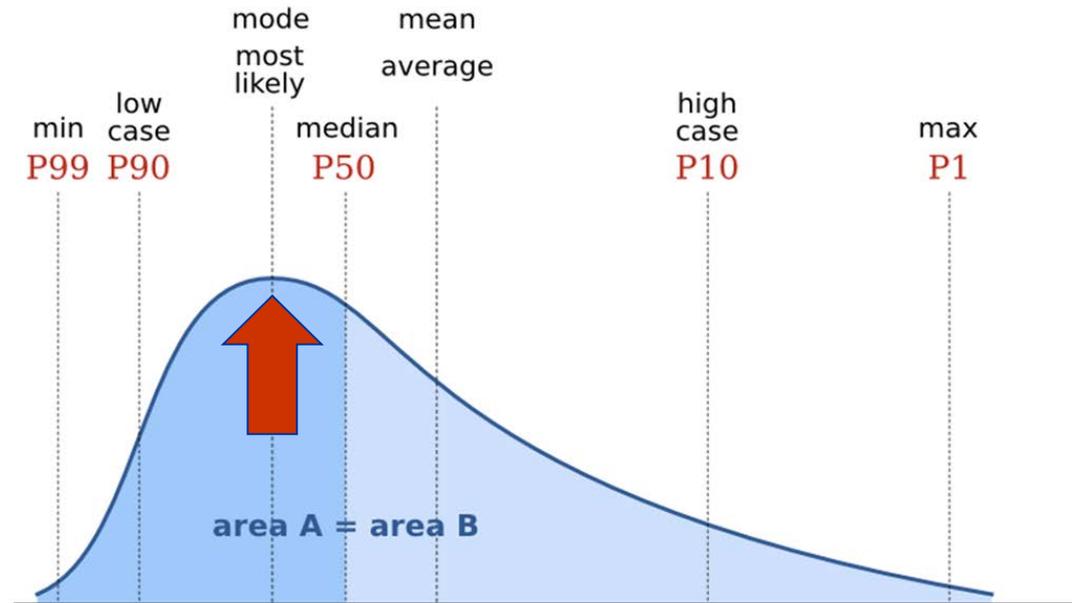
Expert Opinions

- **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:

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



“I can see your error bars using google earth”

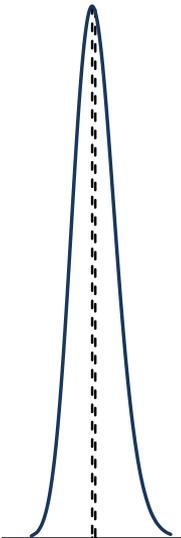
Taken from VADLO.com



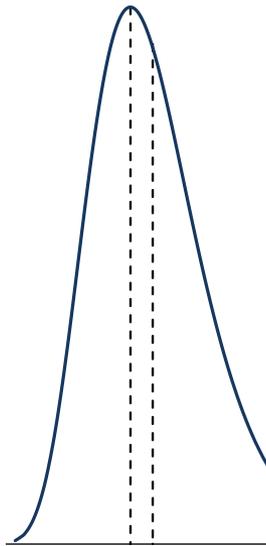
When the CER Result is Modeled as the Mode

- 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

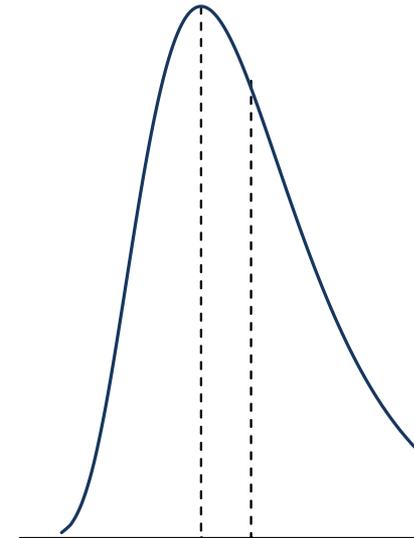
If SPE = 0.1,
 $\mu = 1.015 * \text{mode}$



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



If SPE = 0.5,
 $\mu = 1.250 * \text{mode}$





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



"It sort of makes you stop and think, doesn't it."



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 μ is obtained:

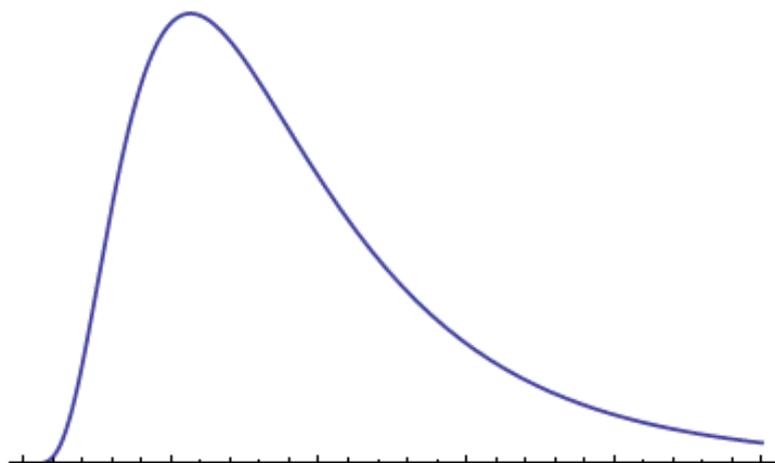
Let $a = \text{Mode}^2$

Let $b = \sigma^2$

$$\mu = \frac{(1/4*a + 1/12*3^{(1/2)}*(3*a^2 + 24*a*b + 2*(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)} - 96/(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)}*a*b^3)^{(1/2)} + 1/12*6^{(1/2)}*(3*a^2 + 24*a*b - (108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)} + 48/(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)}*a*b^3 + 36/(3*a^2 + 24*a*b + 2*(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)} - 96/(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)}*a*b^3)^{(1/2)}*a^2*3^{(1/2)}*b + 72/(3*a^2 + 24*a*b + 2*(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)} - 96/(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)}*a*b^3)^{(1/2)}*a^3^{(1/2)}*b^2 + 3/(3*a^2 + 24*a*b + 2*(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)} - 96/(108*b^4*a^2 + 12*(768*a^3*b^9 + 81*b^8*a^4)^{(1/2)})^{(1/3)}*a*b^3)^{(1/2)}*a^3*3^{(1/2)})^{(1/2)}}{0.5}$$



The Distribution of Choice!

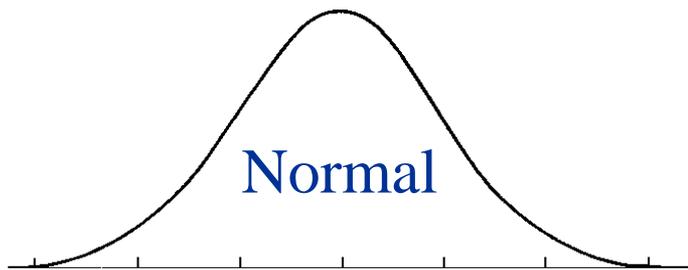


Lognormal Distribution

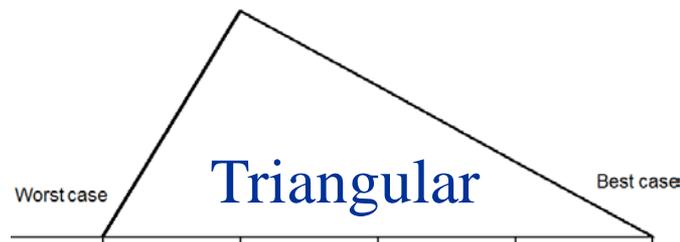
- Used widely in cost estimation
- Costs tend to overrun, rather than underrun
- Has beneficial properties that reflect cost actuals
 - ✓ Skewed to the right
 - ✓ Does not allow values less than 0



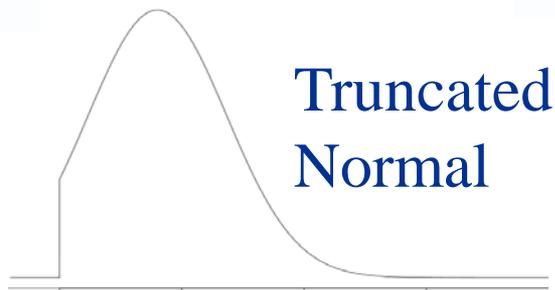
Various Other Distributions



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



- Can be symmetric
- Has a definite upper bound



- Skewed to the left

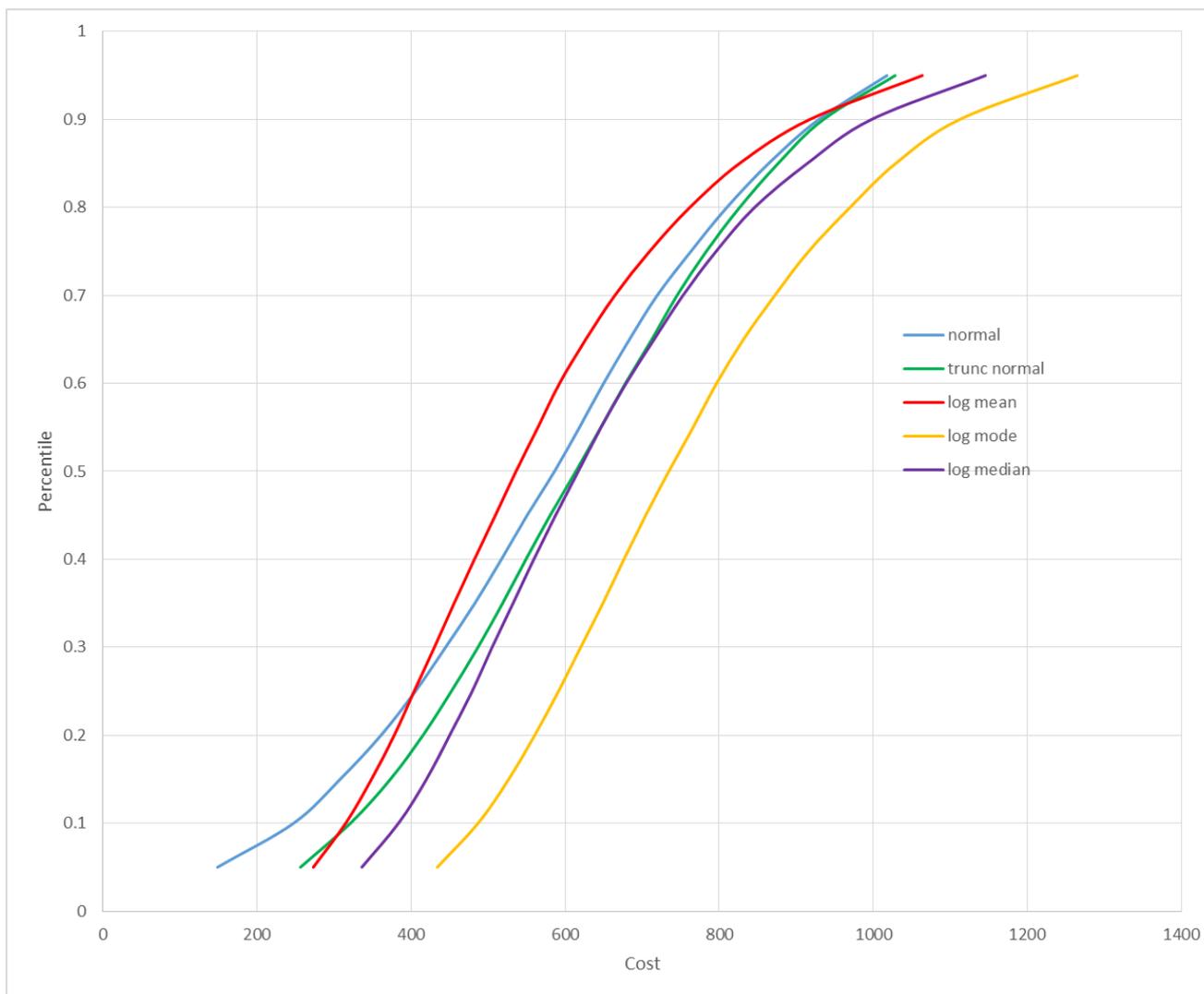
Paranormal



-Of course there are non-standard distributions...

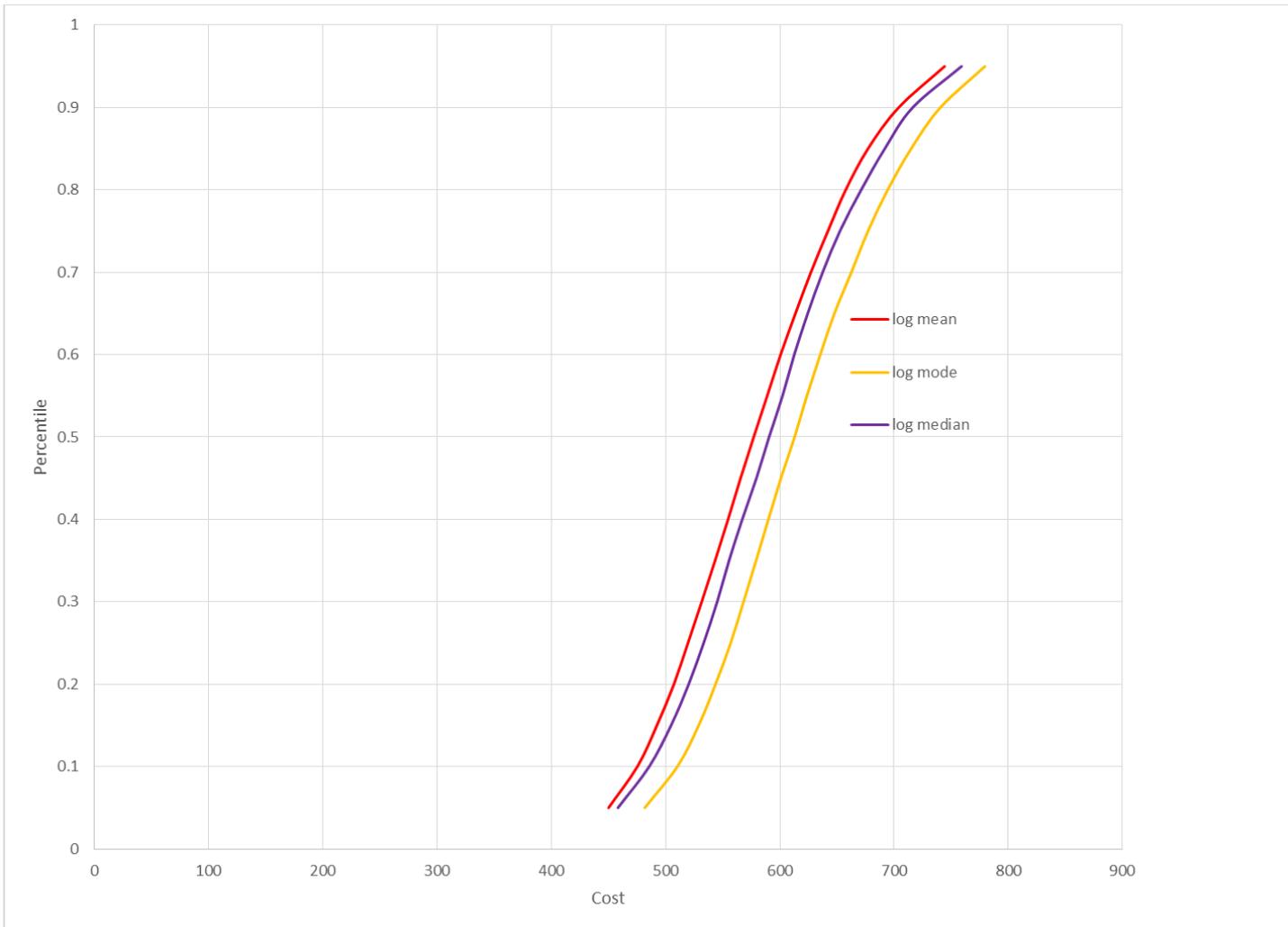


Effect of Assumed Distribution





Lognormal Distributions with Low Error





Lognormal Distributions with More Typical Error

