

Build Your Own Distribution Finder

2010 ISPA/SCEA Joint Annual Conference

Alfred Smith
Tecolote Research, Inc.

ABSTRACT

A basic step in a cost uncertainty analysis is to define the distribution of every uncertain element in the cost model. Identifying and then defending these distributions is a fundamental challenge. If data is available, the preference is always to perform a statistical analysis to arrive at an objective assessment. Several commercial tools are available for finding the distribution that best describes the shape and dispersion of the sample data set, but they are not in agreement.. Are they all correct? Have these different methods been subjected to an independent validation and verification (IV&V)? Is there a “best method” to derive the “best fit”?

In order to easily analyze hundreds of data sets in a consistent manner and present the results in a tailored form, a prototype utility was built in Excel to derive the parameters for the lognormal, normal, triangular and beta distribution that best fit a sample dataset. A variation on Excel’s PercentRank function is introduced and forms a key building block of the utility. Excel’s solver is used in the prototype and the motivation to search for an alternative is presented. The Chi squared statistic is used by at least one commercial tool as the metric for optimizing the distribution parameters. We examine it and several others as metric to optimize such as sum of squared error (SSE) and standard percent error (SPE). The pros and cons of each are presented and the rationale for the one selected is provided. The applicability of other “goodness of fit” tests are discussed.

We present the fit results in a compact and thorough format. Fitted distribution parameters compare favorably to commercial tools, and the math is provided for validation. The Chi squared test is used to assess the significance of the fitted distribution. The associated assumptions, math, and weaknesses of this “goodness of fit” test are discussed.

References:

1. *NIST/SEMATECH e-Handbook of Statistical Methods*, <http://www.itl.nist.gov/div898/handbook/>, 30 November 2009
2. *Guide to Using @Risk Risk Analysis and Simulation Add-In for Microsoft® Excel, Version 5.5*, May 2009, Palisade Corporation
3. Oracle® Crystal Ball, Fusion Edition, User Manual, Release 11.1.1.1.00, September 2008
4. *Numerical Recipes in C: The Art of Scientific Computing*, Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, New York: Cambridge University Press, 1988



TECOLOTE
RESEARCH, INC.
Bridging Engineering and Economics
Since 1973

Build Your Own Distribution Finder

2010 ISPA/SCEA Joint Conference & Training Workshop

8 - 11 June 2010

Alfred Smith CCEA

- Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara
- Albuquerque ■ Colorado Springs ■ Ft. Meade ■ Ft. Monmouth ■ Goddard Space Flight Center ■ Ogden ■ Patuxent River ■ Silver Spring ■ Washington Navy Yard
- Cleveland ■ Dahlgren ■ Denver ■ Johnson Space Center ■ Montgomery ■ New Orleans ■ Oklahoma City ■ San Antonio ■ San Diego ■ Tampa ■ Tacoma ■ Vandenberg AFB



Acknowledgements

- **Jeff McDowell, Dr Lew Fichter and Dr Shu-Ping Hu for their valuable insights, guidance and testing of the concept and Gina Fennell for her help with the presentation content**
- **John Sandberg for resolving many technical issues while converting the concept into a robust utility complete with many features that very much improved accuracy, speed of calculation and the user interface**
- **Army ODASA-CE funded the concept development**
- **AFCAA allowed us to test the concept in the development of the Cost Risk and Uncertainty Metrics Manual (CRUAMM)**
 - See the AFCAA CRUAMM presentation at 1415 Thursday.

- **What is a Distribution Finder?**
- **Why create one from scratch?**
 - What about commercial tools?
- **Core elements of a Distribution Finder process**
- **Choosing a Goodness-of-Fit test**
 - Details of the Chi-squared test
- **Recommended fit options and defaults**
- **Sample results**
 - Fitted parameters and fitted distribution formulas
 - Comparing results to Crystal Ball and @Risk for some datasets
- **Conclusions**

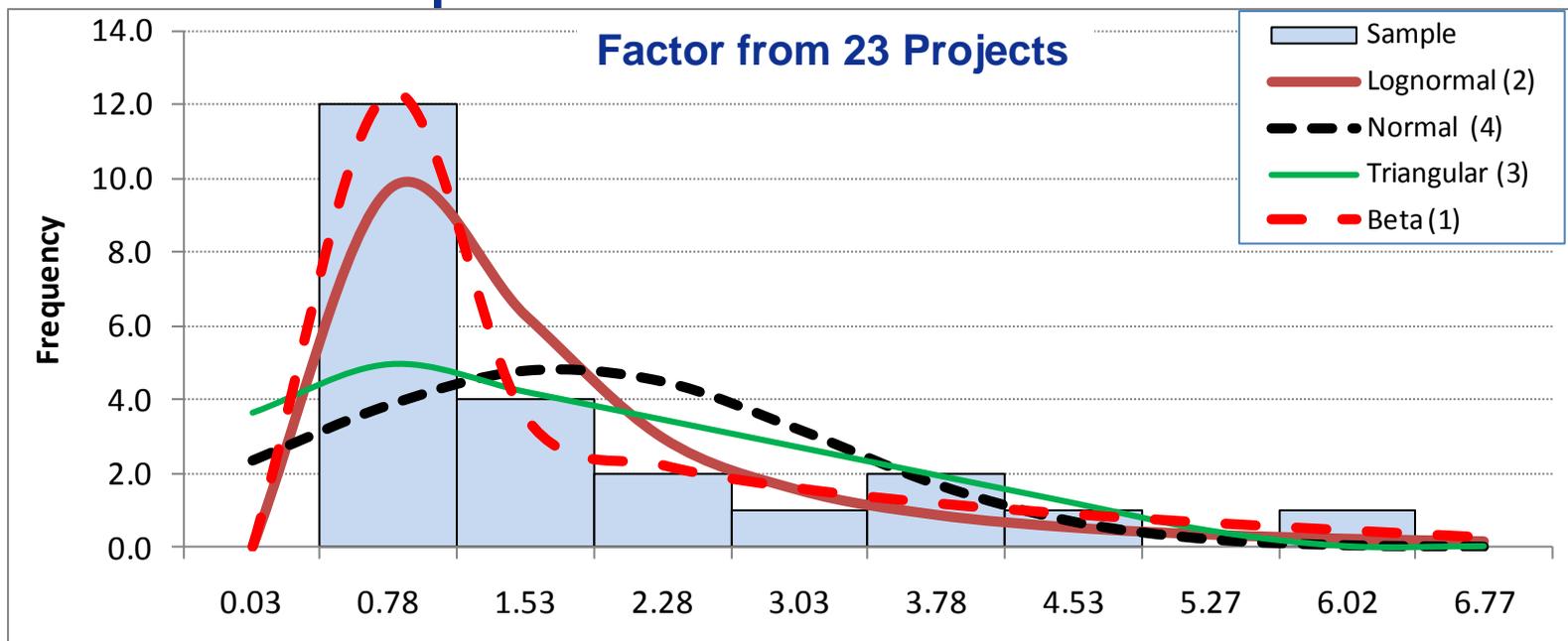
What is a Distribution Finder and Why do we need one?

- **A key step in performing cost uncertainty analysis is to define the distribution for every uncertain element in the cost model**
 - Identifying and defending uncertainty distributions is a fundamental challenge
- **Analysts should have a defensible, repeatable process to find a distribution that adequately describes the uncertainty of a cost estimating relationship (CER) and/or cost drivers (CER inputs)**
- **Assuming normal or triangular can be inappropriate – example to follow**
- **A Distribution Finder utility finds the distribution shape and it's parameters to best represent the sample data**



Real Data, Real Example, Real Consequences

- A common factor relationship was calculated for 23 similar projects
- A univariate analysis computed the mean, confidence intervals and prediction intervals for use in estimating new projects
- However, the assumption that the factors are “Normally” distributed is clearly not supported by the histogram
- In this case, beta or lognormal fits the data much better than normal and the mode is quite different from the mean



Why Create a Distribution Finder from Scratch?

- **Several commercial tools provide for distribution fitting, however:**
 - All are equally adamant that their method is “the best”, yet all yield different answers given the same dataset
 - Sometimes not possible to validate the methods used (if the methods are published at all)
 - Customizing their results into a desired format is not trivial
 - Analyzing hundreds of data sets in a repeated, consistent manner is cumbersome
 - Unclear or impossible to enforce economic or physical realities such as: $Low > 0$, $Low \leq Lowest \text{ Sample Point}$ or $High \Rightarrow Highest \text{ Sample point}$, etc.
 - Are we limited to these commercial tools to perform this analysis?
- **We were motivated to build a transparent utility within Excel that could be easily modified by the user as requirements changed**



**TECOLOTE
RESEARCH, INC.**
*Bridging Engineering and Economics
Since 1973*

Creating a Distribution Finder



Key Elements of a Distribution Finder Process

- **Goal: Fit Lognormal, Normal, Triangular and Beta to the sample data**
 - Focus is on commonly used distributions. Others can be added or leave the occasional more esoteric distribution fitting to the commercial tools.
- **Store selected source data on a single, easy to access spreadsheet**
 - Include filters to allow the analyst to easily stratify data
- **Allow analysts to easily create one or more Analysis sheets to:**
 - Select filtered (stratified) data for analysis
 - Inspect selected data for possible outliers, exclude as appropriate
 - Choose plotting options, fit constraints, fit method, basis for goodness-of-fit measure and histogram bin number
 - Render results in both tabular and graphical form
 - Report goodness-of-fit test results to identify the significance levels of fits
- **Summary Sheet**
 - Be able to tabulate fit statistics from a variety of Analysis sheets



Core Steps on the Analysis Sheet (1/2)

- We considered several distribution fit approaches and settled on the following
- Sort sample data in ascending order
- Assign a cumulative percentile
 - Several methods available, we have considered:
 - NIST¹, Excel, and a “correction for continuity” (CoC) method
 - See the next few slides for definitions and comparisons
- Use the sample descriptive statistics to provide a starting point for fit parameters
- Assess the difference between the sample and fit using either:
 - Sum Squared Error
 - Sum Squared Percent Error
 - n = number of data points
 - y = a sample data point
 - \hat{y} = a fitted point

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

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

1. NIST= National Institute of Standards and Technology

How to Estimate a Sample Data Point's Cumulative Percentile

- Three methods to estimate sample data point percentiles were considered:

<p>NIST¹</p> $Rank = p(n + 1)$ $p = \frac{Rank}{n + 1}$	<p>Excel¹ (PercentRank)</p> $Rank = 1 + p(n - 1)$ $p = \frac{Rank - 1}{n - 1}$
<p>Correction For Continuity² (CoC)</p> $p = \frac{0.5 * ObsFreq + NumObsBelow}{n}$	

p = percentile
n = number of observations

ObsFreq = number of times a specific observation occurs

NumObsBelow = number of observations below the value of sample point being assessed

- Observations**

- Excel reports the lowest point as 0% and highest as 100%. This is inconsistent with how cost estimators tend to view sample data and is a problem when trying to fit lognormal and normal
- The May, Alan method deals with duplicate data with a “correction for continuity” (CoC)

1. The NIST and Excel formulas can be found at: NIST/SEMATECH e-Handbook of Statistical Methods, www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm
NIST formulation is available in Excel 2010, see: <http://blogs.msdn.com/excel/archive/2009/09/14/function-consistency-improvements-in-excel-2010.aspx>

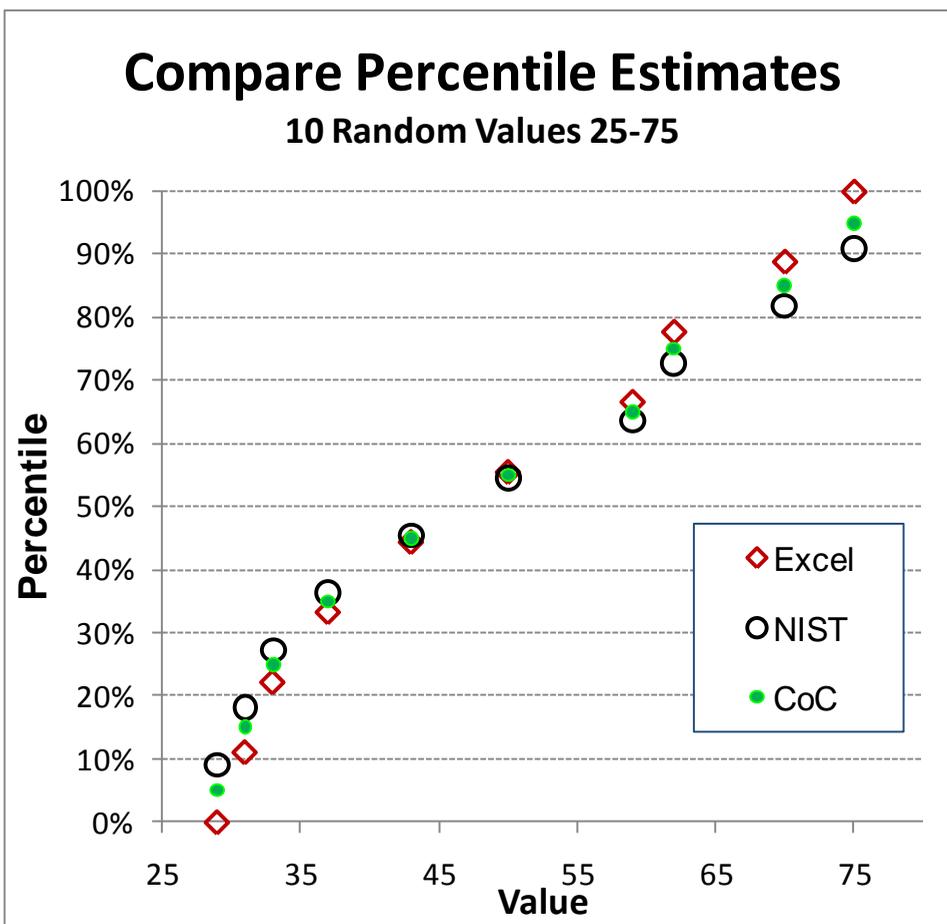
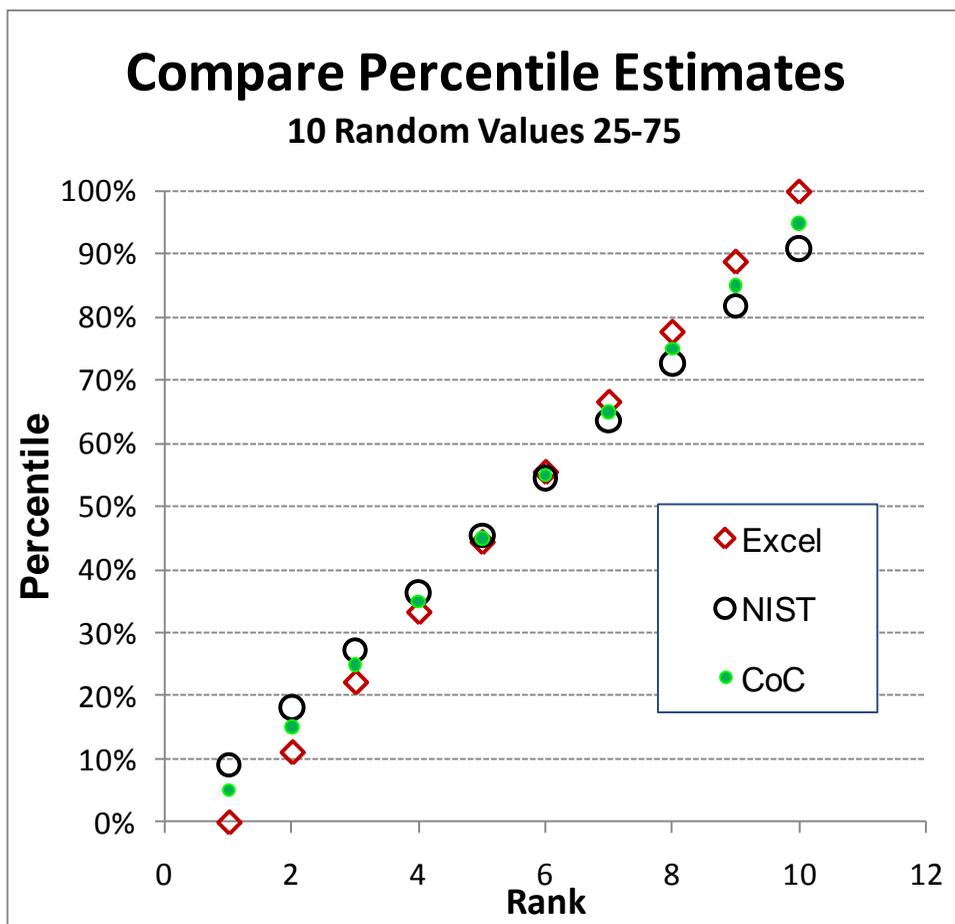
2. From “Reliability and Information Functions for Percentile Ranks” Kim May and W. Alan Nicewander, Journal of Educational Measurement, Vol. 31, No. 4 (Winter, 1994), pp. 313-325



Comparing the Percentile Methods

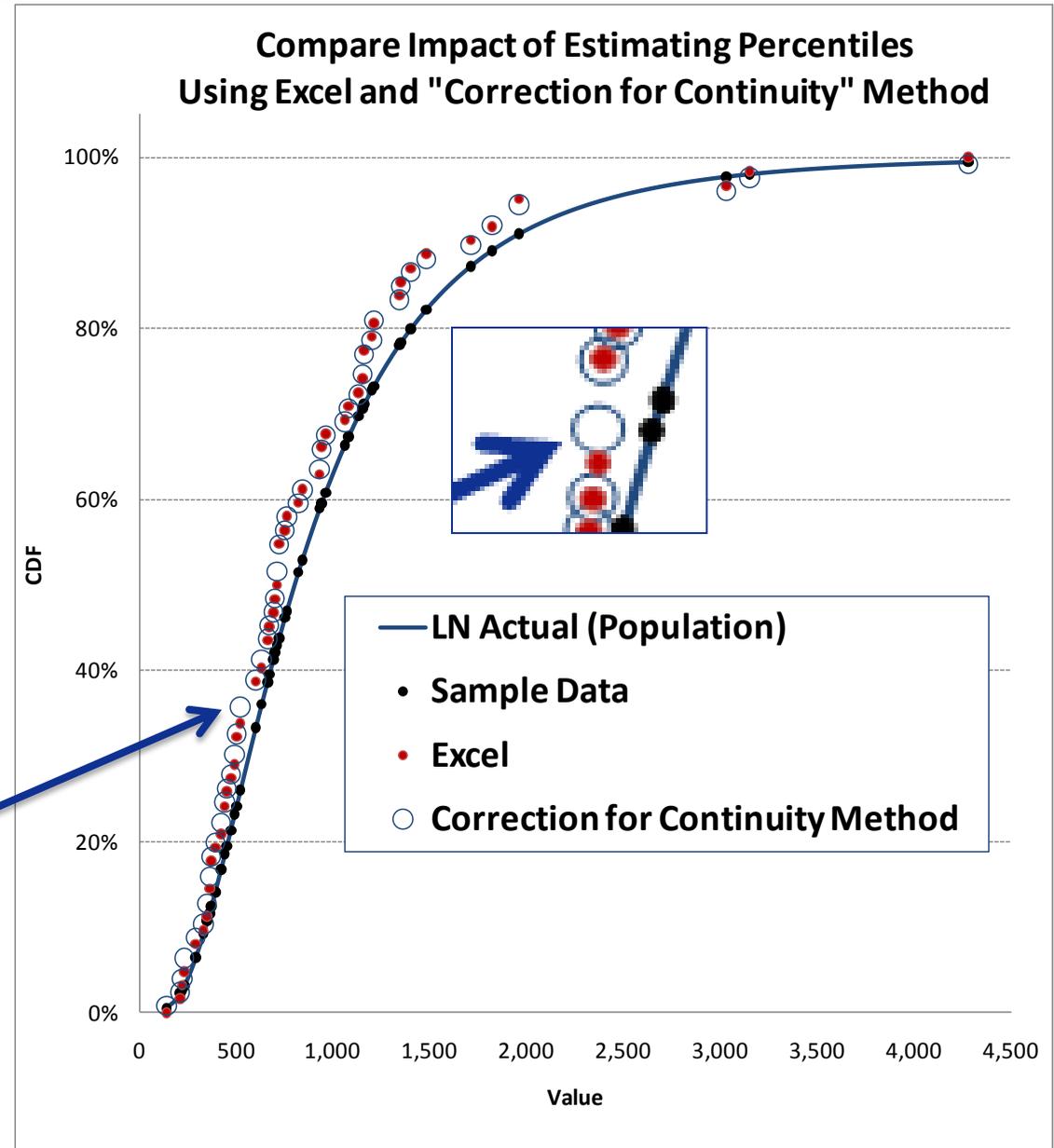
Observations

- The CoC method evenly splits the difference between NIST and Excel
- All three methods line up in the center of the data
- Biggest differences are at the end points
- Differences diminish as the number of sample points increase (not shown)



Compare CoC to the Excel Percentile Method

- Samples were drawn at random from a known lognormal population
- Data were rounded to cause several duplicates to be contained in the sample in order to illustrate how the CoC method differs from Excel or NIST:
 - Excel and NIST report the percentile of the first occurrence for all duplicates
 - The CoC method averages the percentile of the point prior and the point after the duplicates, which tends to smooth out the curve (ie removes “gaps”)
- We use the CoC method. How this choice impacts results is presented later.





Core Steps on the Analysis Sheet (2/2)

- **Use *Excel Solver to find the fit parameters that minimize SSE or SSPE**
 - Set optional constraints such as: Low>0, High>HighestSamplePoint
 - Select SSE or SSPE as error to be minimized
 - SSE is highly influenced by very large sample points (as compared to mean)
 - SSPE is highly influenced by fitted points close to zero (divide by zero problem)

- **Rank the fits using Standard Error of the Estimate (SEE) or Standard Percent Error (SPE)**
 - Where k = number of estimated parameters in the fit
 - Normal, lognormal k = 2
 - Triangular k = 3
 - Beta k = 4
$$SEE = \sqrt{\frac{SSE}{n - k}} \qquad SPE = \sqrt{\frac{SSPE}{n - k}}$$
 - This is a preferred method to rank the fits (rather than SSE or SSPE directly) because it accounts for the degrees of freedom

- **Use a Goodness-of-Fit test to determine the significance level of the fit**
 - Minimum SSE or SSPE alone does not necessarily mean the fit is meaningful

* Microsoft Excel Solver was developed by Frontline Systems, Inc. and distributed with MS Excel

■ Kolmogorov-Smirnov:

- The sample CDF is compared to the fitted CDF and the maximum vertical distance between them is found. This generally happens in the middle, making K-S a preferred test if you are interested in accuracy at the center of the distributions.

■ Anderson-Darling:

- Measures total area between the sample and fit CDF and with weightings that can focus on the fit in the tails. A-D is a preferred test if you need accuracy in the tails.

■ Both K-S and A-D

- Do not require binning of the data
- Limited in the number of distributions for which a p value can be calculated (i.e. beta, triangular and uniform are not addressed)

■ Chi-Squared:

- Compares the sample frequency to the fitted frequency by bin
- The most common test because it is the easiest to calculate and is fast
- Can be applied to any continuous or discrete distribution
- **Weakness:** Relies on “binning” data and that there are no clear guidelines for selecting the number and location of the bins

■ Conclusion: Pick a consistent way to define bins and use Chi-squared

Chi-Squared Test Details

- **Null Hypotheses = Sample data fits the selected distribution**
- **Bins can be set on equal interval or equal percentile. D'Agostino & Stephens¹ recommend equiprobable bins have no bias and greater power compared to equal interval – we follow this recommendation**
- **For each bin, calculate the test statistic:**
 - $(\text{SampleFreq} - \text{ExpectedFreq})^2 / \text{ExpectedFreq}$
 - Expected frequency is calculated based upon the fitted distribution
 - Select bins such that the expected frequency (calculated from the fitted distribution) is > 1 even though some texts prefer it to be > 5
- **Sum the Chi-Square statistic for each bin and compare to critical value**
- **Calculate the critical value using $\text{CHIINV}(\text{SigLvl}, \text{df})$**
 - NIST² advises that degrees of freedom, $\text{df} = \text{Bins} - k - 1$ where k is the number of parameters estimated

1. Goodness-of-Fit Techniques [1986] - D'Agostino & Stephens, pp 69-72
2. NIST/SEMATECH e-Handbook of Statistical Methods, www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm

- **Desirable Fit Options** (many of which may not be available in commercial tools)
 - Force fitted mean and or standard deviation to match the sample
 - Limit normal distribution lower bound so that no more than x% of the distribution is in the negative region
 - Force low of a triangle or beta to greater than or equal to zero
 - Force triangle & beta bounds to be at least as low/high as the sample

- **Goodness-of-Fit Settings (to perform the Chi² test):**
 - Set the level of significance for the test
 - Set the number of histogram bins, choose from:
 - Mann-Wald* $(2 * \text{ObsCount}^2 / (\text{NORMSINV}(\text{ChiSigLvl}))^2)^{0.2}$
 - Mann-Wald/2 – recommended by D'Agostino & Stephens as a good compromise
 - Sturges' formula $(1 + 3.3219 * \text{Log } n)$ where n is the number of data points
 - Scott's Choice and Freedman-Diaconis (see en.wikipedia.org/wiki/Histogram)
 - Manual – let the user select

* Crystal Ball uses Mann-Wald or Mann-Wald/2



**TECOLOTE
RESEARCH, INC.**
*Bridging Engineering and Economics
Since 1973*

Running the Distribution Finder



■ **Plot all four curves**

■ **Set Constraints**

- Max 1% of normal < 0
- Min 0 for Tri and Beta
- Surround Tri and Beta

Plot	Select to Plot			
4	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Constrain Mean	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Constrain StdDev	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Force Min=>Zero	Limit Normal <0: 1%	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Surround Sample		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
	Fit LN	Fit Nor	Fit Tri	Fit Beta
	Lognormal	Normal	Triangular	Beta

Use these check boxes to force the fitted curve's to encapsulate entire sample, i.e., fitted minimum <= smallest obs and fitted maximum >= largest obs.

■ **Optimize on SSE**

- If sample percentiles are known, use the checkbox to minimize on percentile error rather than on the values
- Our default method estimates percentiles so our default is to minimize on values

Minimization Settings for Curve Fitting

Minimize on:

Minimize error of the sample %-tile instead of the sample value

■ **For Chi^2 test**

- 0.05 sig,
- use Mann-Wald/2 to set bin count

■ **Set histogram bins for display to 10 (user can adjust without affecting Chi^2 test)**

Goodness of Fit Statistic Parameters

	Manual Sturges Mann-Wald Mann-Wald / 2 Scott Freedman Diaconis	Lvl of Sig 0.05
Freedman Diaconis 6	Bin Selector: Mann-Wald / 2	Chi^2 Test Bins 10
Scott Bins 5	Sturges Bins 7	Mann-Wald Bins 20

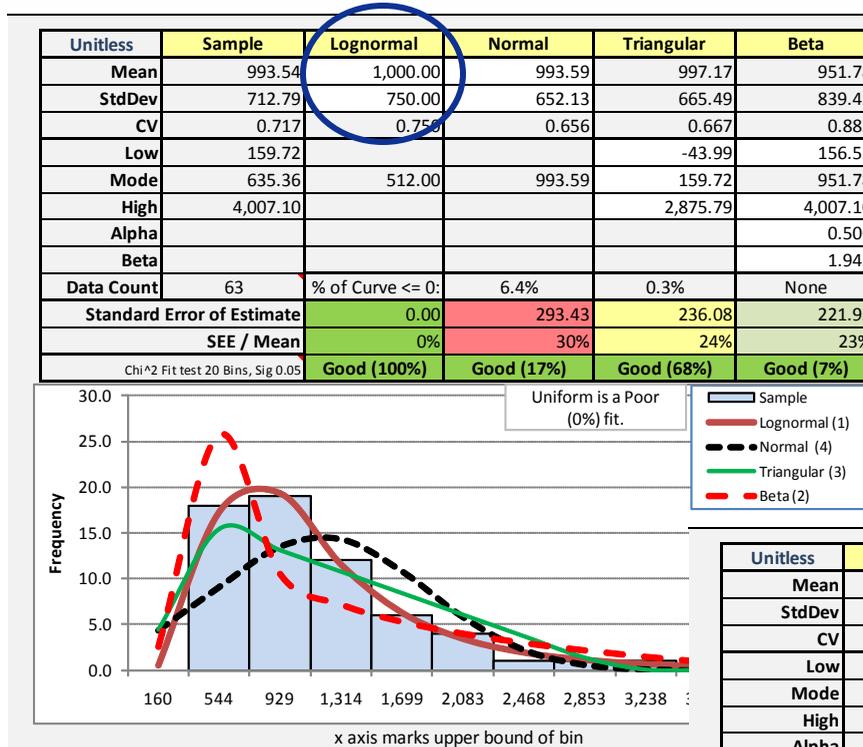
Force integer freq per bin (default unrounded)



Sample Distribution Finder Results

- Both examples are fit to 63 data points from a known population
- Top image: data is from defined percentiles

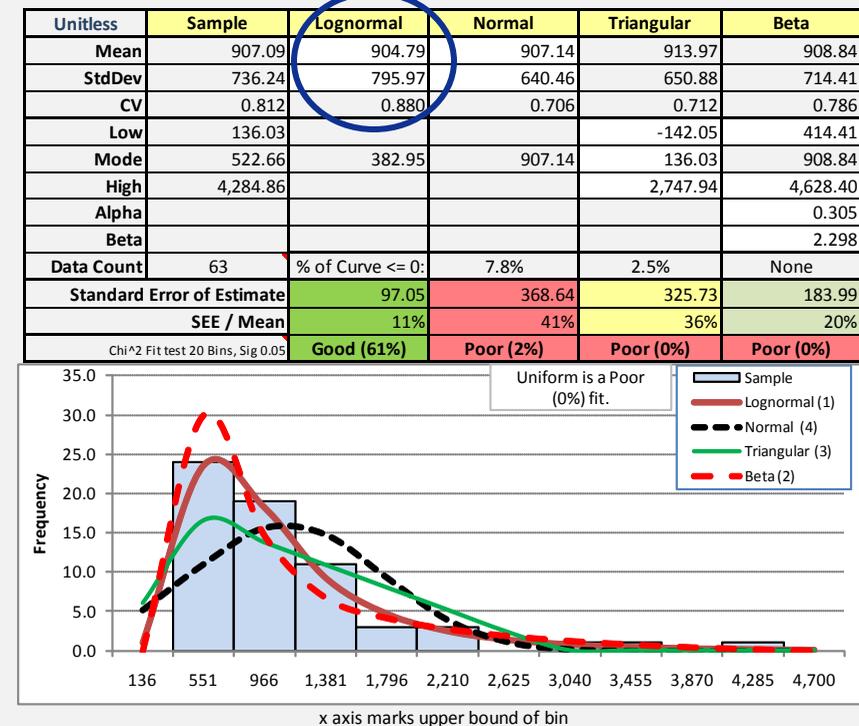
Note that the LN fit is perfect



- Bottom image: data percentiles are estimated.

Note that LN is found to be the best fit

- Similar results when data drawn from known Normal, Triangular or Beta populations

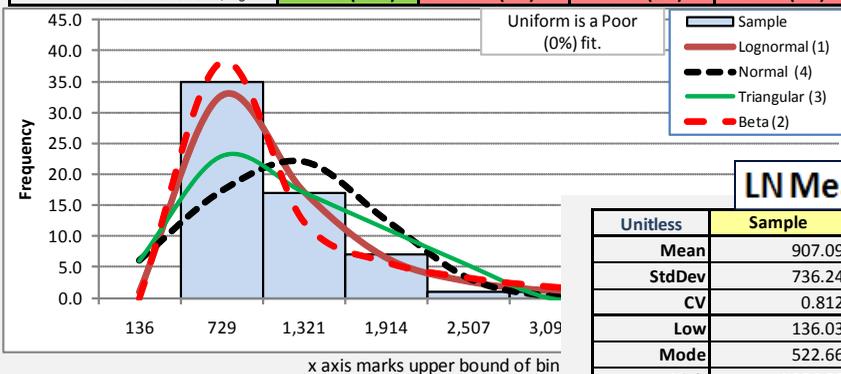




A Seemingly Trivial Difference In Calculating Percentile Has a Huge Impact

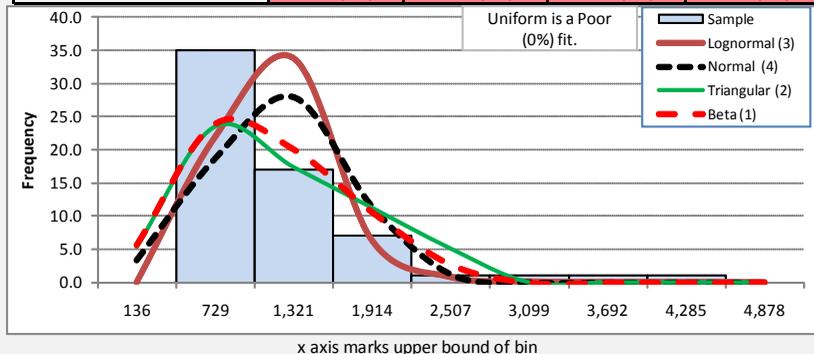
LN Mean 1000, Stdev 750 (CoC %tile)

Unitless	Sample	Lognormal	Normal	Triangular	Beta
Mean	907.09	904.79	907.14	913.97	908.84
StdDev	736.24	795.97	640.46	650.88	714.41
CV	0.812	0.880	0.706	0.712	0.786
Low	136.03			-142.05	414.41
Mode	522.66	382.95	907.14	136.03	908.84
High	4,284.86			2,747.94	4,628.40
Alpha					0.305
Beta					2.298
Data Count	63	% of Curve <= 0:	7.8%	2.5%	None
Standard Error of Estimate		97.05	368.64	325.73	183.99
SEE / Mean		11%	41%	36%	20%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (92%)	Poor (0%)	Poor (1%)	Poor (0%)



LN Mean 1000, Stdev 750 (Excel %tile)

Unitless	Sample	Lognormal	Normal	Triangular	Beta
Mean	907.09	903.16	913.26	912.11	847.63
StdDev	736.24	349.13	490.71	635.51	561.91
CV	0.812	0.387	0.537	0.697	0.663
Low	136.03			-104.00	134.00
Mode	522.66	732.89	913.26	136.03	847.63
High	4,284.86			2,704.40	2,704.40
Alpha					1.557
Beta					3.039
Data Count	63	% of Curve <= 0:	3.1%	1.6%	2.6%
Standard Error of Estimate		312.94	378.15	304.66	304.25
SEE / Mean		35%	41%	33%	36%
Chi^2 Fit test 10 Bins, Sig 0.05		Poor (0%)	Poor (0%)	Poor (1%)	Poor (0%)



■ Random sample from a known LN

■ Using CoC %tile

- LN is correctly identified
- Mean and Stdev reasonable

■ Using Excel %tile

- Beta is identified as best fit
- Mean and Stdev unacceptable
- No fit passes Chi^2

Fitted Parameters

1. Sample descriptive statistics accounting for excluded outliers
2. “Fitted” mean, standard deviation for Lognormal and Normal
3. “Fitted” low, mode and high for Triangular
4. “Fitted” low, high, alpha and beta for Beta
5. % of the Normal, Triangular, and Beta below zero
(can be forced to be x% for normal, 0 for triangular or beta)

Unitless	1 Sample	Lognormal	Normal	Triangular	Beta
Mean	907.09	2 904.79	907.14	913.97	908.84
StdDev	736.24	795.97	640.46	650.88	714.41
CV	0.812	0.880	0.706	0.712	0.786
Low	136.03			3 -142.05	414.41
Mode	522.66	382.95	907.14	136.03	908.84
High	4,284.86			2,747.94	4,628.40
Alpha					4 0.305
Beta					2.298
Data Count	63	5 % of Curve <= 0:	7.8%	2.5%	None

Fitted Distribution Equations

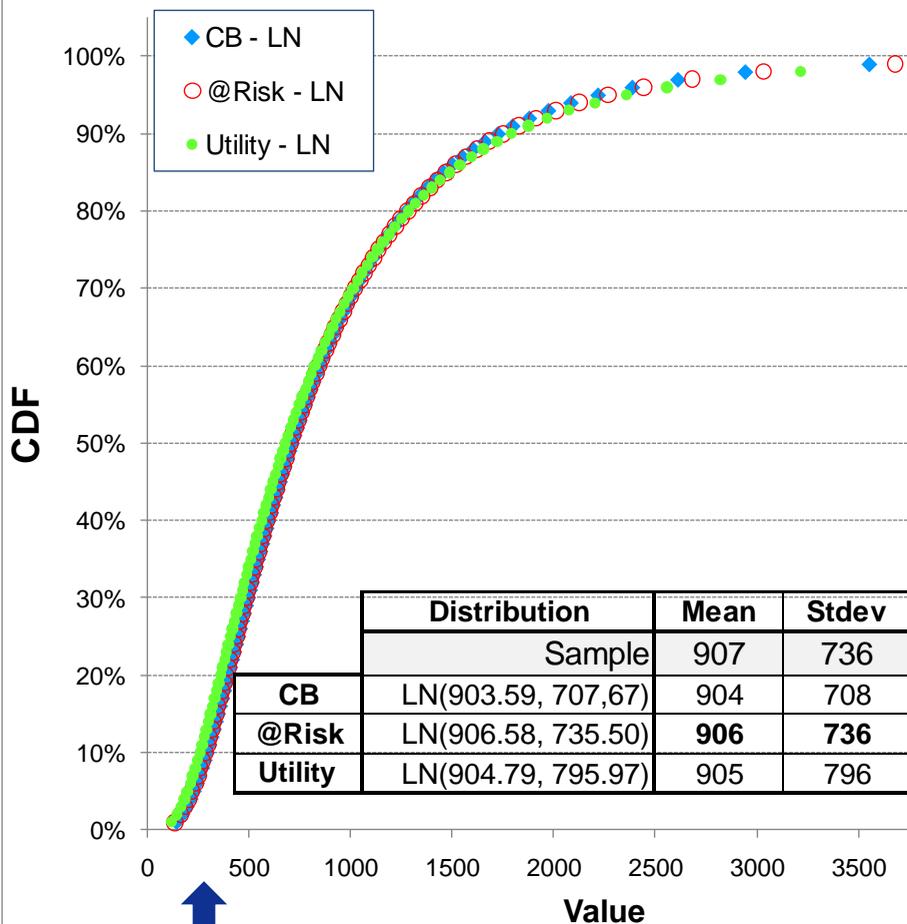
1. **LOGINV(Percentile, Mean, StdDev)**
2. **NORMINV(Percentile, Mean, StdDev)**
3. **For Triangular, if 1st equation < mode then use it, else use 2nd**
 1. $(\text{Percentile} * (\text{High} - \text{Low}) * (\text{Mode} - \text{Low}))^{0.5} + \text{Low}$
 2. $-(((1 - \text{Percentile}) * (\text{High} - \text{Low}) * (\text{High} - \text{Mode}))^{0.5} - \text{High})$
4. **BETAINV(Percentile, Alpha, Beta, LowBeta, HighBeta)**

G	H	I	J	K	L	O	V
%	Sorted Data	Lognormal Estimate	Squared Error	Squared Percent Error	Normal Estimate	Triangular Estimate	Beta Estimate
0.79%	136.03	109.41	708.68	0.06	-637.54	-62.19	414.41
2.38%	210.94	151.64	3,517.36	0.15	-361.46	-3.72	414.42
3.97%	216.00	179.99	1,296.83	0.04	-216.48	36.53	414.45
5.56%	225.47	203.34	489.73	0.01	-113.26	69.25	414.53
7.14%	227.73	224.03	13.66	0.00	-31.29	97.54	414.68
8.73%	289.40	243.06	2,147.71	0.04	37.67	122.83	414.92



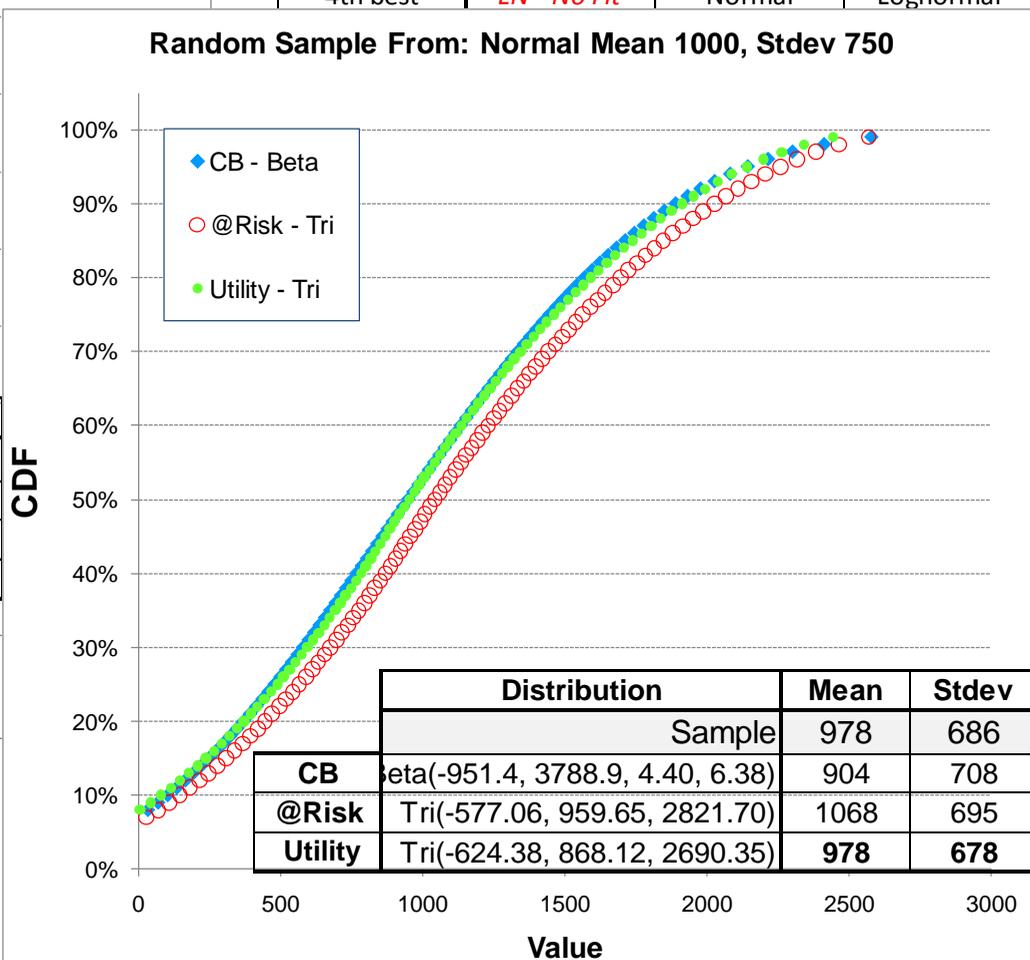
Compare To Fits From Other Tools

Random Sample From: LN Mean 1000, Stdev 750



	CB	@Risk	Utility
Plotted ->	Beta	Triangular	Triangular
2nd best	Normal	Beta	Normal
3rd best	Triangular	Lognormal	Beta
4th best	<i>LN - No Fit</i>	Normal	Lognormal

Random Sample From: Normal Mean 1000, Stdev 750



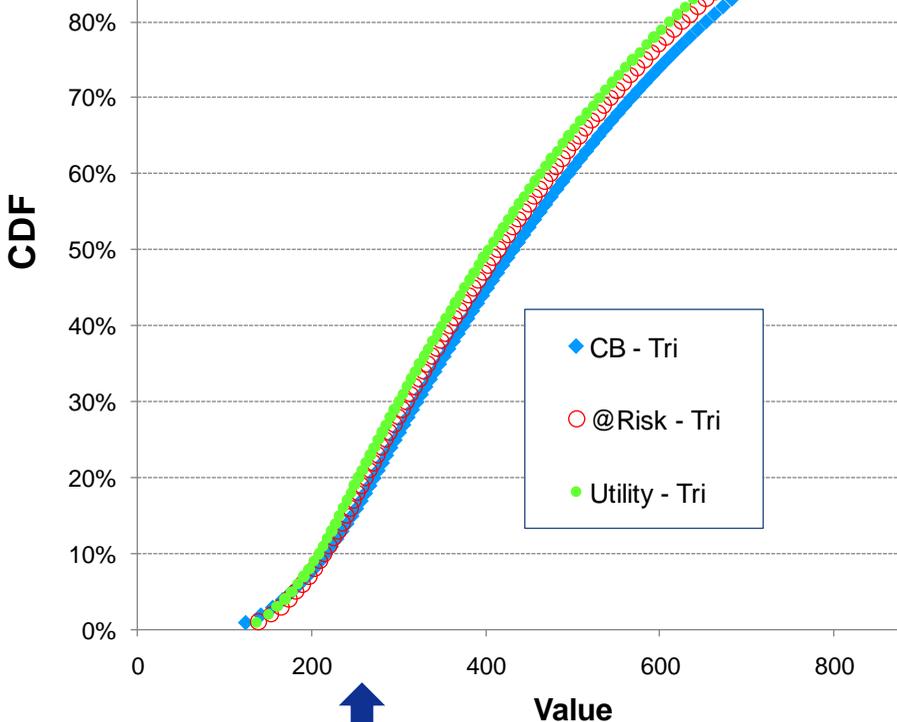
	CB	@Risk	Utility
Plotted ->	Lognormal	Lognormal	Lognormal
2nd best	Beta	Normal	Beta
3rd best	Normal	Triangular	Triangular
4th best	Triangular	<i>Beta - No Fit</i>	Normal



Compare To Fits From Other Tools

Random Sample From: Triangle 100, 300, 1000

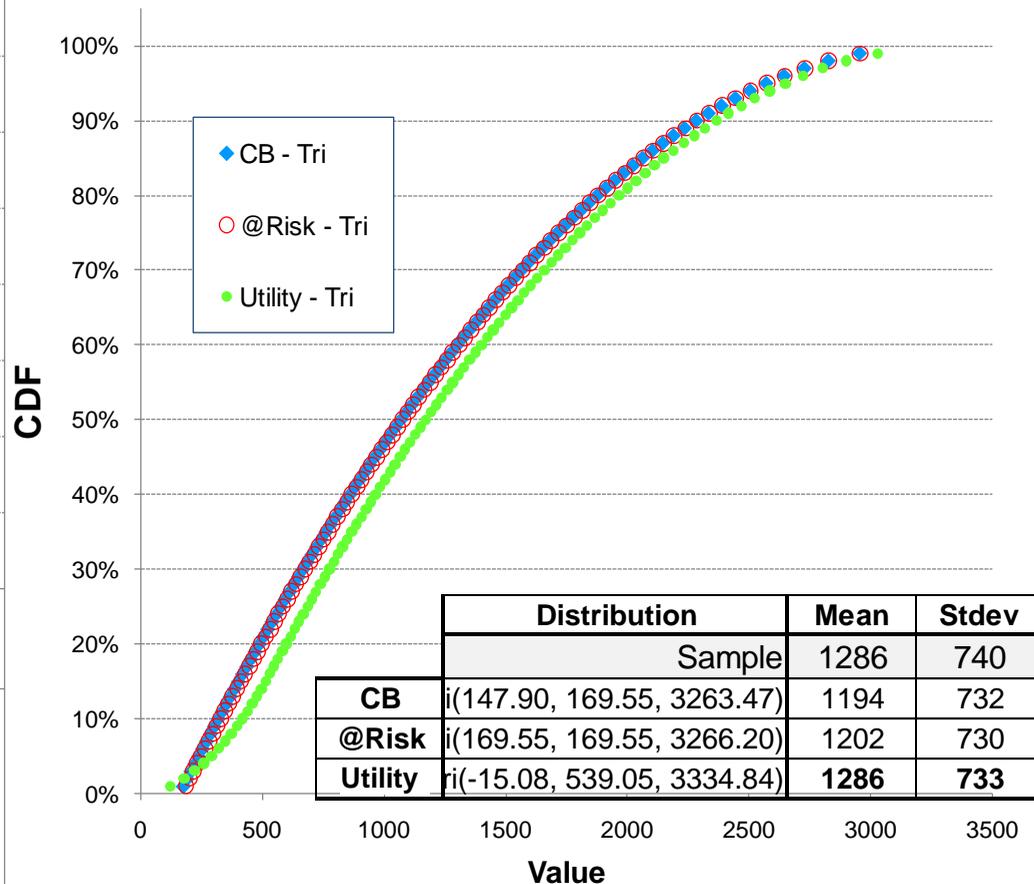
	Distribution	Mean	Stdev
	Sample	434	193
CB	ri(82.15, 271.29, 1034.55)	463	206
@Risk	ri(103.48, 242.62, 987.97)	445	194
Utility	ri(102.68, 230.57, 967.92)	434	191



	CB	@Risk	Utility
Plotted ->	Triangular	Triangular	Beta
2nd best	Beta	Beta	Triangular
3rd best	Normal	Normal	Lognormal
4th best	Lognormal	Lognormal	Normal

	CB	@Risk	Utility
Plotted ->	Triangular	Triangular	Beta
2nd best	Beta	Beta	Triangular
3rd best	Normal	Normal	Normal
4th best	Lognormal	Lognormal	Lognormal

Random Sample From: Beta 150, 3000, 1.2, 1.5



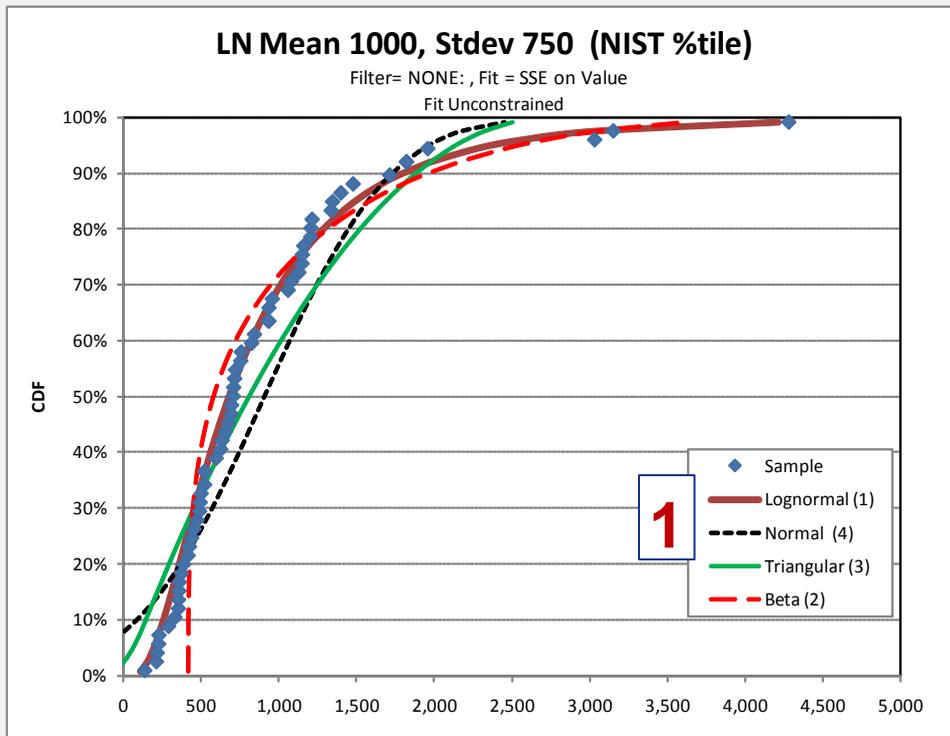
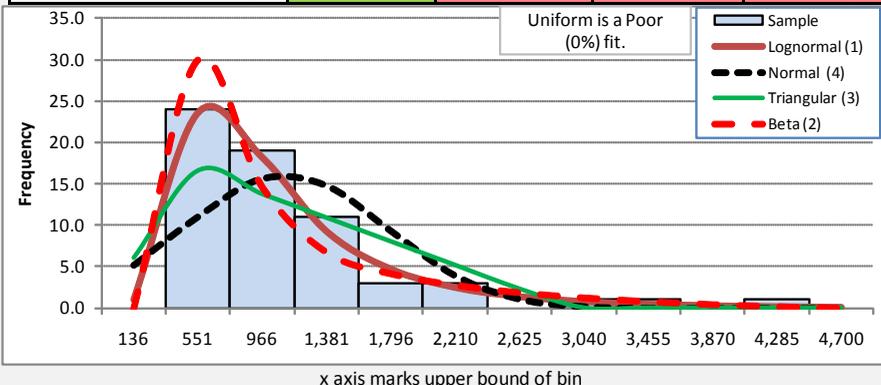
	Distribution	Mean	Stdev
	Sample	1286	740
CB	i(147.90, 169.55, 3263.47)	1194	732
@Risk	i(169.55, 169.55, 3266.20)	1202	730
Utility	ri(-15.08, 539.05, 3334.84)	1286	733



A Compact Result Format

- 1. Fits are numbered based on SEE, lowest (best) to highest (worst)**
- 2. Lowest SSE is colored dark green, next best light green**
 - In this case Lognormal and Beta respectively
 - Chi²-Test is green when the p-value \geq the critical value, red when not
- 3. Note that the fitted normal and triangular have a negative tail in this example. Also, the unconstrained triangle high tail falls far short of the sample high making this fit undesirable. Beta low is too high!**

Unitless	Sample	Lognormal	Normal	Triangular	Beta
Mean	907.09	904.79	907.14	913.97	908.84
StdDev	736.24	795.97	640.46	650.88	714.41
CV	0.812	0.880	0.706	0.712	0.786
Low	136.03			-142.05	414.41
Mode	522.66	382.95	907.14	136.03	908.84
High	4,284.86			2,747.94	4,628.40
Alpha					0.305
Beta					2.298
Data Count	63	Curve \leq 0: 7.8%	2.5%	None	
Standard Error of Estimate		97.05	368.64	325.73	183.99
SEE / Mean		11%	41%	36%	20%
Chi ² Fit test 20 Bins, Sig 0.05		Good (28%)	Poor (1%)	Poor (0%)	Poor (0%)



■ Strengths

- Easy to use with transparent calculations
- Compact and comprehensive report
- User can optimize on one of several objective functions
- Ability to constrain fit to match sample mean, standard deviation or to match financial or physical realities for low and/or high bounds
- Can be fully integrated into your Excel workbooks and reports

■ Weaknesses

- Fit process relies on estimating percentiles
- Selected Goodness-of-Fit test (CHI^2) relies on bin count for which there is no known optimum
 - we settled on Mann-Wald/2 but will use Mann-Wald if bins fall below 6
- Currently only four distributions assessed, but it is easy to add others

■ Conclusion

- Correction for continuity (CoC) method to calculate percentiles is best for this application
- Distribution fit results compare well to commercial tools