Build Your Own Distribution Finder

ACEIT Users Workshop
January 26-27, 2009
Alfred Smith

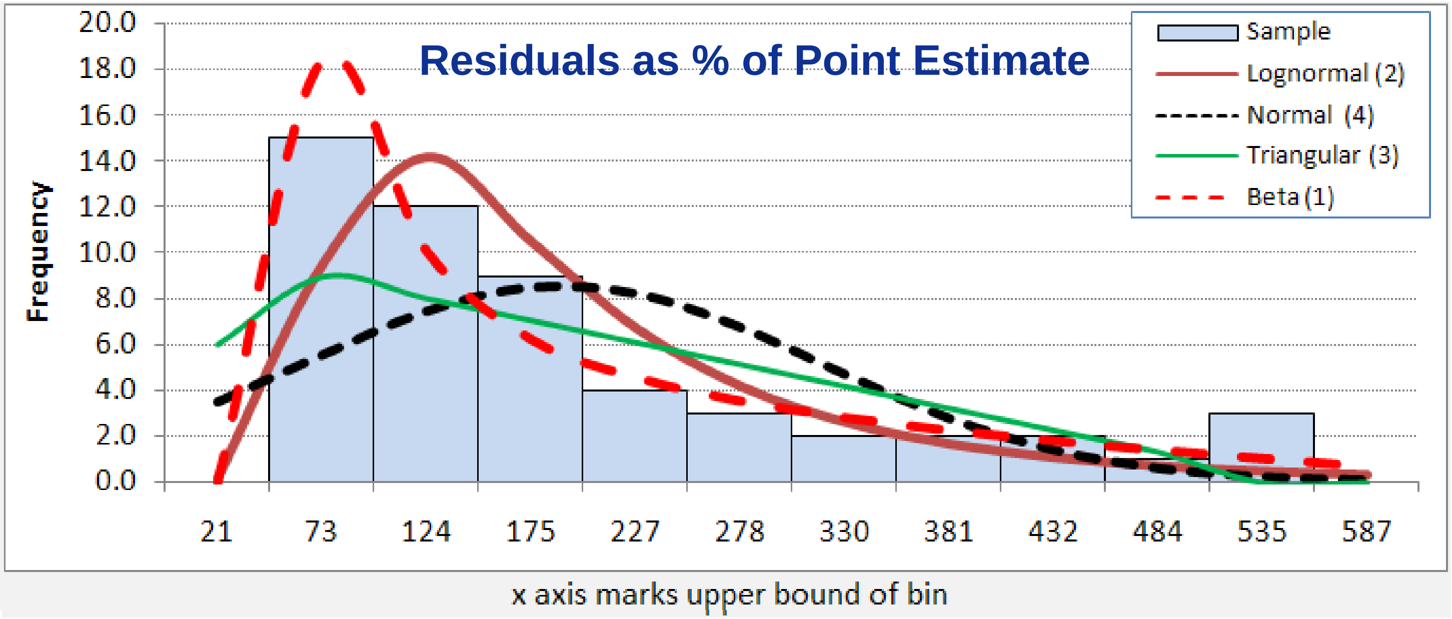
Overview

- **Why do we need a Distribution Finder?**
  - Cost Risk and Uncertainty highly visible
  - Commercial tools have limitations
- **What do we need to do to create a Distribution Finder?**
- **How do we do it?**
  - Enter normalized data
  - Calculate a percentile from sample data
  - Calculate distribution parameters and equations
- **How do we know if it's significant?**
  - Chi-Squared test
- **Live demonstration**
- **Conclusions**

Why do we need a Distribution Finder?

- **Cost Risk and Uncertainty are high priority items in the Cost Community**
- **First step for defining Cost Risk and Uncertainty is to define the distribution for every uncertain element in the cost model**
  - Identifying and then defending these distributions is a fundamental challenge of uncertainty analysis
  - Preference is to perform a statistical analysis to arrive at an objective assessment of the distribution shape and dispersion
- **This briefing will present a tool concept to support uncertainty distribution derivation:**
  - Mathematics/statistics and flow
  - Inputs/outputs
  - "What-if" capability and constraints

- **It is common to assume that the CER error term is “Normally” distributed**
  - However, this is an assumption, not a fact
  - If the error is not normal and the CER was developed using OLS, the implication is that further analysis is required
  - But if it turns out to be the best we have…what can we do?
- **The utility fits distributions to the data, giving their parameters in the form that can be used in ACE RI$K**



Commercial Tool Limitations

- **Crystal Ball and @Risk are examples of commercial tools that provide a curve fitting capability, however:**
  - Neither lend themselves to reporting results in a tailored format
  - Neither will readily analyze hundreds of data sets in a repeated manner without resorting to programming
  - Neither publish the underlying mathematics/statistics that would define how they perform the curve fits, particularly the methods used to perform the Chi-Square test (number of bins, degrees of freedom)
  - They return different results for the same data set
- **Regardless which commercial tool is selected, a large part of the ACEIT community would not be licensed to use it**
- **Therefore, we were motivated to investigate building a simple and transparent tool that would augment CO$TAT**

What do we need to do to create a Distribution Finder?

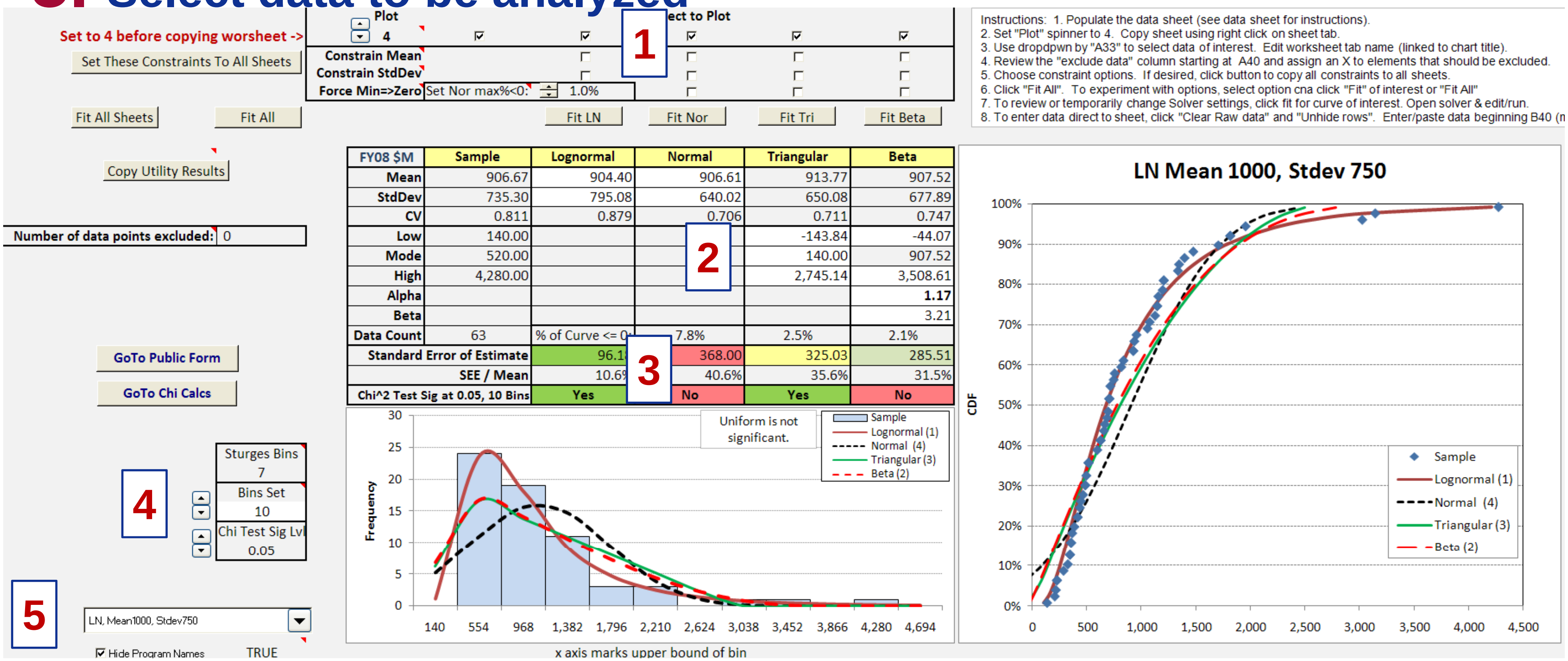
- **Goal:**
  - Fit Lognormal, Normal, Triangular and Beta to sample data
- **Steps:**
  - Sort sample data in ascending order
  - Assign a cumulative percentile using the NIST formula (different than Excel, but Excel 2010 will contain it) and apply a "correction for continuity"
    - Percentile = (0.5*ObsFreq+NumObsBelow)/ObsCount *
  - Use the sample descriptive statistics to provide a starting point for parameters for a lognormal, normal, triangular, Beta
  - For each data point, calculate the squared error: (SampleDataPoint - FittedEstimate)^2
  - Use solver to find the distribution parameters such that the Sum of Squared Errors is a minimum
  - Test for significance using the Chi Square test

*NIST= National Institute of Standards and Technology http://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm

ACEIT
Automated Cost Estimating Integrated Tools

How do we do it?

1. **User options to constrain fit**
2. **White cells are fitted parameters, all others are calculated**
3. **Quality of fit metrics**
4. **Set number of bins for the histogram (and Chi test for significance)**
5. **Select data to be analyzed**



FY08 $M	Sample	Lognormal	Normal	Triangular	Beta
Mean	906.67	904.40	906.61	913.77	907.52
StdDev	735.30	795.08	640.02	650.08	677.89
CV	0.811	0.879	0.706	0.711	0.747
Low	140.00			-143.84	-44.07
Mode	520.00			140.00	907.52
High	4,280.00			2,745.14	3,508.61
Alpha					1.17
Beta					3.21
Data Count	63	% of Curve <= 0	7.8%	2.5%	2.1%
Standard Error of Estimate		96.1	368.00	325.03	285.51
SEE / Mean		10.6	40.6%	35.6%	31.5%
Chi^2 Test Sig at 0.05, 10 Bins		Yes	No	Yes	No

Entering Data for the Fit

1. **Normalized data entered, including any blanks.**
2. **Identify potential outliers (Max shaded red, Pink if > 2 stdev from mean)**
3. **User enters an "X" if data point is to be excluded**
4. **Data is automatically sorted from low to high ("Small" function)**
5. **Percentile of sorted data= (0.5*ObsFreq+ObsNumBelow)/ObsCount**

	2 A	3	C	D 1	E	F	5	4
39	# Stdev Frm Mean	Exclude	Worksheet tab becomes Title for Charts -->	LN Mean 1000, Stdev 750			%	Sorted Data
40	-0.76		Observation 1	350.00		1	0.79%	140.00
41	-0.63		Observation 2	440.00		2	2.38%	210.00
42	-0.66		Observation 3	420.00		3	3.97%	220.00
43	0.03		Observation 4	930.00		4	6.35%	230.00
44	-0.21		Observation 5	750.00		5	6.35%	230.00
45	-0.66		Observation 6	420.00		6	8.73%	290.00
46	-0.59		Observation 7	470.00		7	10.32%	330.00
47	-0.92		Observation 8	230.00		8	12.70%	350.00
48	0.30		Observation 9	1,130.00		9	12.70%	350.00
49	3.05		Observation 10	3,150.00		10	15.87%	360.00
50	1.24		Observation 11	1,820.00		11	15.87%	360.00
51	1.43		Observation 12	1,960.00		12	18.25%	370.00
52	0.33		Observation 13	1,150.00		13	19.84%	390.00
53	0.41		Observation 14	1,210.00		14	22.22%	420.00
54	4.59		Observation 15	4,280.00		15	22.22%	420.00

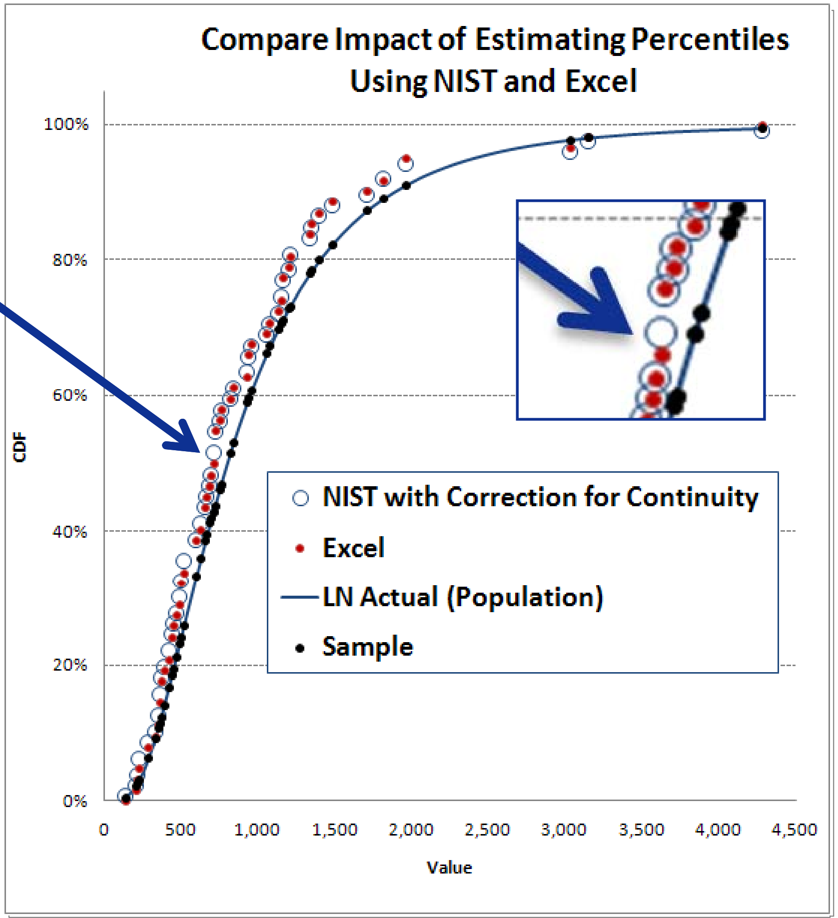
Estimating A Sample Data Point's Percentile

- **When compared to Excel, biggest relative difference is at the low end of the sample**
- **The next biggest difference is with duplicate data**
  - Excel and NIST report the first occurrence
  - The variation we use reports the mid range of duplicates, which tends to smooth out the curve (ie removes "gaps")

Excel → Rank = 1+p(N-1)

NIST → Rank = p(N+1)

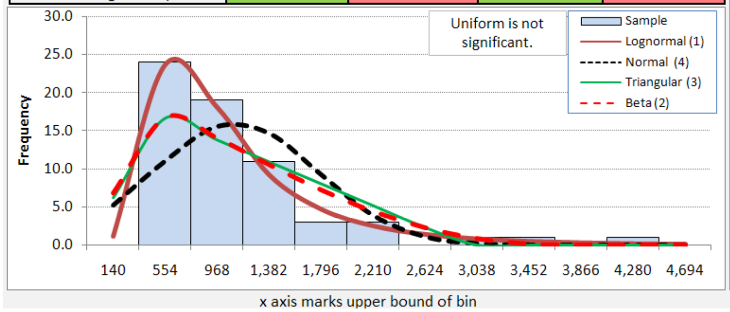
Data	NIST with Correction for Continuity	Excel	NIST/Excel
140.00	0.8%	0.0%	
210.00	2.4%	1.6%	48.8%
220.00	4.0%	3.2%	24.0%
230.00	6.3%	4.8%	32.3%
230.00	6.3%	4.8%	32.3%
290.00	8.7%	8.0%	9.1%
330.00	10.3%	9.6%	7.5%
350.00	12.7%	11.2%	13.4%
350.00	12.7%	11.2%	13.4%

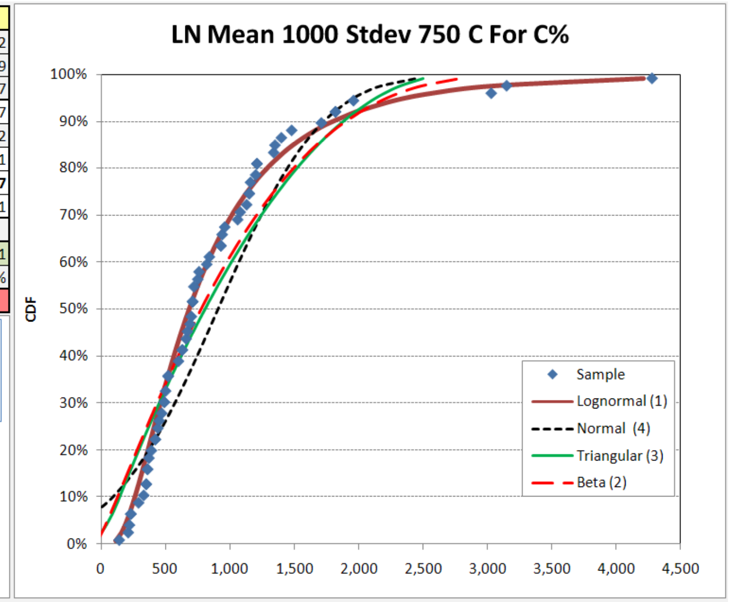


"How Percentile is Calculated" Impact on Fits

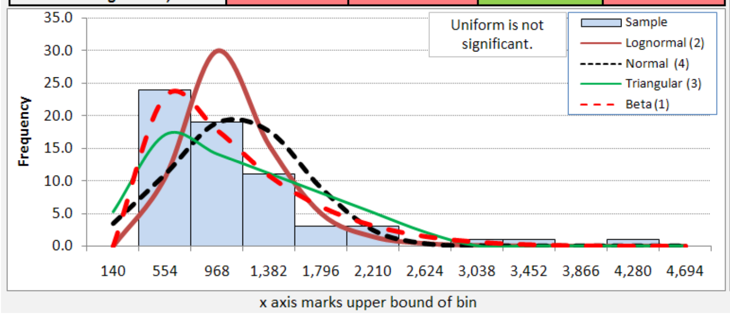
- **Using NIST**
  - LN is correctly identified
  - LN fit is statistically significant
- **Using Excel**
  - Beta is identified as best fit
  - LN is second
  - Neither is statistically significant

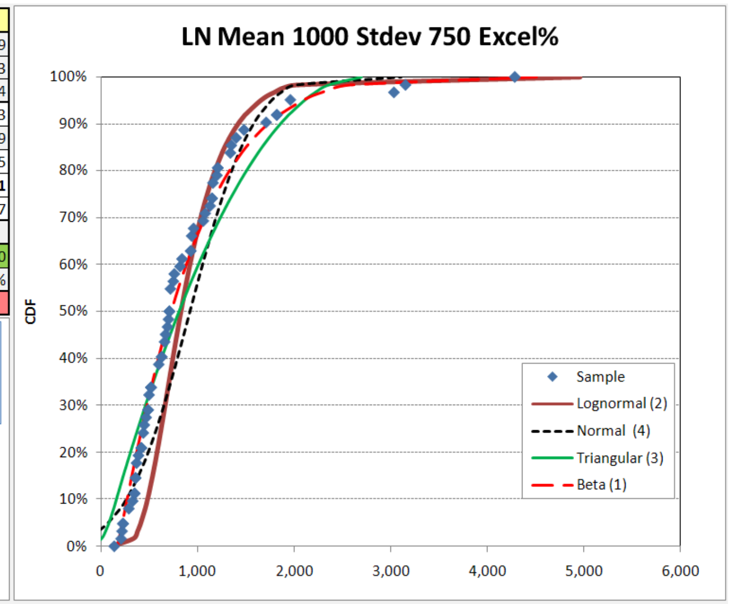
$	Sample	Lognormal	Normal	Triangular	Beta
Mean	906.67	904.40	906.61	913.77	907.52
StdDev	735.30	795.08	640.02	650.08	677.89
CV	0.811	0.879	0.706	0.711	0.747
Low	140.00			-143.84	-44.07
Mode	520.00			140.00	907.52
High	4,280.00			2,745.14	3,508.61
Alpha					1.17
Beta					3.21
Data Count	63	% of Curve <= 0:	7.8%	2.5%	2.1%
Standard Error of Estimate		96.18	368.00	325.03	285.51
SEE / Mean		10.6%	40.6%	35.6%	31.5%
Chi^2 Test Sig at 0.05, 10 Bins		Yes	No	Yes	No





$	Sample	Lognormal	Normal	Triangular	Beta
Mean	906.67	905.93	919.63	917.90	886.29
StdDev	735.30	396.82	513.20	634.09	597.63
CV	0.811	0.438	0.558	0.691	0.674
Low	140.00			-92.64	172.63
Mode	520.00			140.00	886.29
High	4,280.00			2,706.34	5,579.85
Alpha					1.11
Beta					7.27
Data Count	63	% of Curve <= 0:	3.7%	1.3%	None
Standard Error of Estimate		274.69	371.58	303.30	141.90
SEE / Mean		30.3%	40.4%	33.0%	16.0%
Chi^2 Test Sig at 0.05, 10 Bins		No	No	Yes	No





Fitted Parameters

1. Sample descriptive statistics, accounting for excluded data
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

G	H	I	L	O	T
FY08 $M	1 Sample	Lognormal	Normal	Triangular	Beta
Mean	906.67	2 904.40	906.61	913.77	1,165.81
StdDev	735.30	795.08	640.02	650.08	619.47
CV	0.811	0.879	0.706	0.711	0.531
Low	140.00			-143.84	-44.07
Mode	520.00			3 140.00	907.52
High	4,280.00			2,745.14	3,508.61
Alpha					4 1.17
Beta					3.21
Data Count	63 5	% of Curve <= 0:	7.8%	2.5%	2.1%

Fitted Distribution Equations

1. **LOGINV(Percentile, Mean, StdDev)**
   1. Squared Error = (LNestimate - SortedData)^2 (similar for other distributions)
2. **NORMINV(Percentile, Mean, StdDev)**
3. **For Triangular, if 1st equation < mode then use it, else use 2nd**
   1. (Percentile*(High-Low)*(Mode-Low))^0.5+Low)
   2. -(((1-Percentile)*(High-Low)*(High-Mode))^0.5-High)
4. **BETAINV(Percentile, Alpha, Beta, LowBeta, HighBeta)**

	G	H	I (1)	J	K	L (2)	O (3)	T (4)
39	%	Sorted Data	Lognormal Estimate	Squared Error		Normal Estimate	Triangular Estimate	Beta Estimate
40	0.79%	140.00	109.50	930.05		-636.99	-63.17	-25.09
41	2.38%	210.00	151.74	3,394.66		-361.10	-4.11	4.69
42	3.97%	220.00	180.09	1,593.08		-216.22	36.55	31.85
43	6.35%	230.00	214.04	254.63		-70.12	84.34	70.46
44	6.35%	230.00	214.04	254.63		-70.12	84.34	70.46
45	8.73%	290.00	243.16	2,194.16		37.75	123.72	107.75
46	10.32%	330.00	261.06	4,752.37		97.84	147.12	132.20

Chi-Square Test For Significance

- **Used to test if sample of data came from a population defined by a specific distribution**
  - Can be applied to any univariate distribution for which you can calculate the cumulative distribution function
- **Applied to binned data, however, for the test to be valid, the <u>expected</u> frequency for any bin should be at least 5**
  - If counts are less than 5, should combine bins *
- **Is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests**
- **Chi-Squared is the most common test to determine the significance of a fitted distribution to the sample data**
- **Critical value is calculated based upon level of significance and degrees of freedom**
  - Degrees of freedom = Bins-Parameters Estimated-1 *

***http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm**

Chi Squared Statistic Is Tricky

1. **"Count" number of sample data between the bin upper bounds**
2. **Use the "DIST" functions to calculate percent of fitted distribution between consecutive upper bounds and multiply it by the sample observation count to estimate "expected" frequency**
3. **The Chi stat is (SampleFreq – ExpectedFreq)^2/ExpectedFreq**
   - But, expected frequency per bin must be >5. In example below, LN should be collapsed to 4 bins!, normal, triangular and beta to 5 bins, that is, the bin above the red line should be wide enough to capture all the data below the red line
   - The sum of the Chi Statistic is compared to a critical value

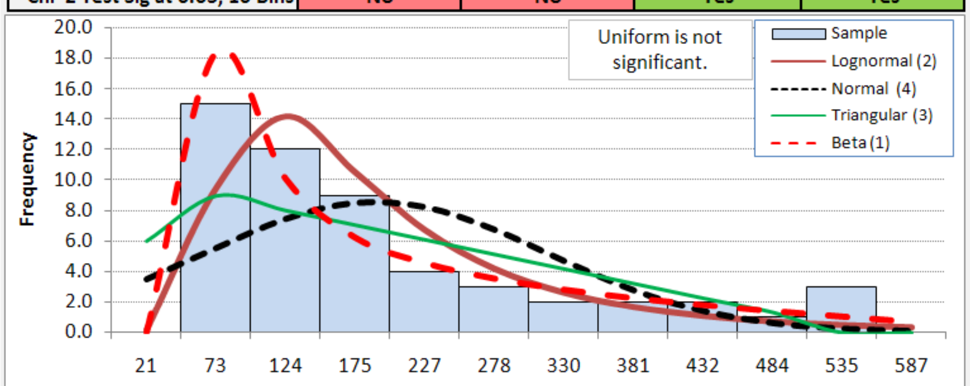
414.00	1	2 Frequency (Count <= Bin Upper Bound)			
Bin Upper Bound	Sample	Lognormal (1)	Normal (4)	Triangular (3)	Beta (2)
554	24.00	23.65	11.05	16.62	16.74
968	19.00	18.04	15.58	13.75	14.05
1,382	11.00	9.19	14.68	10.88	10.55
1,796	3.00	4.70	9.23	8.01	7.23
2,210	3.00	2.51	3.87	5.14	4.41
2,624	0.00	1.41	1.08	2.27	2.25
3,038	1.00	0.83	0.20		0.83
3,452	1.00	0.51	0.03		0.13
3,866	0.00	0.32	0.00		
4,280	1.00	0.21	0.00		

3 Chi Squared Statistic			
Lognormal (1)	Normal (4)	Triangular (3)	Beta (2)
0.005182	15.190349	3.275419	3.146247
0.051203	0.748665	2.002671	1.746913
0.355970	0.921638	0.001266	0.019083
0.612649	4.204048	3.136344	2.471393
0.093707	0.196634	0.893456	0.449485
1.413583	1.084006	2.274333	2.251939
0.034808	3.146160		0.036713
0.481863	37.803341		5.948556
0.318898	0.002079		
3.044420	8743.704878 ← !		

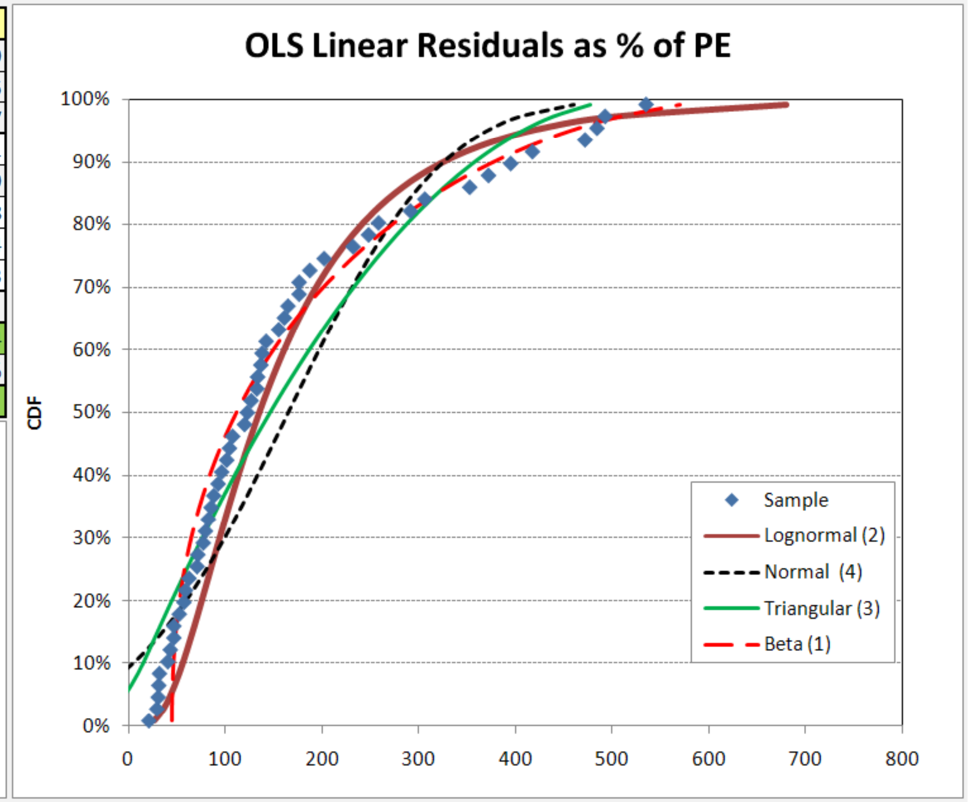
Live Demonstration

1. **Fits are numbered based on SSE, lowest (best) to highest (worst)**
2. **Lowest SSE is colored dark green, next best light green**
   - In this case Beta and Lognormal respectively
   - Chi-Test is green when "significant", red when not
     (caution: Chi-Test is not yet fully functional in the prototype)

%	Sample	Lognormal	Normal	Triangular	Beta
Mean	165.02	171.34	165.02	166.95	165.19
StdDev	136.80	133.11	125.67	128.90	134.95
CV	0.829	0.777	0.762	0.772	0.817
Low	21.09			-49.50	44.71
Mode				21.09	165.19
High	535.18			529.25	644.28
Alpha					0.44
Beta					1.73
Data Count	53	% of Curve <= 0:	9.5%	6.0%	None
Standard Error of Estimate		36.45	55.28	36.74	16.64
SEE / Mean		21.3%	33.5%	22.0%	10.1%
Chi^2 Test Sig at 0.05, 10 Bins		No	No	Yes	Yes



x axis marks upper bound of bin



Other Validation Runs

- **Bold SEE identifies "best fit"**
- **Utility found the distribution form that created the data for 4 of 6 validation runs**
  - The second column labeled normal, data below zero was excluded so it is not unexpected that normal was not the best fit

Notional Data	LN, Mean1000, Stdev750	LN, Mean1000, Stdev250	Nor, Mean1000, Stdev750	Nor, Mean1000, Stdev750	Beta, Mean3251, Stdev1216	Beta, Mean814, Stdev1035	Chi, Mean 5.6, Stdev 3.4	Chi, Mean 10.8, Stdev 4.8	Gamma, Mean 47.4, Stdev 16
Mean	906.67	952.38	868.10	3,251.11	3,251.11	814.13	5.58	47.40	46.45
Std Dev	735.30	253.91	776.71	1,216.06	1,216.06	1,034.56	3.42	16.05	22.53
CV	0.811	0.267	0.895	0.374	0.374	1.271	0.612	0.339	0.485
Lognormal Mean	904.40	953.05	933.96	3,305.96	3,305.96	893.96	5.67	47.48	46.69
Lognormal StdDev	795.08	261.04	662.99	1,023.66	1,023.66	1,041.12	3.38	16.10	22.68
CV	0.879	0.274	0.710	0.310	0.310	1.165	0.595	0.339	0.486
SEE	**96.18**	**21.86**	305.76	648.86	648.86	317.51	0.51	**1.50**	**2.31**
Normal Mean	906.61	951.71	868.14	3,265.98	3,265.98	785.85	5.58	47.40	46.45
Normal StdDev	640.02	252.45	775.31	1,157.49	1,157.49	973.98	3.30	15.73	21.76
CV	0.706	0.265	0.893	0.354	0.354	1.239	0.592	0.332	0.468
SEE	368.00	57.63	**71.82**	479.41	479.41	502.58	0.91	3.38	6.08
Triangular Absolute Low	-143.84	481.78	-970.20	51.72	51.72	-995.44	0.05	17.01	8.64
Triangular Mode	140.00	729.30	779.33	4,280.00	4,280.00	140.00	1.73	34.55	22.91
Triangular Absolute High	2,745.14	1,647.66	2,792.46	5,334.14	5,334.14	3,370.97	14.98	90.64	107.83
CV	0.711	0.263	0.886	0.354	0.354	1.103	0.598	0.331	0.471
SEE	325.03	48.82	97.07	**333.98**	**333.98**	451.62	0.63	2.88	4.83
Beta Absolute Low	-44.07	605.45	-1,142.94	-673.16	-673.16	137.20	0.62	15.56	7.84
Beta Mode	907.52	953.45	866.31	3,256.17	3,256.17	876.09	5.56	47.41	46.48
Beta Absolute High	3,508.61	1,712.15	2,696.56	5,846.22	5,846.22	4,180.00	19.38	99.68	120.95
Alpha	1.17	1.01	2.79	4.02	4.02	0.43	1.27	2.17	1.70
Beta	3.21	2.19	2.54	2.65	2.65	1.91	3.55	3.56	3.27
CV	0.747	0.263	0.880	0.354	0.354	0.976	0.617	0.332	0.473
SEE	285.51	48.69	113.53	422.68	422.68	**261.46**	**0.47**	2.84	4.61

Conclusions

- **SSE appears to be most stable**
  - SSE seems to generate results comparable to commercial tools
  - Several other "objective functions" (SPE, Chi) were explored
- **Constraining the fits to "match sample mean and/or standard deviation" or ensure the fit does not go below zero are highly desirable options**
  - Not available in the commercial tools
- **There is no known "optimum" bin count to perform the Chi test**
  - Sturges Rule (3.322 * Log10(N) +1) provides a start, but generally user needs to adjust manually to see the data "take shape" in the histogram
- **The utility is a reasonable basis for developing a "Distribution Finder" in CO$TAT**
  - Would allow ACEIT users to have a fully integrated tool to develop and use data driven uncertainty distributions in their ACE models