Presented at the 2012 SCEA/ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

Applying the Pareto Principle to Distribution Assignment in Cost Risk and Uncertainty Analysis

James R. Glenn, Business Analyst Leader
Computer Sciences Corporation

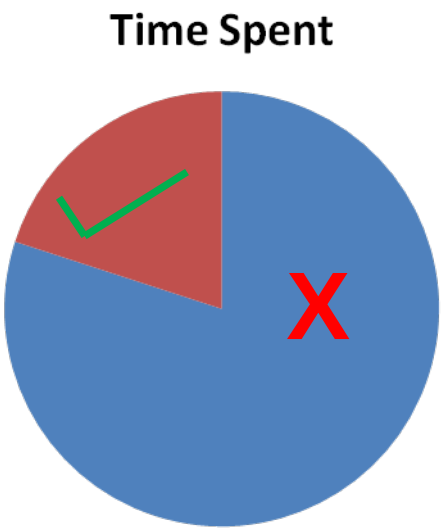
Christian Smart, Ph.D., CCEA, Director
Hetal Patel, CCEA, Cost Lead
Lawrence Johnson, CPA, Cost Analyst
Missile Defense Agency
Cost Estimating and Analysis Directorate (MDA/DOC)

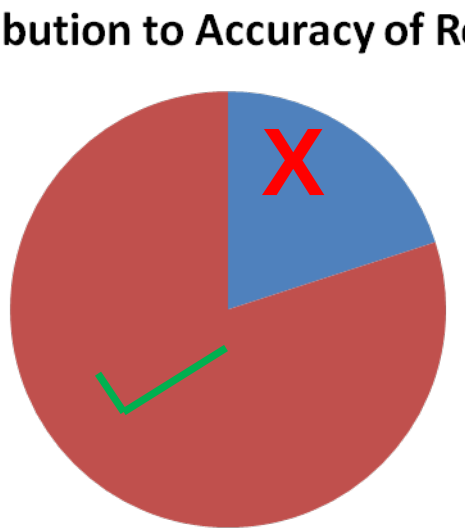
Outline

- Motivation Behind Investigation
- Pareto's Principle and its applicability to cost risk and uncertainty analysis
- Iterative Analysis
- Monte-Carlo Simulation Analysis
- Case Study
- Conclusion
- References

Motivation

- Significant effort is required to properly complete risk and uncertainty analysis for a cost estimate
- Time is limited and effort should only be spent on tasks which appreciably influence results





Input-Based Risk and Uncertainty Analysis Steps

- Distribution Assignment
- Application of Correlation
- Derivation of the aggregate Probability Distribution Function (PDF). The aggregate PDF can be computed using:
  - Probability Theory
  - An approximation technique such as Monte Carlo simulation
- Communication of results. The aggregate Cumulative Distribution Function (CDF) or S-Curve is a popular way cost risk and uncertainty results are communicated in the cost community

This presentation focuses on more efficiently completing Distribution Assignment to yield high-quality risk and uncertainty analysis results

Distribution Assignment

- The *Air Force Risk and Uncertainty Analysis Handbook* categorizes distribution assignment in two categories: subjective and objective

- Objective Distribution Assignment includes:
  - Computation of prediction intervals from parametric Cost Estimating Relationships (CER's)
  - Input modeling of appropriate distributions for datasets using goodness of fit tests such as Chi-Squared or Kolmogorov-Smirnov (K-S)

- Subjective Distribution Assignment includes:
  - Use of expert opinion to define distribution minimum and maximum values and distribution types
  - Use of default subjective distribution bounds

Default subjective distribution bounds provide the least amount of insight into the uncertainty of an estimate

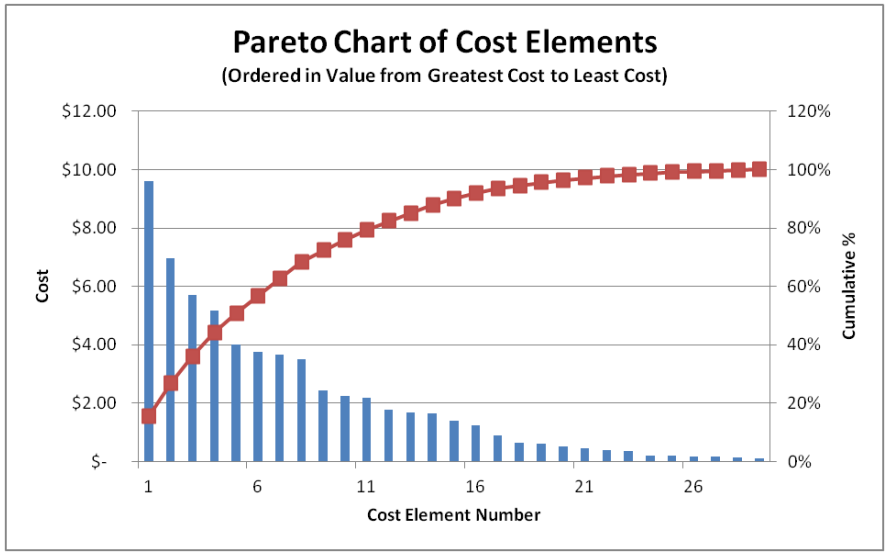
Pareto's Principle

- The Pareto Principle, named after the Italian economist, Vilfredo Pareto states that 80% of the effects come from 20% of the causes
- Applied in the context of cost analysis, this means that the top 20% of cost elements in an estimate can be expected to contain approximately 80% of total cost
- Pareto's Principle is just that, a principle. It is not a mathematical proof nor can it expected to hold true in all cases. The remainder of this presentation focuses upon applying Pareto's principle to cost risk and uncertainty analysis

Hypothesis: Pareto's principle can be applied to cost risk and uncertainty analysis

Experimental Design

- To test the hypothesis that Pareto's principle can be applied to cost risk and uncertainty analysis, the investigation begins with a sanitized version of an actual MDA estimate
- The estimate contains 29 total cost elements
  - The Top 6 cost elements (21%) account for 57% of total cost
  - The Top 12 cost elements (42%) account for 82% of total cost
- The Pareto Chart of the estimate being investigated is shown below



Iterative Analysis

- The investigation begins by iteratively adding uncertainty distributions to the cost elements in the estimate (in order from greatest cost to least cost)

- Assumptions
  - The population distribution for each cost element has a Coefficient of Variation (CV) of 50%
  - The population correlation between each of the cost elements is +0.2
  - Uncertainty distributions included are the population distributions

- Risk Measure
  - Percent difference from the population standard deviation

Standard Deviation Computation

- The population standard deviation of the aggregate distribution and the standard deviation of the aggregate distribution after iteratively adding uncertainty distributions can be computed by hand

- Key Points
  - **No correlation (0):** The covariance term cancels out and the variance of the aggregate distribution is the summation of the variance of the individual elements
  - **Perfect Positive Correlation (+1.0):** The standard deviation of the aggregate distribution is the summation of the standard deviation of the individual elements
  - If correlation is not equal to 0 nor equal to +1.0, then it is more difficult to compute by hand

Standard Deviation Computation (Formulas)

$$E(U) = \sum_{i=1}^{n} \mu_i$$

$$V(U) = \sum_{i=1}^{n} V(Y_i) + 2\sum\sum_{i<j} \mathrm{cov}(Y_i, Y_j)$$

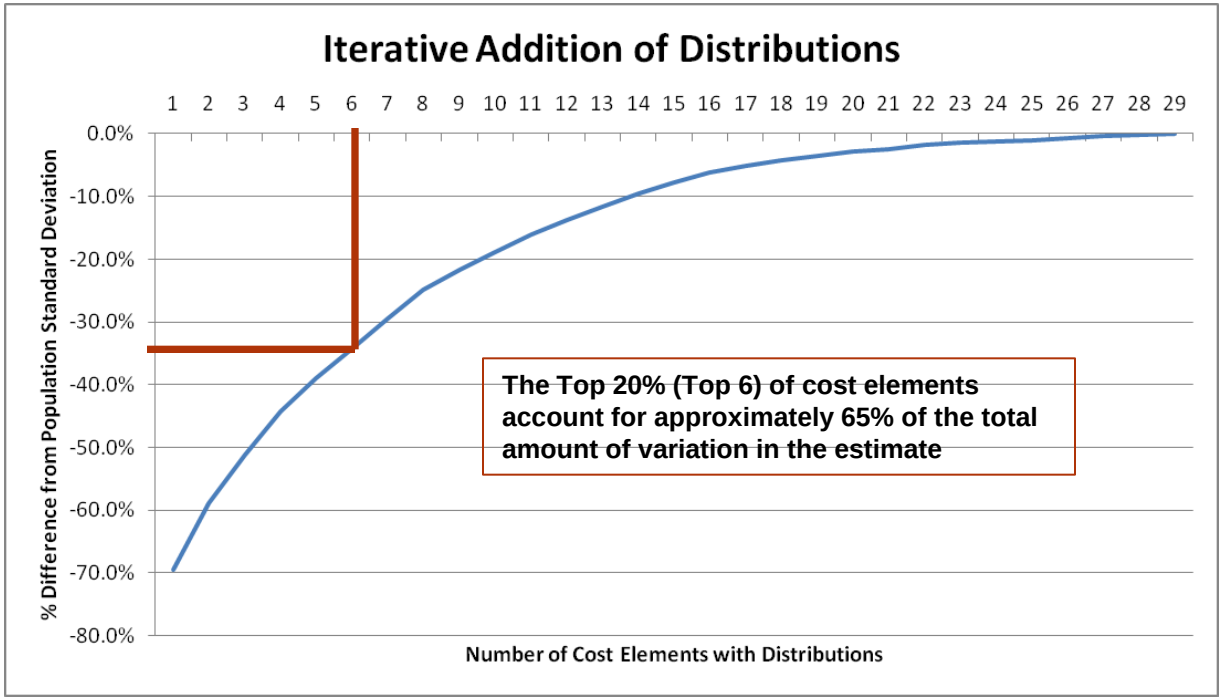
$$p_{(i,j)} = \frac{\mathrm{cov}(Y_i, Y_j)}{\sqrt{V(Y_i) \times V(Y_j)}}$$

where U is the aggregate distribution and

$Y_1, \ldots\ldots, Y_n$ are random variables

Iterative Analysis Results

- The marginal effect of adding additional uncertainty distributions decreases with each additional distribution, but there is no point of inflection
- In other words, there is not a clear point where adding additional uncertainty distributions no longer seems to be a worthwhile use of resources



A More Realistic Example

- Two issues exist with the previous example:
  - Distribution modeling almost certainly will not yield the actual population distribution
  - Applying distributions to only the top drivers may potentially significantly understate variation

- The next example addresses these noted issues with the previous example

- In this example, a distribution will be assigned to all cost elements in the estimate. Cost elements will receive either:
  - **Modeled Distributions:** Objectively defined distributions or use of expect opinion to subjectively define distributions
  - **Default Distributions:** Default subjectively defined distributions. These distributions do not require any information gathering specifically for the risk and uncertainty analysis

Experimental Trials

- A total of 12,000 experimental trials were run in Crystal Ball by varying the parameters shown in the table below
  - Only the experimental trials where modeled distributions are closer to the population distribution than default distributions are kept for further analysis
  - 5,000 replications were completed in each experimental trial

Independent Variable	Range
# of Drivers receiving modeled distributions	0 to 29
Degree Modeled Distributons vary from the actual population	Variance is varied by -50% to +50% in Increments of 5%
Degree Default Distributons vary from the actual population	Variance is varied by -50% to +50% in Increments of 5%

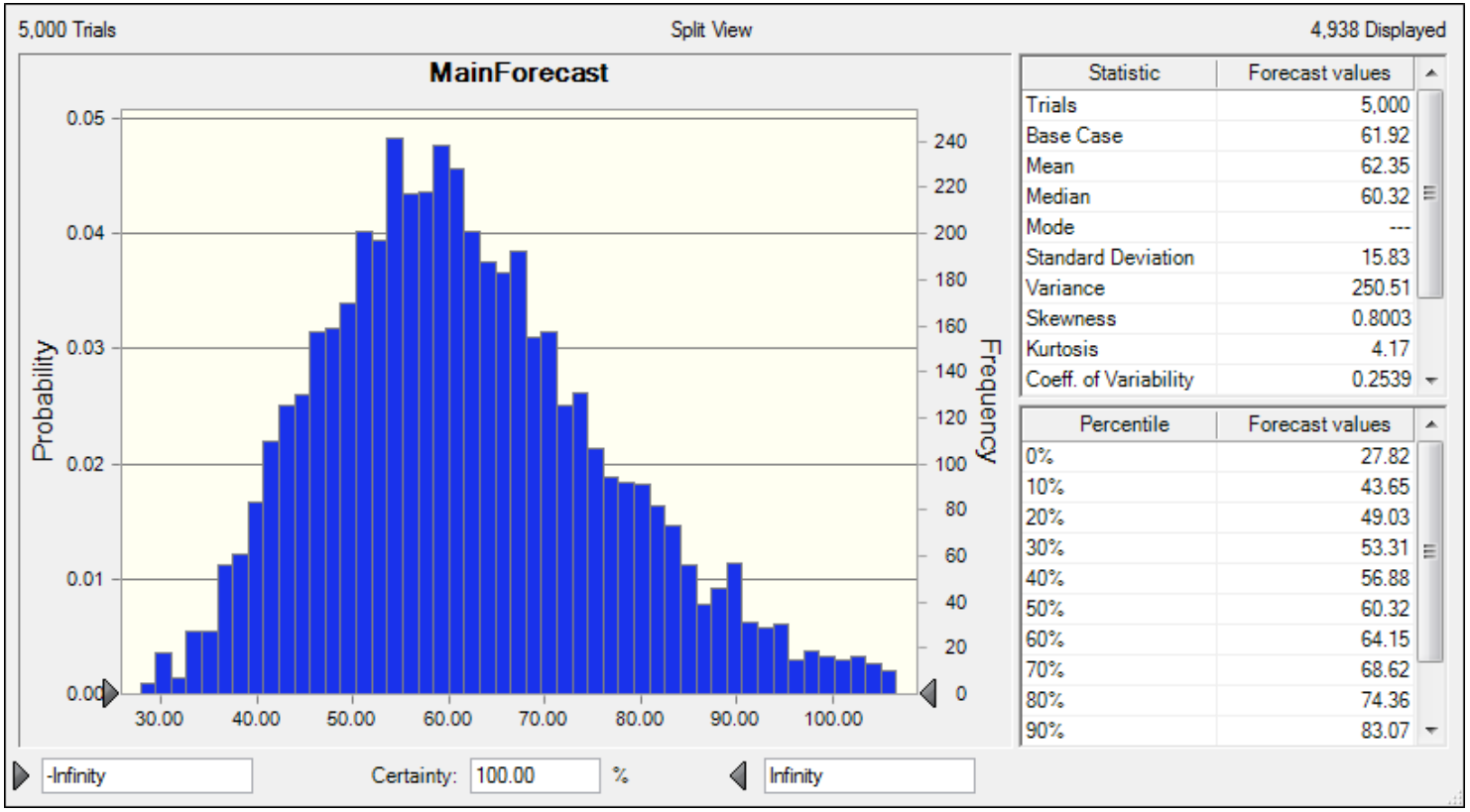
Experimental Trials (cont'd)

- Assumptions
  - The population distribution for each cost element is log normally distributed and is in-turn right skewed
  - The population distribution for each cost element has a CV of 50%
  - Location parameter of the population distributions for each cost element is 0
  - Population correlation between each of the cost elements is +0.2
  - Mean of population distributions is equivalent to the point estimate
  - Modeled distributions always more closely match actual population distributions than default distributions

- Risk Measure
  - The 80th percentile of each experimental trial is compared with the 80th percentile of the aggregate population distribution

Population PDF

- The (approximated) population PDF is shown below:
  - 80th Percentile = $74.36
  - CV = 25.4%



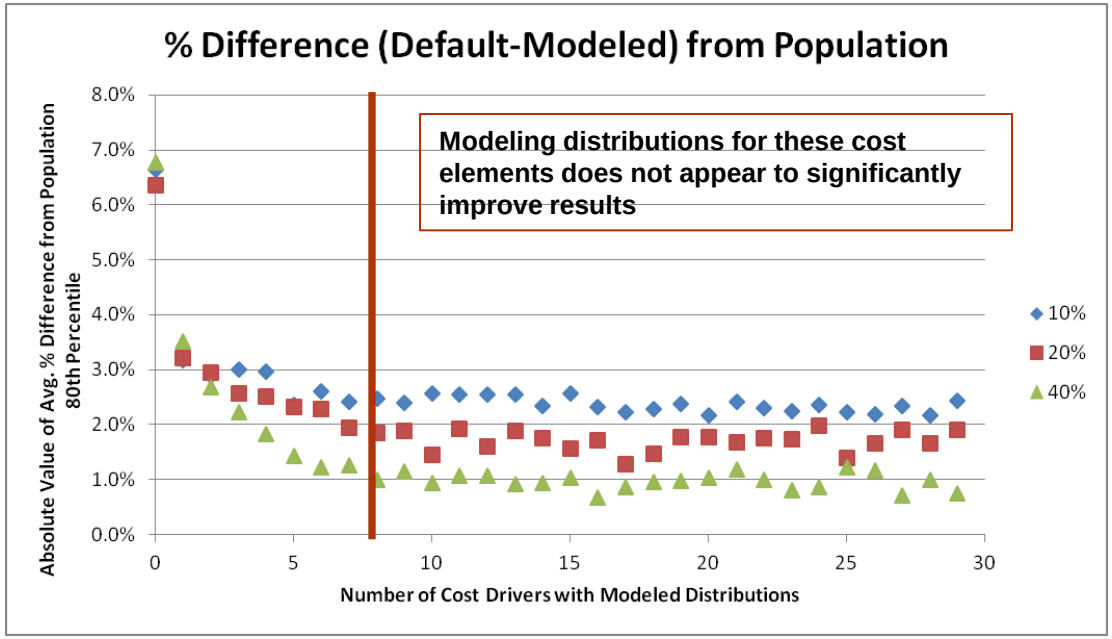
Statistic	Forecast values
Trials	5,000
Base Case	61.92
Mean	62.35
Median	60.32
Mode	---
Standard Deviation	15.83
Variance	250.51
Skewness	0.8003
Kurtosis	4.17
Coeff. of Variability	0.2539

Percentile	Forecast values
0%	27.82
10%	43.65
20%	49.03
30%	53.31
40%	56.88
50%	60.32
60%	64.15
70%	68.62
80%	74.36
90%	83.07

Effect of Adding Modeled Distributions

- The chart below shows the benefits of modeling additional distributions
  - Each data series includes all experimental trials where the delta between the default distribution and modeled distribution deviation from the population distribution is the specified value

- Benefits of adding modeled distributions appear to diminish by approximately the 8th cost driver
  - Equates to 28% of the cost drivers in the estimate

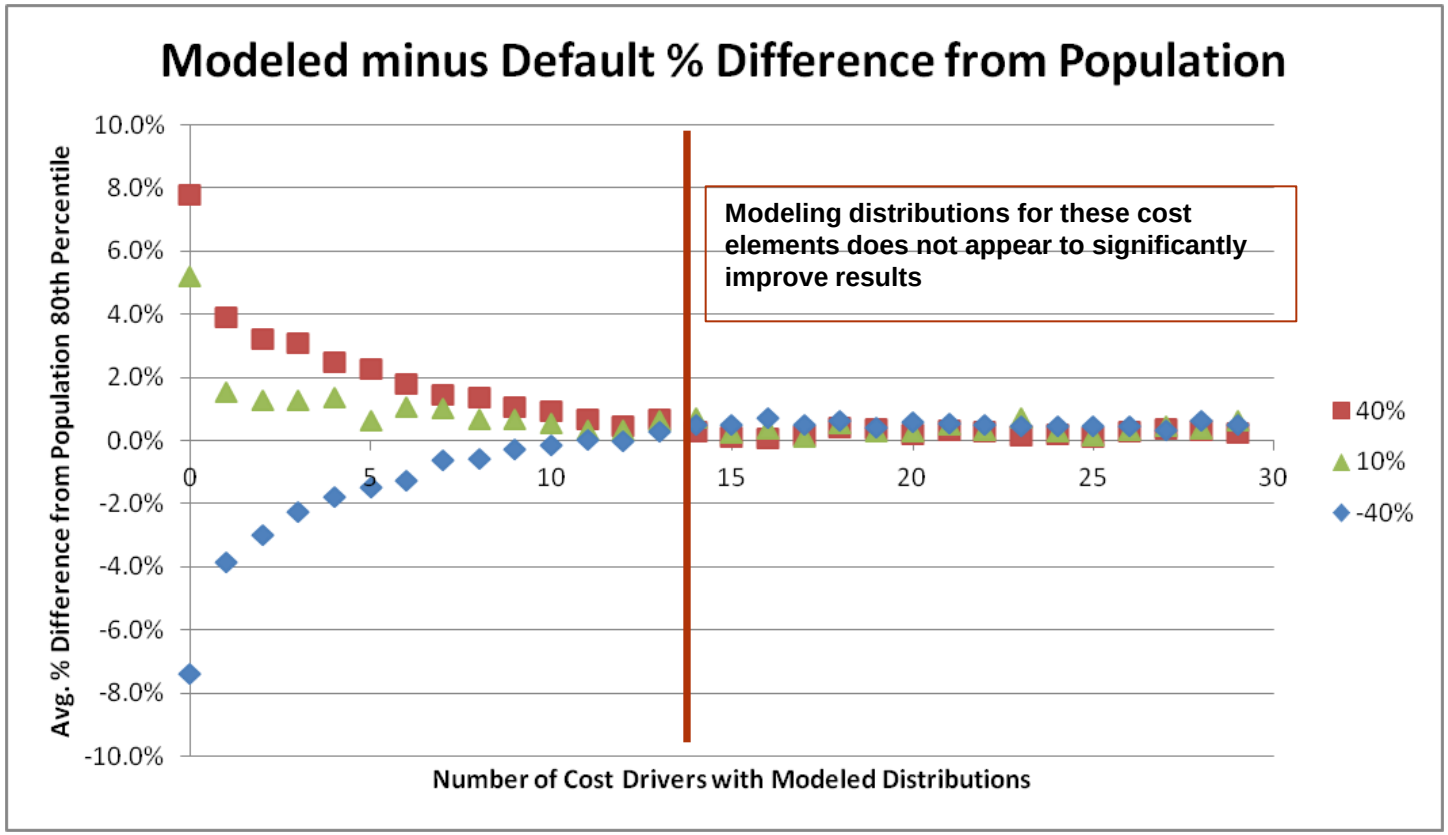


Effect of Adding Modeled Distributions (cont'd)

- The next chart provides a different view of the experimental results
  - The deviation of the default distributions to the population distributions is kept constant in each data series
  - For instance, the data series labeled 40% includes all experimental trials where the default distributions deviate 40% from the population

- In this case, it appears it is beneficial to model approximately the top 12 cost drivers
  - This equates to 42% of the cost drivers in the estimate
  - However, in this particular estimate the top 12 cost drivers still only contain 82% of total cost

Effect of Adding Modeled Distributions (cont'd)



Case Study

- To further validate the results of the iterative analysis and the Monte Carlo simulation, a case study was completed using a sanitized version of another MDA estimate
  - This estimate contains 51 total cost elements

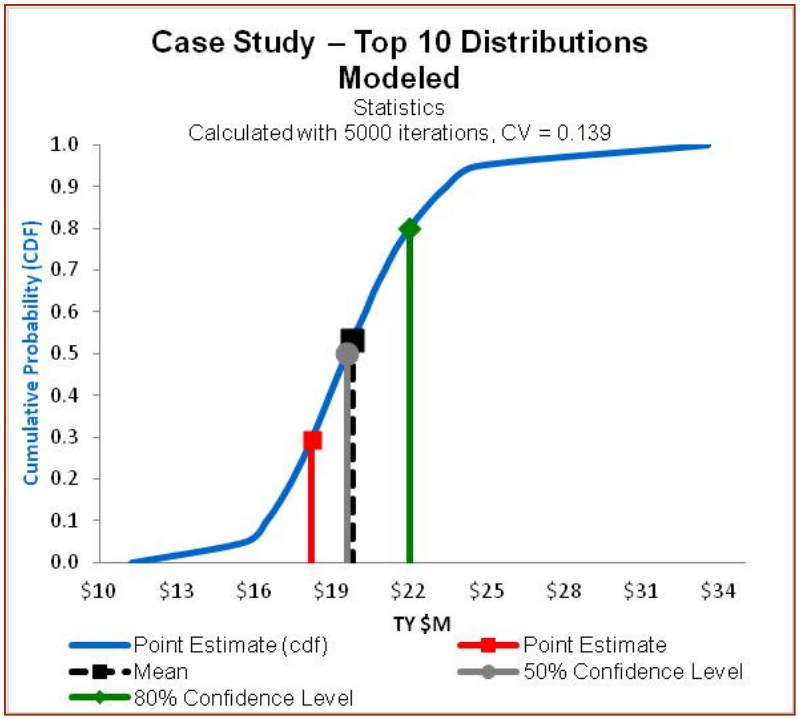
- Two variations of risk and uncertainty analysis were completed on this estimate:
  - **Variation 1:** Modeled distributions were placed on the Top 10 cost elements
  - **Variation 2:** Default subjectively defined distributions were placed on all 51 cost elements

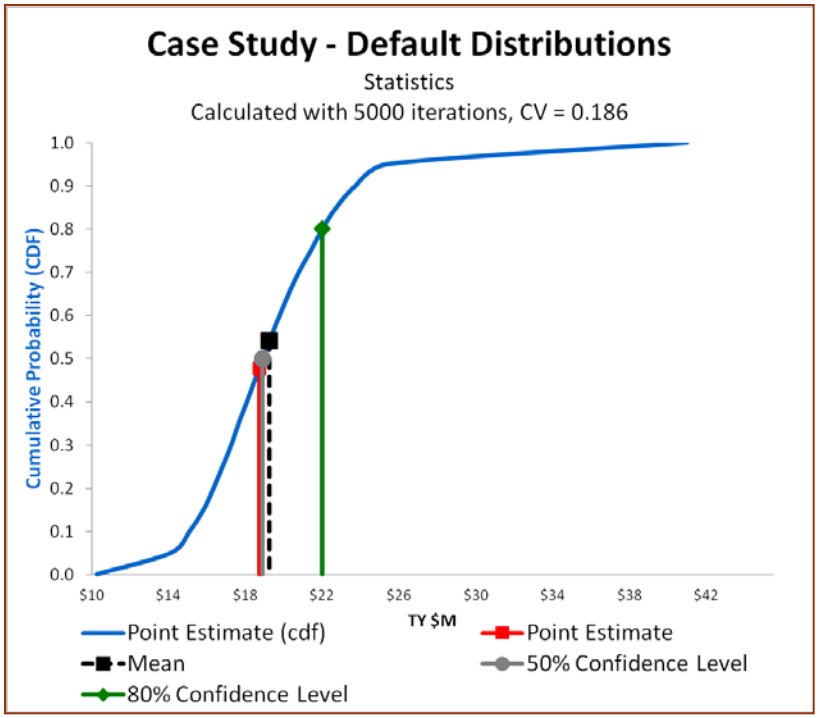
- Correlation of +0.2 was applied between all cost elements
  - Previous studies (Smart, 2009; MDA, 2012) have shown this as a good value to use for default correlation

Case Study – Cumulative Distribution Functions

- The CDF's for both variations of the risk and uncertainty analysis are shown below

- The CV of the risk variation with default distributions (0.186) is greater than the CV of the risk variation with the Top 10 cost elements modeled (0.139), initially suggesting that the variation with the Top 10 cost elements modeled is underestimating risk





Case Study – Risk Statistics

- However, a look at the risk statistics tells a different story

- The 50th percentile of the variation in which the Top 10 distributions are modeled is over 6% higher than the variation in which default distributions are applied to all cost elements
  - The modeled distributions have properly accounted for skewness

Marker	Top 10 Distributions Modeled (TY$M)	Default (TY$M)
Point Estimate	$18.21	$18.21
Mean	$19.80	$18.73
Std. Deviation	$2.74	$3.48
5th Percentile	$15.68	$13.61
10th Percentile	$16.48	$14.58
20th Percentile	$17.49	$15.83
30th Percentile	$18.25	$16.75
40th Percentile	$18.91	$17.56
50th Percentile	$19.58	$18.38
60th Percentile	$20.30	$19.27
70th Percentile	$21.04	$20.26
80th Percentile	$21.99	$21.49
90th Percentile	$23.43	$23.22
95th Percentile	$24.63	$24.89

Conclusions

- The Pareto principle cannot directly be used to complete distribution assignment in cost risk and uncertainty analysis
  - The Top 20% of cost drivers in an estimate may account for less than 80% of the total point estimate
  - Including uncertainty distributions on only the top 20% of cost drivers will likely result in an understatement of variation

Conclusions (cont'd)

- The following guidelines can be used to more efficiently complete input-based cost risk and uncertainty analysis
  - Subjectively defined default distributions should be included on all cost drivers where it is not possible to include modeled distributions
  - There does not appear to be a substantial benefit to continue to include modeled distributions once modeled distributions have been included on the cost drivers which account for approximately 80% of the total cost in an estimate

References

- Missile Defense Agency Cost Estimating and Analysis Directorate (2012). *Cost Estimating Handbook.*
- Scheaffer, R.L. (1995). *Introduction to Probability and its Applications*. Belmont, CA: Duxbury Press.
- Smart, C. (2009). Correlating Work Breakdown Structure Elements. *National Estimator,* Spring 2009, 8-10.
- U.S. Air Force Cost Analysis Agency (2007). *Cost Risk and Uncertainty Analysis Handbook*.