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

James Glenn, Computer Sciences Corporation
Christian Smart, Missile Defense Agency
Hetal Patel, Missile Defense Agency
Lawrence Johnson, Missile Defense Agency

Abstract

Significant effort and statistical knowledge are required in order to complete an accurate uncertainty analysis for a cost estimate. First, a representative sampling of data points must be collected in order to model appropriate distributions for each random variable in the estimate. Additionally, correlation between the random variables in the estimate must be properly assessed and applied. Finally, to generate the cumulative distribution of total costs (i.e., S-Curve) a simulation method such as Monte-Carlo simulation or an analytic approximation such as the method of moments must be used.

However, collecting the data required to accurately model the distribution of each random variable and correlation between the random variables is difficult and time consuming. To help cost estimators develop uncertainty analyses with limited time, a variety of default distributions for random variables and default correlation application techniques between random variables have been proposed by authors and are incorporated in the leading cost estimating software applications. However, an overreliance on default distributions and default values for correlation is likely to provide inaccurate results since it is not a sampling from the actual population of interest.

This is where the Pareto Principle can be applied. The Pareto Principle, named after the Italian economist Vilfredo Pareto, comes from income distribution studies conducted in the 19th century. Pareto found that 80% of income goes to the top 20% of the population. This principle has since been found to have broader application, and has been generalized to state that roughly 80% of effects come from 20% of the causes. For example, 80% of the work done in an office is contributed by only 20% of employees. Applied to cost risk analysis, this would lead to focusing efforts on assigning accurate cost risk distributions to the major cost drivers, while not assigning cost risk to the many WBS elements that only minimally drive cost. The purpose of this paper is to examine the potential effects of working to develop cost risk distributions for major cost drivers, and compare this to the overall effect of assigning cost risk distributions to all elements with default distributions. Assumptions are varied for the proportion of random variables that receive modeled distributions and the proportion of random variables that are assigned default distributions. Also, assumptions are varied for how closely both the modeled distributions and default distributions match the actual population mean, variance, and skewness. Finally, observations are provided to help the cost community manage their limited resources to generate improved uncertainty analyses in cost estimates.

Introduction

One of the most important steps in creating a credible cost estimate is completion of a thorough cost risk and uncertainty analysis. However, in practice cost estimators often find themselves with insufficient time and resources needed to complete cost risk and uncertainty analysis to allow an in-depth analysis. In fast-paced program office environments, quick-turn estimates often have to be completed. When working quick-turn estimates, it can be challenging enough to generate point cost estimates in the short amount of time that is granted, let alone complete any risk and uncertainty analysis.

Program managers tend to be optimistic and believe that their programs will not experience the same obstacles as prior programs, which is a primary reason for the importance of a sound cost uncertainty analysis . Output-based risk and uncertainty analysis can be used when risk and uncertainty analysis must be completed in a short amount of time. A number of different output-based approaches have been proposed, including those by Smart (2011a, 2011b) and Braxton et al. (2011). While these approaches are very useful, particularly when developing independent cost estimates, or early, Milestone A estimates, greater fidelity is required for program office estimates in order to communicate risk to program managers and to allocate total risk to individual WBS elements for inclusion in budgets. As a result, it is useful to determine if input-based cost risk and uncertainty analysis can be completed on an estimate with only a fraction of the work that is required to develop detailed cost risk estimates for each WBS element.

The Pareto Principle is named after the Italian economist Vilfredo Pareto. In his income studies conducted in the 19th century, Pareto found that 80% of income goes to the top 20% of the population. This principle has since been found to have broader application, and has been generalized to state that roughly 80% of effects come from 20% of the causes. For example, 80% of the work done in an office is contributed by only 20% of employees. Applied to the context of input-based cost risk and uncertainty analysis, this suggests that it may be possible to place uncertainty distributions on only 20% of cost drivers and still explain 80% of the potential variation in costs. This paper will investigate the Pareto Principle and its potential application to input-based cost risk and uncertainty analysis.

Distribution Assignment

The first step in input-based cost risk and uncertainty analysis is distribution assignment. The Air Force Cost Risk and Uncertainty Analysis Handbook (U.S. Air Force, 2007) categorizes distribution assignment into subjective and objective distribution assignment.

Objective Distribution Assignment

Objective distribution assignment refers to analyzing data to characterize the uncertainty associated with a cost driver. Objective distribution assignment includes:

- Computation of prediction intervals from parametric CERs.

- Input-modeling of an appropriate distribution to describe a dataset using goodness of fit tests such as Chi-Squared or Kolmogorov-Smirnov .

Subjective Distribution Assignment

Subjective distribution assignment refers to assigning an uncertainty distribution to a cost driver without analyzing data. Subjective distribution assignment includes:

- Use of expert opinion to define distribution minimum and maximum values and distribution type.
- Use of default subjective distributions which includes rules of thumb to describe the expected amount of variation (i.e., low variation = 15% CV; medium variation = 25% CV, etc.) (commonly used cost estimating software applications include default subjective distributions to make distribution assignment very easy if no other information is available).

Benefits to Each Approach

It is reasonable to expect that objective distribution assignment will produce better results than subjective distribution assignment, provided that enough data points are available and the analysis has been done correctly. However, objective distribution assignment can be substantially more time consuming. Obtaining a sufficient number of data points to derive a parametric CER or fit a valid distribution around can be the most challenging part.

It is also reasonable to expect that subjective distributions derived using expert opinion will produce better results than default subjective distributions, provided that the expert has provided a basis for the values (for example, design criteria or actual value from a previous development effort). Additionally, it may not be possible to objectively define distributions for all cost drivers leaving expert opinion as the best available option.

The following terms are defined for use in the remainder of this paper:

- Modeled Distribution - objectively defined distributions or subjectively defined distributions using expert opinion. All of these distribution assignment approaches are referred to as modeled distributions since a basis can be provided to defend at least one of the distribution parameters.
- Default Distribution - default subjectively defined distributions.

Sequential Analysis

The investigation begins by sequentially applying uncertainty distributions to the top cost elements in an estimate. Typically, uncertainty distributions are applied at the input variable level but for simplicity in this example only the cost elements are investigated.

Example Estimate

The estimate under investigation is a sanitized version of an actual estimate. A Pareto chart of the cost elements in this estimate is displayed in Figure 1.

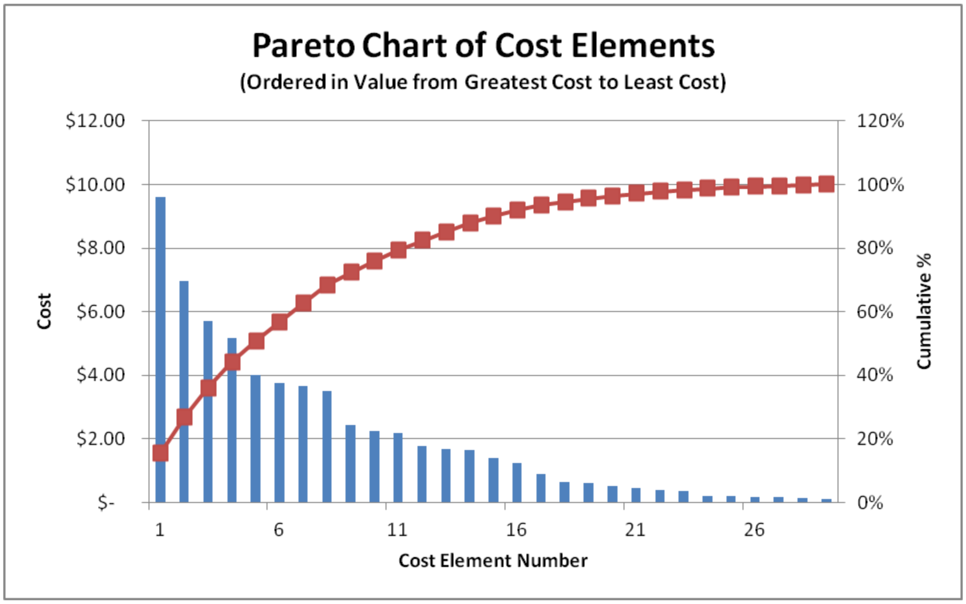


Figure 1: Pareto Chart

This example cost estimate has 29 total cost elements. If the Pareto Principle applies to this estimate, it would suggest that applying uncertainty distributions to the top six cost elements (21% of the 29 total cost elements) should explain roughly 80% of the variation. However, several observations can be made from the Pareto chart.

- The top six cost elements only comprise 57% of total cost in the point estimate.
- The top twelve cost elements comprise 82% of total cost in the point estimate.

Whether applying uncertainty distributions to the top six cost elements adequately explains the overall uncertainty in the estimate is addressed next.

Results

The following assumptions are made to analyze the effect of sequentially adding uncertainty distributions to the top elements in the example estimate:

- The included uncertainty distributions match the actual population distribution.

- The included uncertainty distributions have a Coefficient of Variation (CV) of 50%. Note that the CV is ratio of the standard deviation to the mean, and thus represents the relative amount of uncertainty for the estimate.

- Correlation of +0.2 is applied to included uncertainty distributions. This correlation matches the actual population correlation.

When sequentially adding the uncertainty distributions to each cost element, the resulting standard deviation of the total aggregate distribution is compared with the standard deviation of the total aggregate distribution with uncertainty distributions and correlation applied between all of the cost elements. Computation of these results does not require simulation as the formulas in Figure 2 can be applied.

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

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

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

where U is the aggregate distribution and

Y_1 ,......, Y_n are random variables

Figure 2: Sums of Means and Variances

Figure 3 shows the results of sequentially adding uncertainty distributions to each cost element in the example estimate.

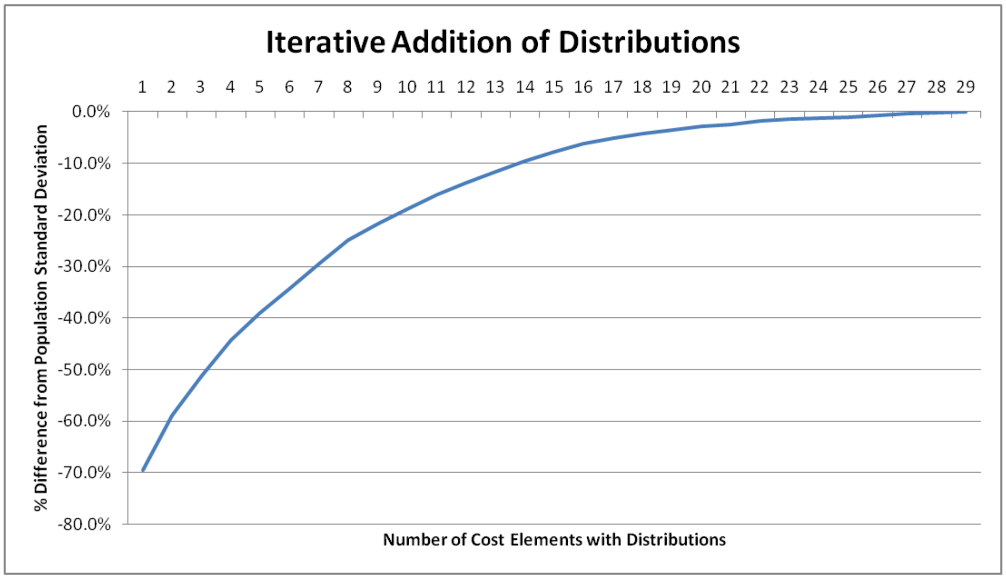


Figure 3: Results of Sequentially Adding Uncertainty Distributions

The following observations are drawn from Figure 3:

- Applying uncertainty distributions to the top six (21%) of cost elements only accounts for approximately 70% of the total variation in the estimate. The Pareto Principle does not apply in this example. Thus, the more general conclusion applies that the Pareto Principle cannot be used to assigned risk to only the top 20% of cost drivers while expecting to estimate 80% of the risk.

- Uncertainty distributions must be applied to approximately 10 cost elements (or approximately 30% of the cost elements) to account for 80% of the variation.

- Even though the marginal amount of variation explained decreases with the addition of each uncertainty distribution, there is no point of inflection which indicates where it no longer makes sense to include additional distributions.

While the analysis discussed above provides insight, the following limitations exist:

- It is highly unlikely that uncertainty distributions included in a risk and uncertainty analysis will match the actual population distribution.

- Only applying uncertainty distributions to the top cost elements in an estimate will likely result in an underestimate of variation.

As a result, the next analysis that will be completed investigates the effect of including modeled distributions on some cost elements and default distributions on the remaining cost elements with the assumption that the included distributions do not match the actual population. Including distributions on all cost elements will reduce the natural tendency to understate variation when distributions are only included on the top cost elements.

Experimental Trials Using Monte-Carlo Simulation

Visual Basic for Applications (VBA) and Monte-Carlo simulation was used to run a large number of experimental trials to analyze the effect of simultaneously including modeled distributions and default distributions. The example cost estimate used is the same as the one in the previous section. The 12,000 total experimental trials which were computed are described in Table 1. For each experimental trial, 5,000 replications were completed.

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%

Table 1: Experimental Trials Run

The assumptions underlying this analysis are:

- Modeled distributions are more accurate than default distributions. As a result, only the experimental trials where the modeled distributions more closely match the assumed population distribution are kept for further analysis. Additionally, the X modeled distributions that are included are applied to the X top cost elements.

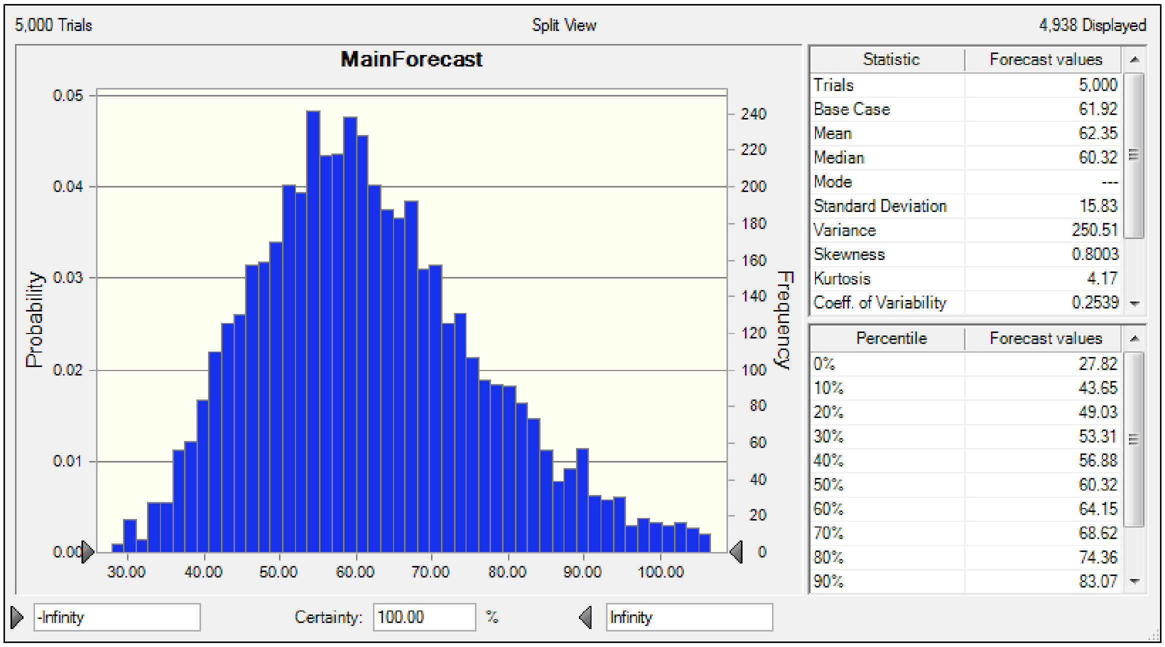
- The population distribution for each cost element is lognormally distributed and is therefore right skewed.

- The location parameter of the population distributions for each cost element is 0.

- The population distribution for each cost element has a CV of 50%.

- The population correlation between each of the cost elements is +0.2.

In order to determine the accuracy of each experimental trial, the 80th percentile of each trial is compared to the 80th percentile of the population density. The total aggregate population probability distribution function, as approximated with a Monte Carlo simulation, is displayed in Figure 4. The approximate population 80th percentile is 74.36.



5,000 Trials Split View 4,938 Displayed

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

-Infinity Certainty: 100.00 % Infinity

Figure 4: Aggregate Population PDF

Results

Figure 5 shows the effect of including modeled distributions with the remaining cost elements receiving default distributions. Each data series includes all experimental trials where the delta between default distributions and modeled distributions deviation from the population distributions is the specified value. For instance, the data series labeled "10%" includes all experimental trials where:

- Default distribution deviation = 50% and Modeled distribution deviation = 40%.
- Default distribution deviation = 40% and Modeled distribution deviation = 30%.
- And all other possible combinations yielding a 10% delta.

Including modeled distributions clearly appears to improve results when added to the first several cost elements, but the benefits appear to become negligible after approximately the 8th cost element (or 28% of the total cost elements in the estimate).

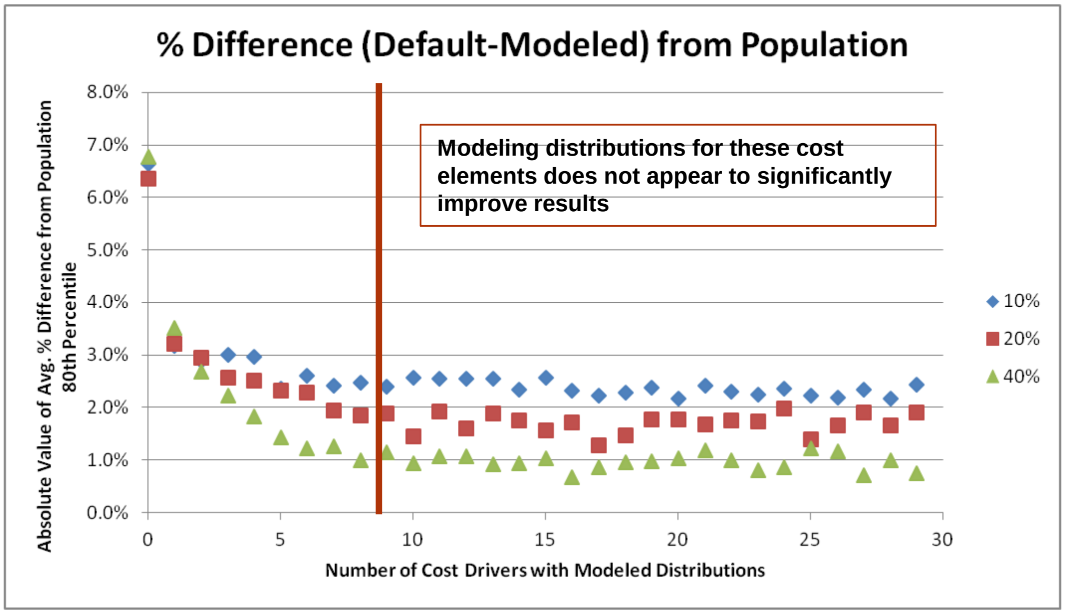


Figure 5: Results for Specified Deltas Between Default and Modeled Distributions

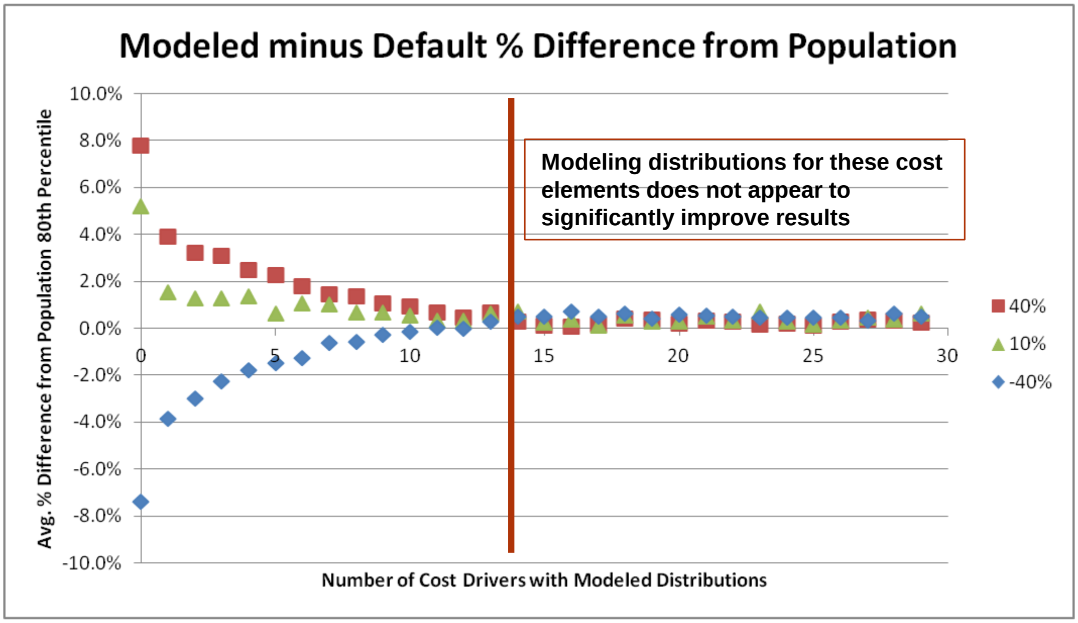


Figure 6: Constant Default % Difference to Population

Figure 6 provides a different view of all the experimental trials. Each data series shows all experimental trials for a constant default % deviation from the population. In this case, including additional modeled distributions still appears to improve results up until approximately the 12th cost element. It is important to point out that that the top 12 cost elements were required to account for over 80% of total cost in the point estimate.

In summary, the experimental trials suggest that when default distributions are added to the cost elements which do not receive modeled distributions a point can be reached where adding additional modeled distributions does not dramatically improve results.

Case Study

To further validate the results of the previous two sections, two different variations of risk and uncertainty analysis were completed on a sanitized version of another actual program office cost estimate. The two risk and uncertainty analysis variations are summarized below:

- **Variation 1:** Modeled distributions were placed on the Top 10 input variables which affect cost (equates to roughly 20% of the total input variables in the estimate). No uncertainty was assigned to any other inputs variables.

- **Variation 2:** Default subjectively defined distributions were placed on all input variables in the estimate.

In both of the variations, subjective correlation of +0.2 was applied between all of the input variables with uncertainty distributions. Previous studies (Smart, 2009; MDA, 2012) have identified this as a safe value to use when it is not feasible to objectively assess correlation.

The cumulative distribution functions (or S-Curves) of these two risk and uncertainty analysis variations are shown in Figure 7. The CV for Variation 2 is equal to 18.6% while the CV for Variation 1 is significantly lower at 13.9%. Thus it appears that Variation 1 may be underestimating risk by not including risk for all WBS elements.

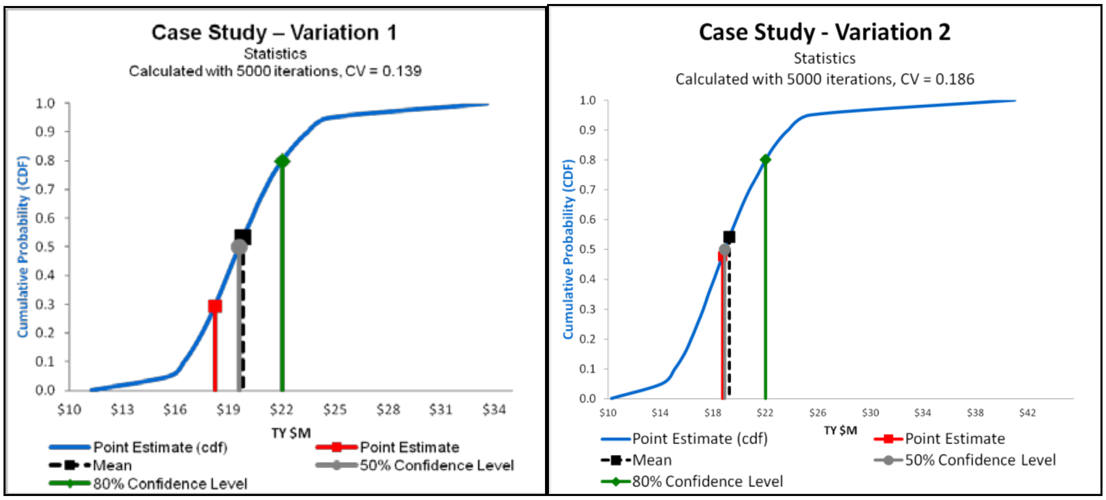
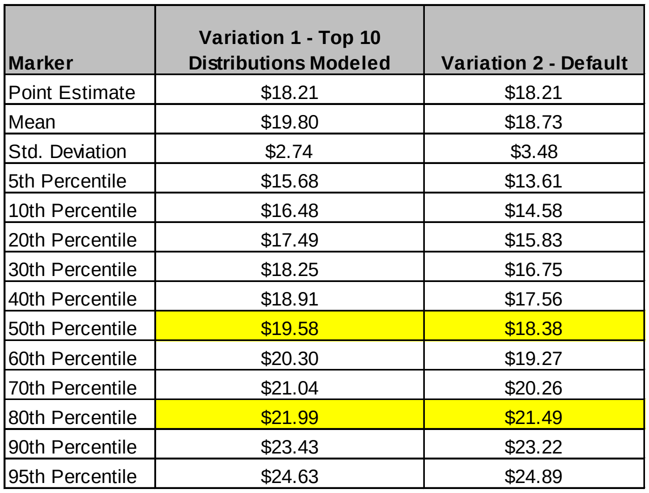


Figure 7: ***CDF Comparison***

However, review of the risk statistics for the two risk and uncertainty analysis variations tells a different story (Table 2). Both the 50th and 80th percentiles for Variation 1 are greater, even though the CV is lower than in Variation 2. In particular, the 50th percentile for Variation 1 is over 6% greater than the 50th percentile for Variation 2. The explanation for this is that the modeled distributions better account for skewness in the dataset. In summary, it appears that modeling the uncertainty with the Top 20% of cost drivers (or approximately Top 10 in this case) is much more critical than including uncertainty on the remaining cost elements.



Marker	Variation 1 - Top 10 Distributions Modeled	Variation 2 - Default
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

Table 2: Summary of Risk Statistics

Variation 1 is more accurate, while Variation 2 captures more uncertainty. Thus for the case study the best overall solution may be to model the top drivers in-depth, while incorporating default uncertainty on the remaining variables.

Conclusions

The results in this paper show that the Pareto Principle cannot directly be used to complete distribution assignment for cost risk and uncertainty analysis for the following reasons:

- The top 20% of cost drivers in an estimate may account for less than 80% of the total point estimate value.
- Only including uncertainty distributions on the top 20% of cost drivers will likely result in an understatement of variation. This is not a desired outcome in a profession that is more often known for underestimating costs rather than overestimating costs.

However, the results in this paper do identify guidelines which can be used to help cost estimators more efficiently complete an input-based cost risk and uncertainty analysis:

- Subjectively defined default distributions should be included on all cost drivers which do not receive modeled distributions. Subjectively defined default distributions can very quickly be applied to cost drivers in commonly used cost estimating software applications.

- Once modeled uncertainty distributions are applied to the cost drivers which account for approximately 80% of the total cost in the estimate, there does not appear to be a significant benefit to modeling additional distributions. This observation holds true as long as default distributions are included on the remaining cost drivers.

References

Braxton, P.J., Lee, R.C. and Coleman, R.L. (2011). The NCCA S-Curve Tool, presented to the Hampton Roads chapter of SCEA.

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.

Smart, C. (2011a). Covered with Oil: Incorporating Realism in Cost Risk Analysis, presented at the 2011 Joint Annual ISPA-SCEA Conference.

Smart, C. (2011b). Quick-turn Risk Analysis, presented to the Greater Alabama chapter of SCEA.

U.S. Air Force Cost Analysis Agency (2007). *Cost Risk and Uncertainty Analysis Handbook*.