# Robust Default Correlation for Cost Risk Analysis

Christian Smart, Ph.D., CCEA

Director, Cost Estimating and Analysis

Missile Defense Agency

Presented at the 2013 ICEAA Professional Development and Training Workshop

June, 2013

# Introduction

- **Correlation is an important consideration in cost risk analysis**
- **When correlation is ignored, you are making the de facto assumption that all risks are independent**
  - **"Even when you choose not to decide, you still have made a choice" (Rush, *Free Will*)**
- **Assuming no correlation results in a vast understatement of risk**
- **In 1996, Don Mackenzie wrote that "One of the more difficult chores in cost risk analysis is establishing appropriate levels of correlation… " (Mackenzie 1996)**
  - **Seventeen years later, this is still true**
- **This presentation is an attempt at making forward progress on this issue**

# Definitions

- **Consider two random variables, *X* and *Y*.**
- **The mean of X, E(X), is denoted by $\mu_x$, and similarly, the mean of Y, E(Y), is denoted by $\mu_y$**
- **The variance of X, Var(X), is denoted by $\sigma_X^2$, and similarly, the variance of Y, Var(Y), is denoted by $\sigma_Y^2$**
- **The variance of X and Y are equal to:**

$$Var(X) = Cov(X,X) = E(X^2) - [E(X)]^2$$

$$Var(Y) = Cov(Y,Y) = E(Y^2) - [E(Y)]^2$$

- **Correlation, denoted by the Greek letter *r* ("rho"), is defined by**

$$\rho_{XY} = Corr(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

# Total System Mean and Variance

- **For *n* WBS elements, the mean and the variance of the total cost are defined by:**

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu_i$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=}^{n} \sigma_i^2 + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} \rho_{ij}\sigma_i\sigma_j$$

# Total Variance with Level Correlation

- **Suppose (for simplicity)**
  - **There are *n* WBS Elements** $C_1, C_2, \ldots, C_n$
  - **Each** $Var(C_i) = \sigma^2$
  - **Each** $Corr(C_i, C_j) = \rho < 1$
  - **Total Cost** $C = \sum_{k=1}^{n} C_i$

$$Var(C) = \sum_{k=1}^{n} Var(C_i) + 2\rho \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sqrt{Var(C_i)\ Var(C_j)}$$

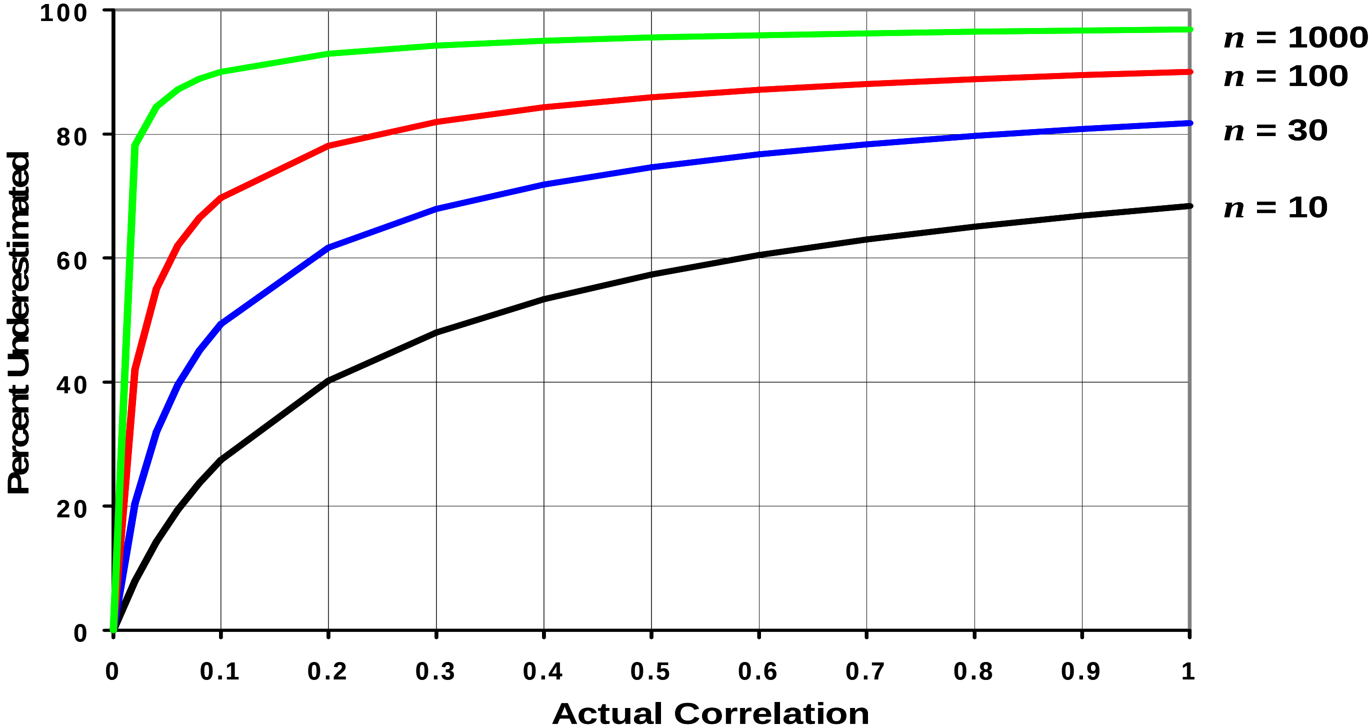
$$= n\sigma^2 + n(n-1)\rho\sigma^2$$

$$= n\sigma^2(1+(n-1)\rho)$$

| Correlation | 0 | ρ | 1 |
|---|---|---|---|
| $Var(C)$ | $n\sigma^2$ | $n\sigma^2(1+(n-1)\rho)$ | $n^2\sigma^2$ |

# Impact of Assuming Independence

- **For a 100 element WBS assuming independence among all WBS elements when the true underlying correlation is equal to 20% results in an underestimate of total system standard deviation equal to 80%!**




**Source: "Why Correlation Matters in Cost Estimating," Advanced Training Session, 32nd Annual DOD Cost Analysis Symposium, Williamsburg, VA, 1999.**

# Example of Impact

- **As an example, consider a system with 10 subsystems, each with mean equal to \$10 million and standard deviation equal to \$3 million**
- **For 100 elements and 1,000 elements, assuming correlation is zero when it is actually 20% results in underestimating the 80th percentile by 8-10%, and if the correlation is 60%, the 80th percentile is underestimated by 15-17%**

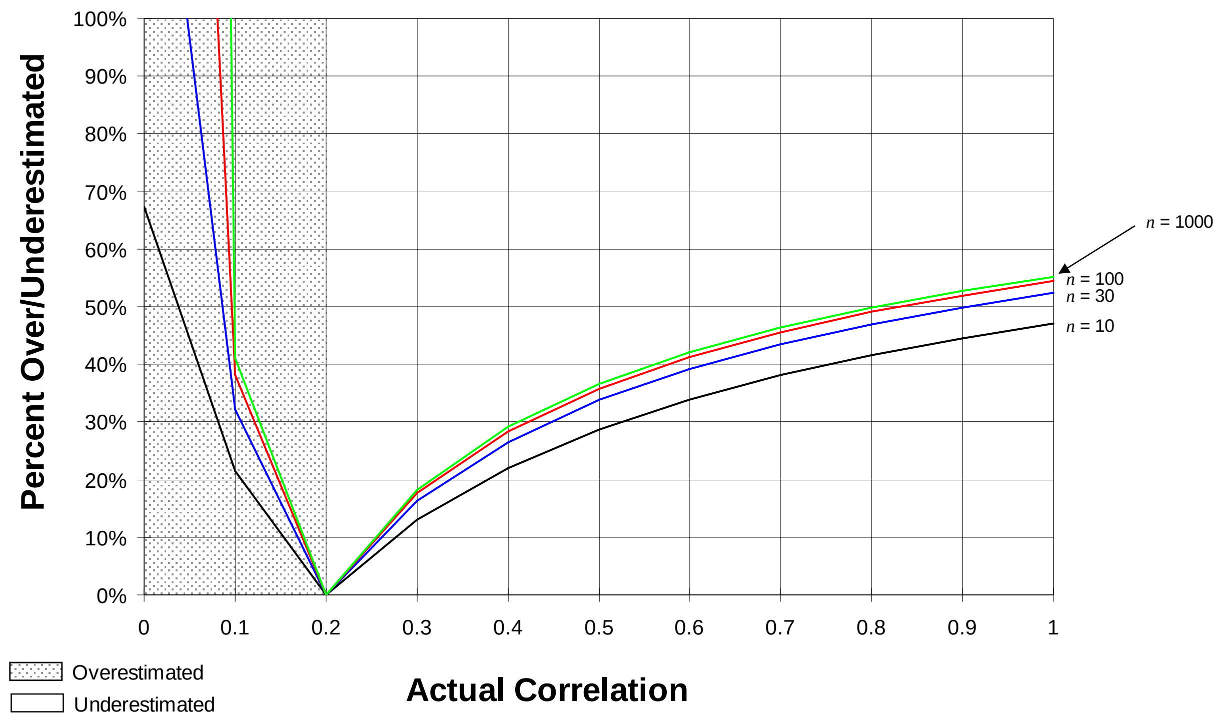
| | 80% Confidence Level (TY$, Millions) | | |
|---|---|---|---|
| Number of WBS Elements | Independence | 20% Correlation | 60% Correlation |
| 10 | $108 | $113 | $119 |
| 100 | $1,025 | $1,111 | $1,182 |
| 1,000 | $10,080 | $11,092 | $11,815 |

# Default Correlation

- **Notice in the graph on the previous chart there is an apparent "knee" in the curve around 20%**
  - **Above 20% correlation the consequence of assuming less correlation begins to dwindle**
  - **This graph is the basis for assuming 20-30% for default correlation for elements between which there is no functional correlation**
  - **Book (Book 1999) recommends 20% as a default correlation value because of this**
- **However, the graph does not tell us how much the total standard deviation is underestimated because correlation is assumed to be 20%, but is actually 60%, for example**

# Underestimating Correlation with the Default 20%

- **For a 100-element WBS, if the correlation is assumed to be 20% but is actually 60%, the total standard deviation is underestimated by 40%**




Source: "Why Correlation Matters in Cost Estimating," Advanced Training Session, 32nd Annual DOD Cost Analysis Symposium, Williamsburg, VA, 1999.

# Robust Approach

- **A more robust approach to assigning correlations would be to use the value that results in the least amount of error in the variance**
  - **It is robust in the sense that without solid evidence to assign a correlation value, it minimizes the amount by which the total standard is misestimated due to the correlation assumption**
  - **This robust default measure of correlation would be a value for correlation that would minimize the error when the assumed correlation differs from the actual underlying correlation**

# Absolute Error

- **We are interested in the absolute value of the error, since if we consider negative and positive values, they may offset each other**
- **Let ε denote the error, then we are interested in |ε|, where |ε| is defined by**

$$|\varepsilon| = \begin{cases} \varepsilon \text{ if } \varepsilon > 0 \\ -\varepsilon \text{ if } \varepsilon < 0 \end{cases}$$

# Expected Value of Absolute Error

- **If we assume that the prior distribution of correlation on the interval (0,1) is uniform, then the expected value of the absolute error |ε| of the variance as a function of the assumed correlation is defined by**

$$\int_0^1 |\varepsilon|\, f(\rho)\,d\rho = \int_0^1 |\varepsilon|\,d\rho$$

**since** $f(\rho) = 1$

- **Thus the approach is to find the value of $\rho$ that minimizes the expected (absolute) error**
- **This equation provides the expected error as a function of $\rho$, and then we minimize this function with respect to $\rho$ using techniques from elementary Calculus**

# What is the Error?

- **Now that we have determined how to determine the minimum error, we need to figure out what to minimize**
- **We present several different choices, calculate the results, and provide pros and cons for each**

# Case 1: Percentage Error (of Actual)

- **Denote the assumed correlation by $\rho_1$ and the actual correlation by $\rho_2$**
- **In this first case, we consider the metric that Book (Book, 1999) looked at when measuring over- and under-estimation of correlation, which is to consider the percentage error in variance as a percentage of the actual correlation**

$$\varepsilon = \frac{\sqrt{n}\sigma\sqrt{1+(n-1)\rho_2} - \sqrt{n}\sigma\sqrt{1+(n-1)\rho_1}}{\sqrt{n}\sigma\sqrt{1+(n-1)\rho_2}} = \frac{\sqrt{1+(n-1)\rho_2} - \sqrt{1+(n-1)\rho_1}}{\sqrt{1+(n-1)\rho_2}}$$

# Case 1: Calculating Expected Absolute Error (1 of 2)

- **The expected absolute error is calculated* as**

$$\int_0^{\rho_1} \frac{\sqrt{1+(n-1)\rho_1}-\sqrt{1+(n-1)\rho_2}}{\sqrt{1+(n-1)\rho_2}} d\rho_2 + \int_{\rho_1}^{1} \frac{\sqrt{1+(n-1)\rho_2}-\sqrt{1+(n-1)\rho_1}}{\sqrt{1+(n-1)\rho_2}} d\rho_2$$

$$= 2\rho_1 + \frac{4}{n-1} - \frac{2}{n-1}\sqrt{1+(n-1)\rho_1}\left(1+\sqrt{n}\right)+1$$

- **This is a function of the number of WBS elements (*n*) and the assumed correlation ($\rho_1$)**
- **Minimizing this with respect to $\rho_1$ we find that**

$$\rho_1 = \frac{\left(1+\sqrt{n}\right)^2 - 4}{4(n-1)}$$

***See the paper for detailed calculations**

# Case 1: Calculating Expected Absolute Error (2 of 2)

- **The limit of this minimum as $n \to \infty$ is 25%**
- **This is close to the 20% default value advocated by Book (Book, 1999)**
- **However, the total error is minimized by this value because of the large penalty assigned when overestimating actual correlations near zero**
- **For example, let *n= 100* and assume the correlation is 40%. The absolute percentage error when the actual correlation is equal to zero is 537%, while the absolute percentage error when the actual correlation is equal to 80% is only 29%**
- **The penalty should not differ greatly whether you are overestimating or underestimating**
  - **An easy way to overcome this issue is to examine the percent error as a function of the assumed correlation, which is considered in Case 2**

# Case 2: Percentage Error (of Assumed) (1 of 2)

- **This case is similar to Case 1, only the denominator is different**

$$\varepsilon = \frac{\sqrt{n}\sigma\sqrt{1+(n-1)\rho_2} - \sqrt{n}\sigma\sqrt{1+(n-1)\rho_1}}{\sqrt{n}\sigma\sqrt{1+(n-1)\rho_1}} = \frac{\sqrt{1+(n-1)\rho_2} - \sqrt{1+(n-1)\rho_1}}{\sqrt{1+(n-1)\rho_1}}$$

- **In this case,**

$$E(|\varepsilon|) = 2\rho_1 - \frac{4}{3(n-1)}\left(1+(n-1)\rho_1\right) + \frac{2}{3(n-1)}\left(1+n^{\frac{3}{2}}\right)\left(1+(n-1)\rho_1\right)^{-\frac{1}{2}} - 1$$

# Case 2: Percentage Error (of Assumed) (2 of 2)

- **The value of $\rho_1$ that minimizes the expected (absolute) error is**

$$\rho_1 = \frac{\left(\frac{1+n^{\frac{3}{2}}}{2}\right)^{\frac{2}{3}} - 1}{n-1}$$

- **The limit of $\rho_1$ as $n \to \infty$ is**

$$\left(\frac{1}{2}\right)^{\frac{2}{3}} \approx 63\%$$

# Impact of Case 2

- **The single recommended value from this approach is 63%**
  - **This is much larger than the 25% value using the other approach, or the 20% rule of thumb widely used in practice**

- **The impact on standard deviation in increasing default correlation from 20% to 63% will result in a significant increase in standard deviation**

| *n* | *% Increase in* $\sigma$ |
|---|---|
| 10 | 54.3% |
| 30 | 68.3% |
| 100 | 74.5% |
| 1,000 | 77.2% |
| 10,000 | 77.5% |

# Case 3: Total Absolute Difference

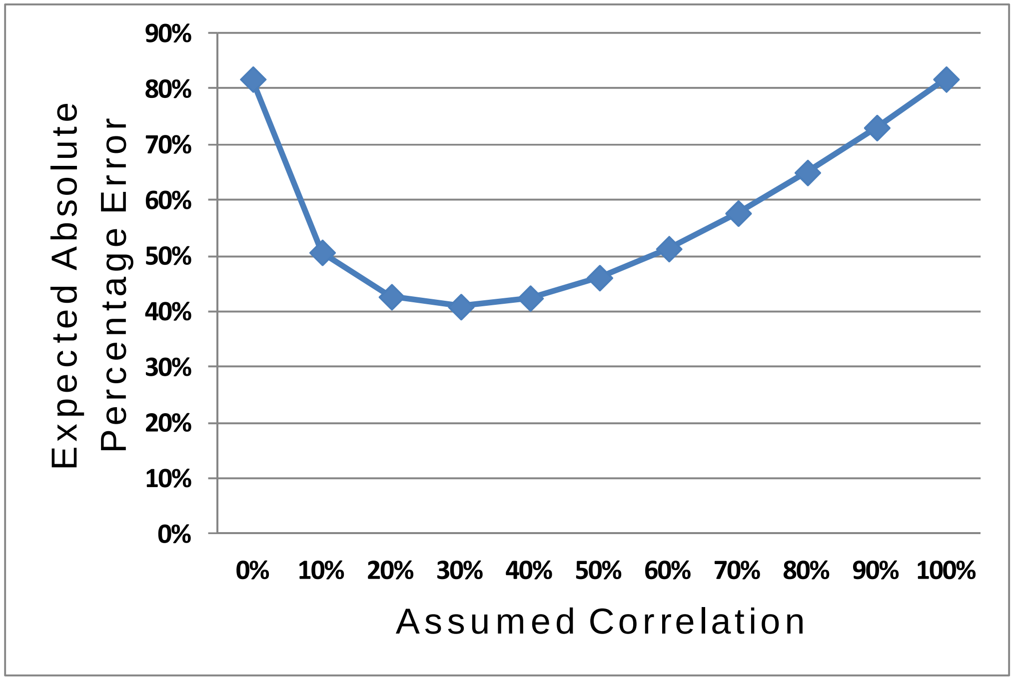
- **The absolute difference could also be considered as a metric**

$$\sqrt{n}\sigma\sqrt{1+(n-1)\rho_2} - \sqrt{n}\sigma\sqrt{1+(n-1)\rho_1}$$

- **In this case, the absolute expected value of the error occurs when $\rho_1$ = 50%**

# Case 4: Case 1 with Truncated Limits (1 of 2)

- **If we consider the first case, much of the reason why the minimum is so low compared to the other cases is the error when the actual correlation is close to 0%**
  - **We know that in most case the correlation is not 0%, and we know that it is not 100%**
- **Absolute percentage error for variance as a percent of the actual correlation for 100 WBS elements:**



# Case 4: Case 1 with Truncated Limits (1 of 2)

- **If we truncate the actual correlation to be uniform in the interval (0.1,0.9) then the expected value of the absolute percent error is minimized when**

$$\rho_1 = \frac{\left( (0.1n+0.9)^{\frac{1}{2}} + (0.9n+.1)^{\frac{1}{2}} \right)^2 - 4}{4(n-1)}$$

- **The limit of this as $n \to \infty$ is 40%**

# Summary of the Four Cases

- **All four cases minimize the expected value of the absolute error in the variance, but use different metrics for measuring error**
- **Case 1:**
  - **Error is measured as a percentage of the variance that results from the actual correlation, result in the limit is 25%**
- **Case 2:**
  - **Error is measured as a percentage of the variance that results from the assumed correlation, result in the limit is 63%**
- **Case 3:**
  - **Error is measured as total difference in variances, result is 50%**
- **Case 4:**
  - **Error is measured as a percentage of the variance that results from the actual correlation, with the correlation range limited to 10-90%; result is 40%**

# Recommendation

- **I recommend a percentage difference approach**
  - **Knowing that the difference between the estimated total standard deviation and the actual total standard deviation is $100 million doesn't tell you much, since it could be large if the standard deviation is $100 million, or relatively small if the total standard deviation is $1 billion**
- **Calculating the error based as a percentage of the assumed correlation is logical**
  - **The issue with looking at the error relative to the actual correlation is that we don't know the actual correlation - we only know the assumed correlation.**
  - **The same is true for CER residuals**
    - **For the Minimum Unbiased Percent Error (MUPE) and the Zero bias Minimum Percent Error (ZMPE) CER methods look at the percentage error from the estimate, not from the actual**
    - **We should use the same metric in looking at correlation**
- **Bottom line: I recommend using a default value for correlation that is equal to 63%**

# Empirical Evidence for Correlation

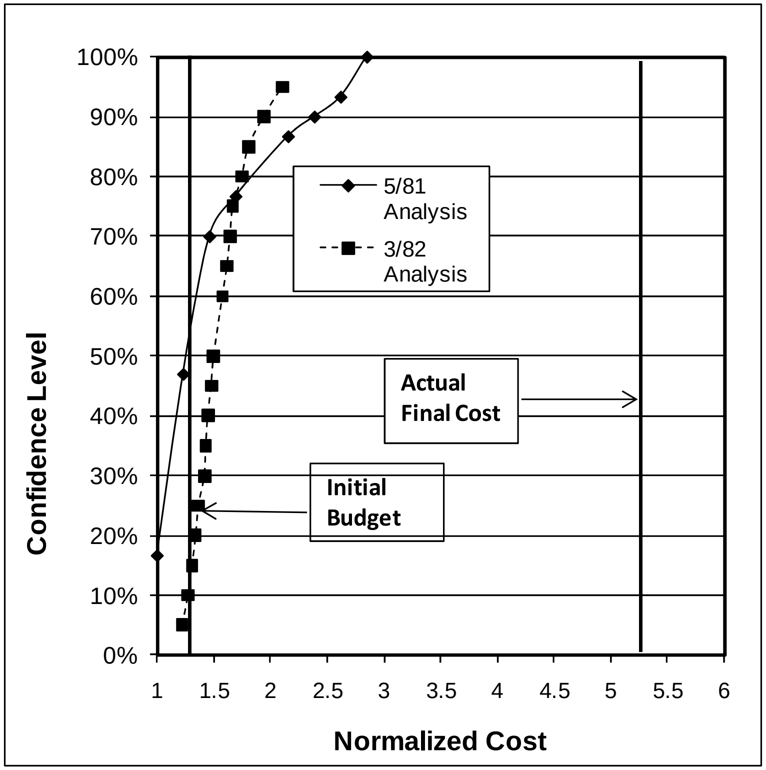
- **There is some limited empirical evidence on correlation for spacecraft**
- **This ranges from 16-40% at the subsystem level**
  - **Smart calculated an average correlation in the range 16-20% for NASA/Air Force Cost Model hardware subsystems (Smart, 2004)**
  - **Covert and Anderson calculated an average correlation equal to 16.8% for Unmanned Spacecraft Cost Model subsystems (Covert and Anderson, 2005)**
  - **Mackenzie and Addison reported correlations in the range 20-40% for average unit cost of subsystems NRO data (Mackenzie and Addison, 2000)**
- **However, this evidence is only for one commodity**

# Summary (1 of 2)

- **20% is often the default value when there is no information to provide informed input**
  - **This level is too low**
- **Using a more robust approach, we have shown that default values in the range 40-63% are more reasonable**
  - **I recommend 63% as a default value**
- **Only downside is potential for overestimation**
  - **However as a profession we do not have a reputation for overestimation**
  - **Increasing default correlation value may help counter this**

# Summary (2 of 2)

- **Example of underestimation of risk**
  - **For a risk analysis conducted for the Tethered Satellite System, the actual cost was more than double the 95th percentile of the original cost risk analysis**



# References


- **Book, S.A., "Why Correlation Matters in Cost Estimating," Advanced Training Session, 32nd Annual DOD Cost Analysis Symposium, Williamsburg, VA, 1999.**
- **Covert, R. and T. Anderson, "Correlation Tutorial, Rev. H," presented at the Cost Drivers Learning Event, November 2005.**
- **Gupton, G.M., C.C. Finger, and M. Bahtia, *CreditMetrics – Technical Document*, 1997, J.P. Morgan, New York, available at: http://www.defaultrisk.com/_pdf6j4/creditmetrics_techdoc.pdf**
- **Mackenzie, D., "Cost Variance in Idealized Systems," presented at the International Society of Parametric Analysts Annual Conference, Cannes, France, June, 1996.**
- **Mackenzie, D., and B. Addison, "Space System Cost Variance and Estimating Uncertainty," presented at the Space Systems Cost Analysis Group meeting, Seattle, WA, October, 2000.**
- **Smart, C., "Average Correlation Values for NAFCOM 2004," unpublished white paper, SAIC, 2004.**
- **Smart, C., "Risk Analysis in the NASA/Air Force Cost Model," presented at the Joint Annual ISPA/SCEA Conference, Denver, CO, June, 2005.**
- **Smart, C., "Mathematical Techniques for Joint Cost and Schedule Analysis," presented at the 2009 NASA Cost Symposium, Cocoa Beach, FL, May, 2009.**
- **Smart, C., "Covered With Oil: Incorporating Realism in Cost Risk Analysis," presented at the 2011 Joint Annual International Society of Parametric Analysts and Society of Cost Estimate and Analysis Conference, Albuquerque, NM, June 2011.**