

# Correlation Coefficients for Spacecraft Subsystems from the USCM7 Database

Raymond P. Covert  
The Aerospace Corporation  
15048 Conference Center Drive, CH1-410  
Suite 600  
Chantilly, VA 20151  
E-mail: raymond.p.covert@aero.org

## ***Abstract***

The Unmanned Space Vehicle Cost Model Revision 7 (USCM 7) contains Cost Estimating Relationships (CER) for spacecraft bus components and subsystems, integration and test, and program level costs. Each of the USCM 7 CERs contains inherent regression errors (expressed in percent error) that represent the uncertainty associated with the CER. These uncertainties are used in risk analysis to capture the total estimating uncertainty of a spacecraft bus estimate. Although the CER uncertainties are not independent, they are either incorrectly modeled in a risk analysis as uncorrelated independent variables or through educated guesses of their true values because the correlation between these subsystem errors have not been published with the USCM 7 CERs. In this paper the correlation coefficients from the differences between normalized actual cost data from the USCM 7 database and results from USCM 7 subsystem CERs are determined and used to determine their effects on the uncertainty in bus cost. The total bus estimate uncertainty using the calculated correlation coefficients are compared against an uncorrelated case to show the effects of the correlation between the random variables.

## **Keywords**

Risk analysis, uncertainty, Monte Carlo, correlation, Unmanned Space Vehicle Cost Model, spacecraft subsystem

## Introduction

All estimates and measurements are subject to uncertainty. The analyst tasked with providing a cost estimate for a spacecraft bus can use libraries of Cost Estimating Relationships (CER), like those in the Unmanned Space Vehicle Cost Model Revision 7 (USCM 7), to provide both a point estimate for each subsystem and an associated error for that estimate. In the case of USCM 7 subsystem CERs, the documentation includes a percentage error of that subsystem cost estimate<sup>1</sup>. This paper answers the following questions:

- How should the uncertainty of the entire estimate be calculated?
- What is a correlation coefficient?
- What are some of the potential causes of correlation?
- How are correlation coefficients derived?
- What are the correlation coefficients for the uncertainties in the USCM 7 database?
- How are the correlation coefficients used in analysis?
- How do the statistical uncertainties of the CERs and correlation coefficients contribute to the uncertainty of the entire estimate?

## ***Uncertainty of an Estimate***

Once a point estimate has been determined using the subsystem CERs in the USCM 7 database, the individual subsystem nonrecurring and recurring estimates is multiplied by its respective percentage error to form a vector of uncertainties. The uncertainty of the entire estimate is then calculated using Equation 1, which provides the total cost variance<sup>2</sup>. The standard deviation, or sigma, of the total cost is merely the square root of the total cost variance. The first terms are merely the sum of the variances (squares of the standard deviation of the individual WBS elements), and the second term calculates the sum of the

---

<sup>1</sup> There are actually two types of CERs available in USCM 7: Mean Percentage Error (MPE) and Mean Unbiased Percentage Error (MUPE). The errors in both of these sets of CERs are expressed as a percentage of the estimate. For this discussion, the MUPE weight-based CERs were used.

<sup>2</sup> Assuming input errors are neglected and that program will behave like others in the database.

covariances. This term is often neglected, but it has a major impact on the total cost variance. We need correlation to accurately capture the statistical effects of adding uncertainties.

**Equation 1:** Total cost variance =  $\sigma_{Total}^2 = \sum_{k=1}^n \sigma_k^2 + 2 \sum_{k=2}^n \sum_{j=1}^{k-1} \rho_{jk} \sigma_j \sigma_k$ , where  $\sigma_j$ ,  $\sigma_k$ , and  $\rho_{jk}$  are the standard deviations of WBS elements j and k respectively and the correlation between them.

## ***The Correlation Coefficient***

There are two types of correlation coefficients used in estimating aggregate uncertainty: Linear (Pearson's product-moment) correlation and (Spearman's) rank correlation. In this paper, linear correlation coefficients are derived and used because the sum of random variables depends on the Pearson's product-moment correlation and not the Spearman rank correlation. Here are brief definitions of the two:

- Pearson's product-moment correlation is a measure of the linearity between two random variables.
- Spearman's rank correlation is a measure of the monotonicity between two random variables.

In the cases where the uncertainties between two variables are linear, there is little difference between the two. On the other hand, when the uncertainties have a non-linear relationship the answers can be remarkably different. Commercial spreadsheet-based Monte Carlo simulations (e.g., Crystal Ball ® and @Risk®) use rank correlation and may give a different answer than analytically derived using linear correlation. In the case of this analysis, the aggregate uncertainty using a spreadsheet-based Monte Carlo tool was 50% higher than analytically determined.

## ***Potential Causes of Correlation***

When correlation is used in an estimate of uncertainty or risk analysis, there is a natural tendency to attribute the causes of the correlation to the value of the coefficients. Correlation does not necessarily imply a causal relationship between two random variables. It is merely a measure of the tendency of one WBS element to cost more than estimated while another WBS element is over or under estimated. In the case where two elements are both under or over estimated, we say that these pairs of WBS elements are positively correlated, and in the case where one is over estimated and the other is under estimated, we say

that they are negatively correlated. While a causal relationship will drive random variables to be correlated, the reverse is not always true.

There are several potential causes of correlation and reasons why WBS elements are correlated. Here are just a few examples encountered:

- Tradeoffs - This is the case usually attributed to negative correlation. There are rather few negatively correlated pairs of uncertainties produced in this paper.
- Schedule / Network dependency - This is a very popular reason attributed to correlation. It certainly has a large effect, but other forces may be more important.
- Common contract, material and labor factors - Labor and material shortages, supply chain problems with a common vendor, may cause components or major subsystems to all overrun (or under run).
- Database normalization and bucketing schemes - This is a cause that is often overlooked by estimators that may have one of the most significant impacts. Learning curve assumptions, bucketing of weights and costs, application of inflation (e.g., Office of the Secretary of Defense, Air Force, Navy or Consumer Price Indexes) significantly affect the conversion of actual costs in then-year dollars to base-year dollars and may skew the data in the database.
- Requirements dependency - This is a cause typically attributed to the correlation between critical bus subsystems and payload subsystems.
- Choice of cost drivers - Different cost drivers used in CERs will produce different uncertainties and thus affect the correlation between pairs of residual errors.
- Over-design and margin - Some systems are over designed with plenty of margin to accommodate requirements creep or performance shortfalls, and some are not.
- External and internal influences - Budget constraints, other programs competing for resources, and catastrophic failures (e.g., launch failures) may affect most or all of the costs of a particular program's subsystem costs in a database. This would clearly influence the correlation.

The cost community cannot yet determine the combined effects of all causes of correlation, and a database that captures all of these effects most likely does not exist. Since it is extremely difficult to enumerate and all these causes, let alone attribute a correlation coefficient to all of them, it is reasonable to assume that

correlation does not necessarily imply causality nor should causality necessarily define correlation. There is a capacity to determine them from data without attributing them to a particular set of causes.

## ***Deriving Correlation Coefficients***

The correlation coefficients developed in this paper were derived using lists of actual subsystem nonrecurring costs, recurring costs, and weights for all 26 spacecraft programs in the USCM 7 database. The first step in the analysis is to use USCM 7 CERs to calculate estimates for subsystem nonrecurring and first unit costs for all of the programs in the database. The next step is to calculate the residuals between actual costs and estimated costs. Once this was completed, pair-wise subsystem residuals were used in Equation 2 to calculate the sample Pearson product -moment correlation,  $r$ .

**Equation 2:**

$$r_{xy} = \frac{\sum (x_i - x_m)(y_i - y_m)}{\sqrt{\sum (x_i - x_m)^2 \sum (y_i - y_m)^2}}, \text{ where } x \text{ and } y \text{ are CER residual pairs, } x_i \text{ and } y_i \text{ are individual program residual data, and } x_m \text{ and } y_m \text{ are the mean of the residuals respectively.}$$

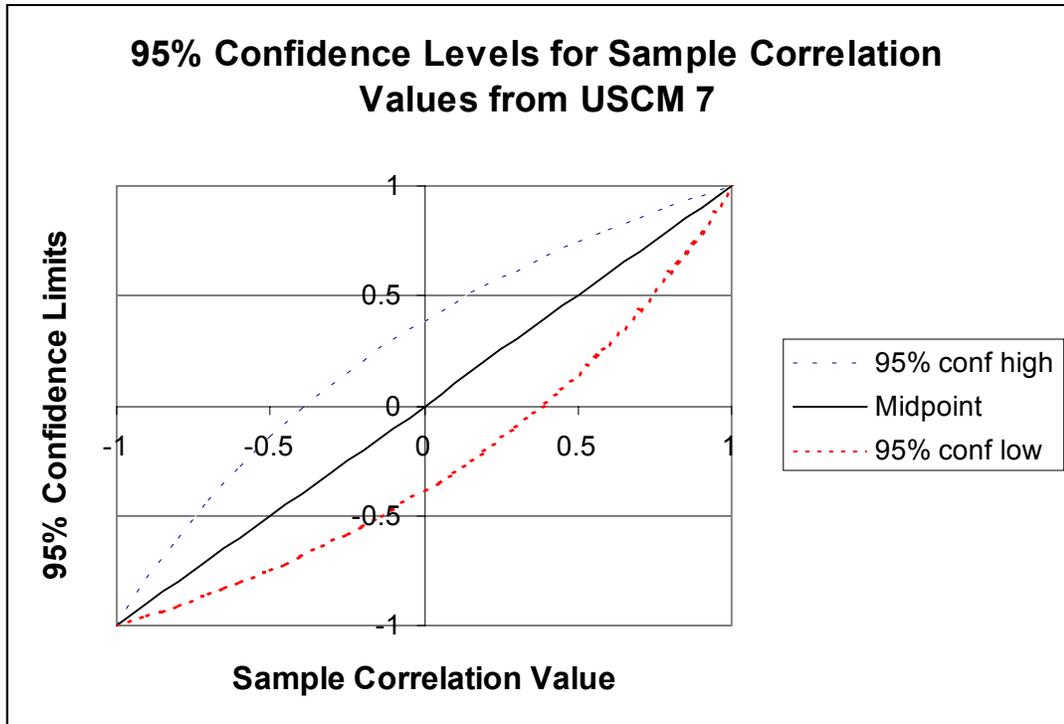
If the two variables exactly follow a linear relationship (with no scatter), then the correlation coefficient  $r = +1$  or  $-1$ . Similarly, if there is no correlation between  $x$  and  $y$ , then the numerator should be zero and so should  $r$ .

Since the number of elements on the sample population ( $N=26$ ) is considered a small sample, the sample correlation values. Once the sample correlation coefficients were derived, their confidence intervals were derived to of the  $s$ . This probability is difficult to compute, but it can be done. Table 1, which shows percentage probabilities, answers this question. The rows represent  $N$ , the number of data points, and the columns are labeled with values for  $\rho$ .

For example, we used 26 data points, so a set of 20 to 30 data points would be uncorrelated at the 0.5% to 2.5% (1  $\sigma$ ) confidence level if their correlation coefficient came out to 0.5. These numbers were taken from Reference 1. The table confirms that the correlation coefficients derived in this analysis are reasonable.

$$Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = 1.1513 * \log\left(\frac{1+r}{1-r}\right)$$

the Z statistic is approximately normally distributed with mean and standard deviation



$$\mu_z = \frac{1}{2} \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513 * \log\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

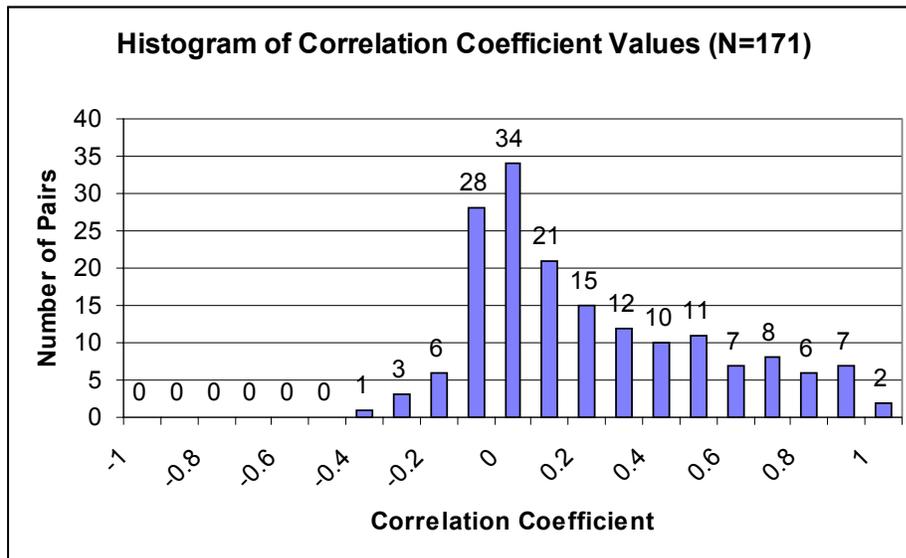
$$\sigma_{z=} = \frac{1}{\sqrt{N-3}}$$

**Table 1. Confidence Levels of Sample Correlation Coefficients**

		r										
N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
3	100	94	87	81	74	67	59	51	41	29	0	
4	100	90	80	70	60	50	40	30	20	10	0	
6	100	85	70	56	43	31	21	12	5.6	1.4	0	
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0	
10	100	78	58	40	25	14	6.7	2.4	0.5	<0.1	0	
15	100	72	47	28	14	5.8	1.8	0.4	<0.1	<0.1	0	
20	100	67	40	20	8.1	2.5	0.5	0.1	<0.1	<0.1	0	
30	100	60	29	11	2.9	0.5	<0.1	<0.1	<0.1	<0.1	0	

### **USCM 7 Correlation Coefficients**

The distribution of the values of sample correlation coefficients in the USCM 7 database is shown in Figure 1 below. The correlation coefficients range from approximately -0.4 to 1.0 with the most likely value being near 0.0. The shape of the histogram is not normal, as might be expected, and the correlation coefficients are predominately positive.



**Figure 1. Histogram of Correlation Coefficient Values**

Table 2 shows the result of the derivation of the correlation Matrix for USCM 7 Weight-Based, MUPE Subsystem CERs.

**Table 2. Sample Correlation Matrix for UCSM 7 Weight-Based, MUPE Subsystem CERs**

	ADCSNR	AGENR	COMMNR	EPSNR	IATNR	PROGNR	STRCNR	THERNR	TT CNR	ADCST1	AKMT1	COMMT1	EPST1	IATT1	LOOST1	PROGT1	STRCT1	THERT1	TT CT1	
ADCSNR	1.000	-0.067	-0.096	-0.035	0.035	0.012	0.413	0.605	0.121	-0.095	0.983	-0.122	0.099	0.564	0.139	0.089	-0.047	-0.057	0.092	
AGENR		1.000	-0.028	0.525	-0.079	0.127	0.091	-0.230	-0.125	0.416	0.001	0.085	-0.043	-0.163	-0.189	0.033	0.146	0.151	0.232	
COMMNR			1.000	0.888	0.884	0.966	0.762	0.281	0.850	-0.166	0.305	-0.176	0.157	0.368	0.884	-0.158	0.109	0.037	-0.004	
EPSNR				1.000	0.265	0.604	0.409	0.003	0.337	0.237	0.011	-0.275	0.076	0.342	0.021	-0.049	0.465	0.123	0.035	
IATNR					1.000	0.721	0.615	0.331	0.747	-0.037	0.391	-0.133	-0.028	0.501	0.265	-0.145	0.113	-0.014	-0.189	
PROGNR						1.000	0.697	0.222	0.868	-0.065	0.145	-0.191	-0.044	0.444	0.329	-0.191	-0.000	-0.125	0.019	
STRCNR							1.000	0.837	0.761	-0.001	0.117	-0.214	-0.113	0.418	0.173	-0.018	0.220	-0.103	0.069	
THERNR								1.000	0.077	-0.200	0.662	-0.171	-0.053	0.514	0.102	-0.010	-0.063	-0.165	0.092	
TT CNR									1.000	-0.149	0.475	-0.118	-0.071	0.519	0.294	-0.178	-0.111	-0.095	0.022	
ADCST1										1.000	-0.100	0.614	0.421	-0.262	-0.354	0.543	0.676	-0.029	0.655	
AKMT1											1.000	-0.006	0.292	0.855	0.286	0.176	-0.003	-0.027	0.052	
COMMT1												1.000	0.266	-0.454	-0.088	0.777	0.729	0.126	0.391	
EPST1													1.000	-0.150	-0.145	0.381	0.388	-0.007	0.520	
IATT1														1.000	0.448	-0.144	-0.224	-0.014	-0.320	
LOOST1															1.000	-0.336	-0.097	-0.074	-0.169	
PROGT1																1.000	0.421	-0.039	0.481	
STRCT1																	1.000	-0.175	0.285	
THERT1																		1.000	-0.140	
TT CT1																				1.000

## **Correlation Used in Analysis**

The correlation matrix and the calculated uncertainties can be combined in Equation 3 to calculate the variance of the total estimate<sup>3</sup>. For an uncorrelated case, the matrix,  $\rho$ , is replaced with the identity matrix,  $\mathbf{I}$ , (ones on the diagonal and zeros in the off-diagonal elements) to form Equation 4.

**Equation 3:**  $\sigma^2_{Total} = \sigma^T \rho \sigma$ , where  $\rho$  is the Correlation matrix (full matrix),  $\sigma$  is the Vector of standard deviations (cost space), and  $\sigma^T$  is the Transpose of vector of standard deviations (in cost space)

**Equation 4:**  $\sigma^2_{Total} = \sigma^T \mathbf{I} \sigma = \sigma^T \sigma$

The standard deviation of the total estimate is merely the square root of the variance of the total estimate. The next step in this analysis was a comparison of the effects of the correlation on the estimate uncertainty.

## **Coefficient of Variation**

The coefficient of variation (COV) is defined in Equation 5 as the ratio of the standard deviation to the mean and is a useful measure of dispersion when comparing different estimates. Table 3 shows the effect of correlation on the COV on three estimates of arbitrarily chosen spacecraft programs in the USCM 7 database. A comparison of uncorrelated case and correlated case in Table 3 shows an increase in the variance and thus the COV for the correlated run. This is to be expected, since the correlation coefficients are predominately positive.

**Equation 5:**  $COV = \sigma/\mu$ , where  $\sigma$  and  $\mu$  are the standard deviation and mean of the total cost estimate.

It is interesting to note that the COV nearly doubles due to the effect of correlation. This means that the uncertainty of the estimate is twice than would be expected if correlation were not accounted for.

---

<sup>3</sup> Equation 3 can be translated to the following Excel functions:

`SIGMA_TOT=SQRT(MMULT(MMULT(TRANSPOSE(SIGMA),RHO),SIGMA))`

`PERCENTILE=NORMDIST(ACTUAL,ESTIMATE,SIGMA_TOT,TRUE)`

**Table 3. Effect of Correlation on COV**

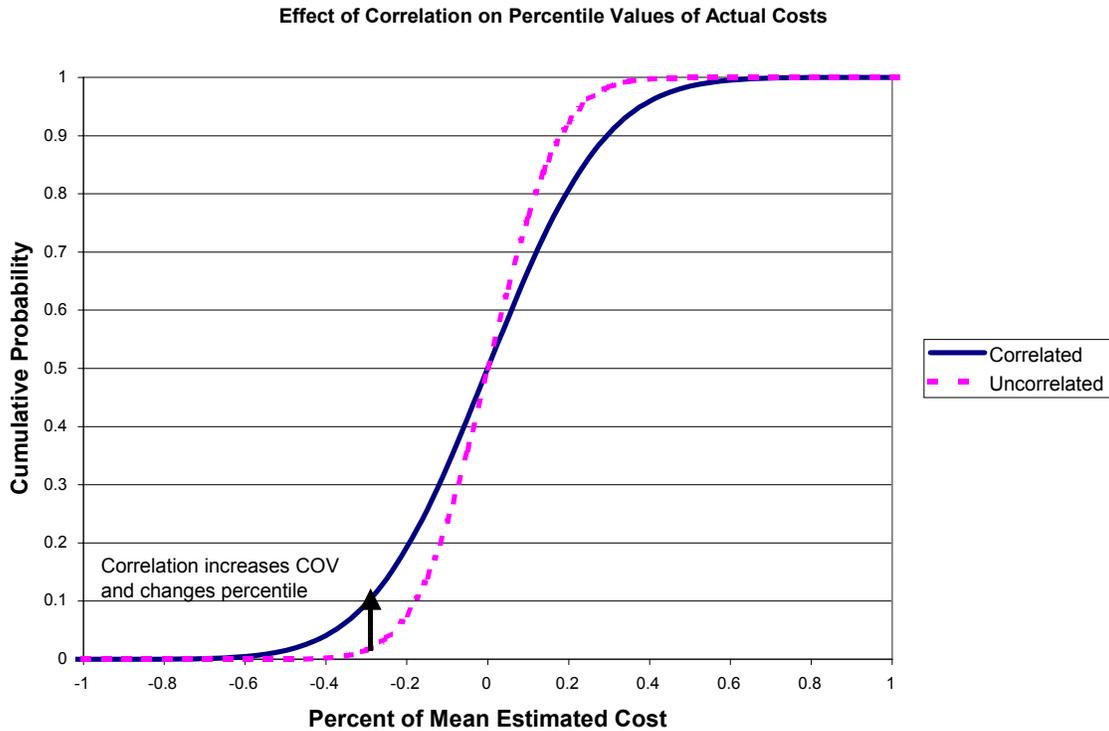
r=0	PROG1	PROG2	PROG3	r<>0	PROG1	PROG2	PROG3
<b>NR</b>	0.15	0.15	0.17	<b>NR</b>	0.27	0.27	0.28
<b>T1</b>	0.14	0.14	0.17	<b>T1</b>	0.23	0.23	0.26
<b>NR+T1</b>	0.11	0.11	0.13	<b>NR+T1</b>	0.20	0.20	0.22

### Percentiles of Actual Costs

The final step in the analysis was to determine where actual costs from the database fall in relation to estimated values with uncertainty. The results using the same three arbitrarily chosen programs from the USCM 7 database are presented in Table 4. As expected, the increase in variance in the correlated case move the actual cost values to the center. If the actual cost fell below the 50<sup>th</sup> percentile in the uncorrelated case, its percentile would increase in the correlated case, and if the actual cost fell above the 50<sup>th</sup> percentile in an uncorrelated case, its percentile would decrease in the correlated case. Note that for program 1 the percentiles of both the nonrecurring total (NR) and the total estimate (NR+T1) actual costs correspond to a very small value (near zero) in the uncorrelated run, but have a small percentage in the correlated run.

**Table 4. Effect of Correlation on Percentile of Actual Costs in the Estimate**

r=0	PROG1	PROG2	PROG3	r<>0	PROG1	PROG2	PROG3
<b>NR</b>	0.0%	90.5%	57.3%	<b>NR</b>	0.8%	75.7%	54.5%
<b>T1</b>	5.1%	77.6%	28.8%	<b>T1</b>	16.2%	67.8%	35.7%
<b>NR+T1</b>	0.0%	93.4%	49.4%	<b>NR+T1</b>	0.5%	79.3%	49.6%



## Potentially Missing Drivers

It was interesting to note that some of the correlation coefficients were near unity, as shown in Table 5.

**Table 5. High Correlation Coefficients Between CER Uncertainties**

Error1	Error2	Correlation
AKMT1	ADCSNR	0.983
PROGNR	COMMNR	0.966
EPSNR	COMMNR	0.888
IATNR	COMMNR	0.884
LOOST1	COMMNR	0.884
PROGNR	TTCNR	0.868
IATT1	AKMT1	0.855
TTCNR	COMMNR	0.850

These high correlation values may indicate the presence of hidden relationships between the errors of the correlated pairs, which may require further study.

## References

1. Taylor, John, *An Introduction to Error Analysis*, University Science Books, Mill Valley, CA 1982.

2. Nguyen, P., et al., *Space and Missile Systems Center Unmanned Spacecraft Cost Model Seventh Edition*, Space and Missile Systems Center, Cost Division, Los Angeles AFB, CA, August 1994.
3. Book, Stephen A., "Cost Risk analysis: A Tutorial", in conjunction with the Risk Management Symposium Co sponsored by USAF Space and Missile Systems Center and The Aerospace Institute, Manhattan Beach, CA, 2 June 1997
4. Book, Stephen A., "Why Correlation Matters in Cost Estimating", 32nd Annual DoD Cost Analysis Symposium, Williamsburg, VA, 2-5 February, 1999
5. Garvey, Paul R, " Do Not Use Rank Correlation in Cost Risk Analysis", 32nd Annual DoD Cost Analysis Symposium, Williamsburg, VA, 2-5 February, 1999
6. Lurie, Philip M. and Goldberg, Matthew S., " Simulating Correlated Random Variables", 32nd Annual DoD Cost Analysis Symposium, Williamsburg, VA, 2-5 February, 1999

$$P_N(|r| \geq |r_0|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}[(N-2)/2]} \int_{|r_0|}^1 (1-r^2)^{(n-4)/2} dr$$