EVERYTHING YOU WANT TO KNOW ABOUT CORRELATION BUT WERE AFRAID TO ASK

FRED KUO

MOTIVATION

- Correlation as a source of confusion
  - Some of the confusion may arise from the literary use of the word to convey dependence as most people use “correlation” and “dependence” interchangeably
  - The word “correlation” is ubiquitous in cost/schedule risk analysis and yet there are a lot of misconception about it.
- A better understanding of the meaning and derivation of correlation coefficient, and what it truly measures is beneficial for cost/schedule analysts.
- Many times “true” correlation is not obtainable, as will be demonstrated in this presentation, what should the risk analyst do?
- Is there any other measures of dependence other than correlation?
  - Concordance and Discordance
  - Co-monotonicity and Counter-monotonicity
  - Conditional Correlation etc.

CONTENTS

- What is Correlation?
  - Correlation and dependence
  - Some examples
- Defining and Estimating Correlation
  - How many data points for an accurate calculation?
  - The use and misuse of correlation
  - Some example
- Correlation and Cost Estimate
  - How does correlation affect cost estimates?
  - Portfolio effect?
- Correlation and Schedule Risk
  - How correlation affect schedule risks?
- How Shall We Go From Here?
  - Some ideas for risk analysis

POPULARITY AND SHORTCOMINGS OF CORRELATION

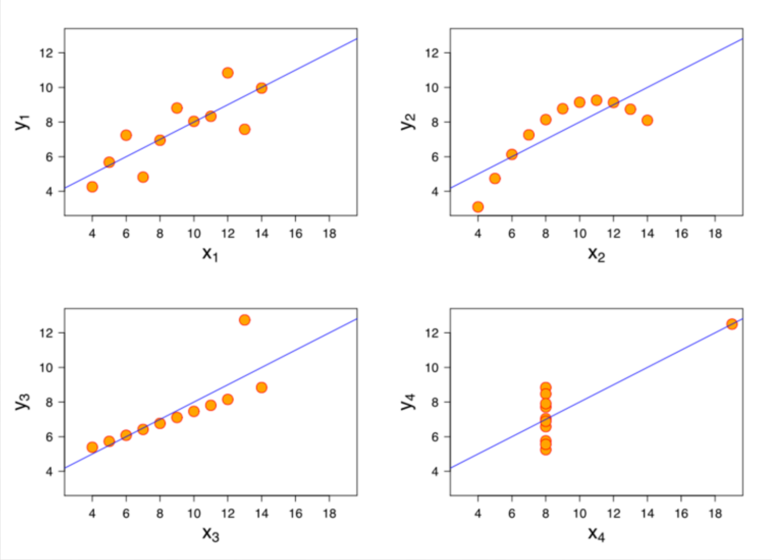
- Why Correlation Is Popular?
  - Correlation is a natural measure of dependence for a Multivariate Normal Distribution (MVN) and the so-called elliptical family of distributions
  - It is easy to calculate analytically; we only need to calculate covariance and variance to get correlation
  - Correlation and covariance are easy to manipulate under linear operations
- Correlation Shortcomings
  - Variances of R.V. X and Y must be finite or "correlation" can not be defined
  - Independence of 2 R.V. implies they are not correlated, but zero correlation does not in general imply independence
  - Linear correlation is not invariant under non-linear transformation

WHAT IS CORRELATION?

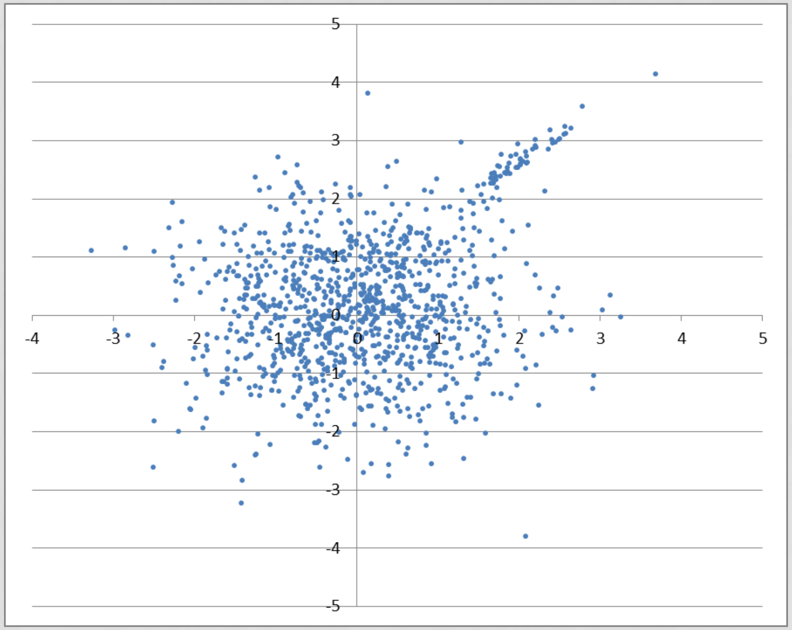
- We generally refer to "Pearson's" product-moment coefficient
- There are other, but less used, definitions for "correlation" such as
  - Rank correlation
  - Kendall's Tau
- It is a measure of only *linear* dependence, only a sliver of information regarding dependence between two random variables.
- It is a very crude measure of dependence.
- It does not necessarily indicate causality:
  - Correlation coefficient of 1 does not imply causality, only " perfect" dependence
  - "perfect" dependence means the ability to express one variable as a deterministic function of the other.
  - Correlation coefficient of 0 does not preclude dependence
- Can you guess the correlation coefficient of the following functions, where x is a random variable?
  - Y = 3 * x
  - Y= 10 * x
  - Y = 3 * x – 1
  - Y = x^2
  - Y = abs(x)
  - Y = Sin(x)

SOME EXAMPLES OF PITFALLS

The famous anscombe example
(same correlation coefficient)



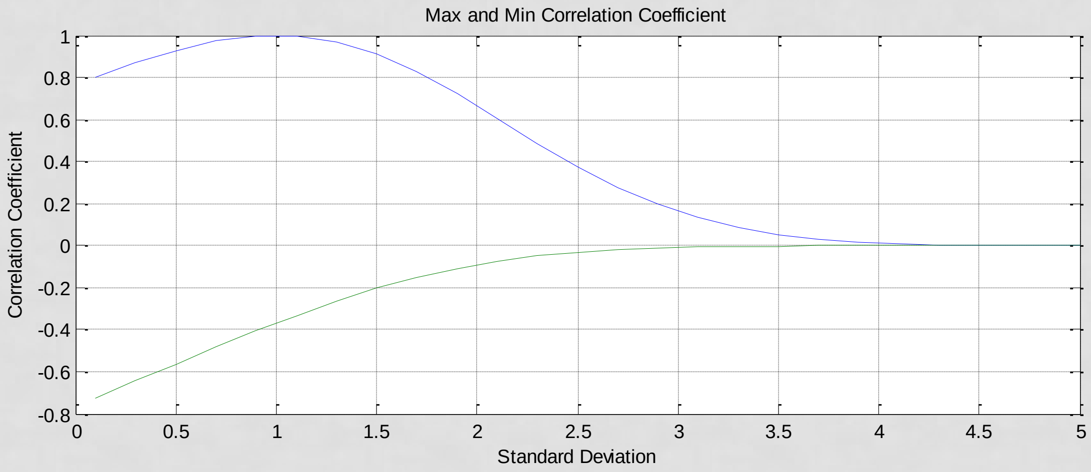
High correlation at the right tail
corr=.3 overall, but corr=.9 at 2 sigma



RANGE OF APPLICABILITY

- Accuracy of correlation is dependent on the variance of the data.
- There is a general degradation of correlation coefficient when the volatility of the data increases, i.e., correlation approaches 0 when volatility approaches infinity.
- For example, lognormal distribution can be founded to be bounded by:

$$\rho_{min} = \frac{e^{-\sigma}-1}{\sqrt{(e-1)(e^{\sigma^2}-1)}};\ \rho_{max} = \frac{e^{\sigma}-1}{\sqrt{(e-1)(e^{\sigma^2}-1)}}$$



DEFINITION OF CORRELATION

- Sample correlation calculation

$$\hat{\rho}_{x,y} = \frac{\text{cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

$$\hat{\sigma}_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2$$

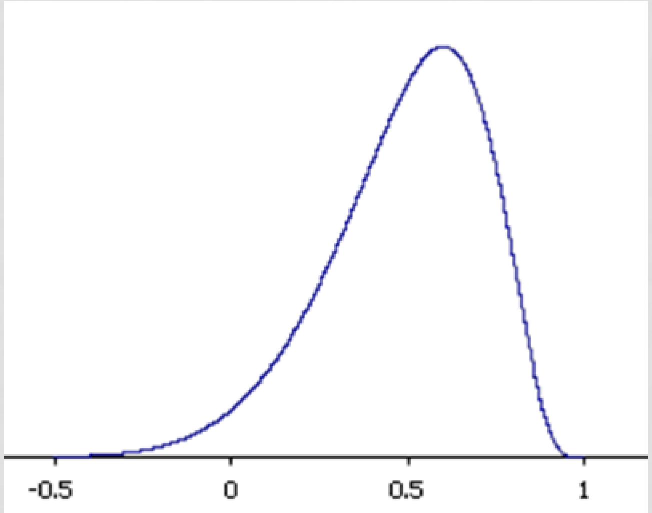
- Cov(x,y) is the covariance σ_{xy}
- Relationship between correlation and covariance is therefore:
- $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$

- There are Excel functions that calculates all these:
  - COVARINCE.P, CORREL.P, STDEV.P

WHAT IS FISHER Z-TRANSFORMATION

- Since "correlation" is a statistical entity, the accuracy of the estimate depends on the number of data points.
- However, Pearson's correlation is not normally distributed so it is hard to calculate standard error.
- Fisher Z transformation is a technique:
- $z = \frac{1}{2} \ln\left[\frac{1+\rho}{1-\rho}\right]; \sigma_z = \frac{1}{\sqrt{N-3}}$
- Which is Normally Distributed with standard error σ_z , which can be used to construct confidence intervals for ρ.

Sampling Distribution of Pearson's ρ
ρ= .6, N= 12



CONFIDENCE INTERVAL FOR PEARSON'S CORRELATION

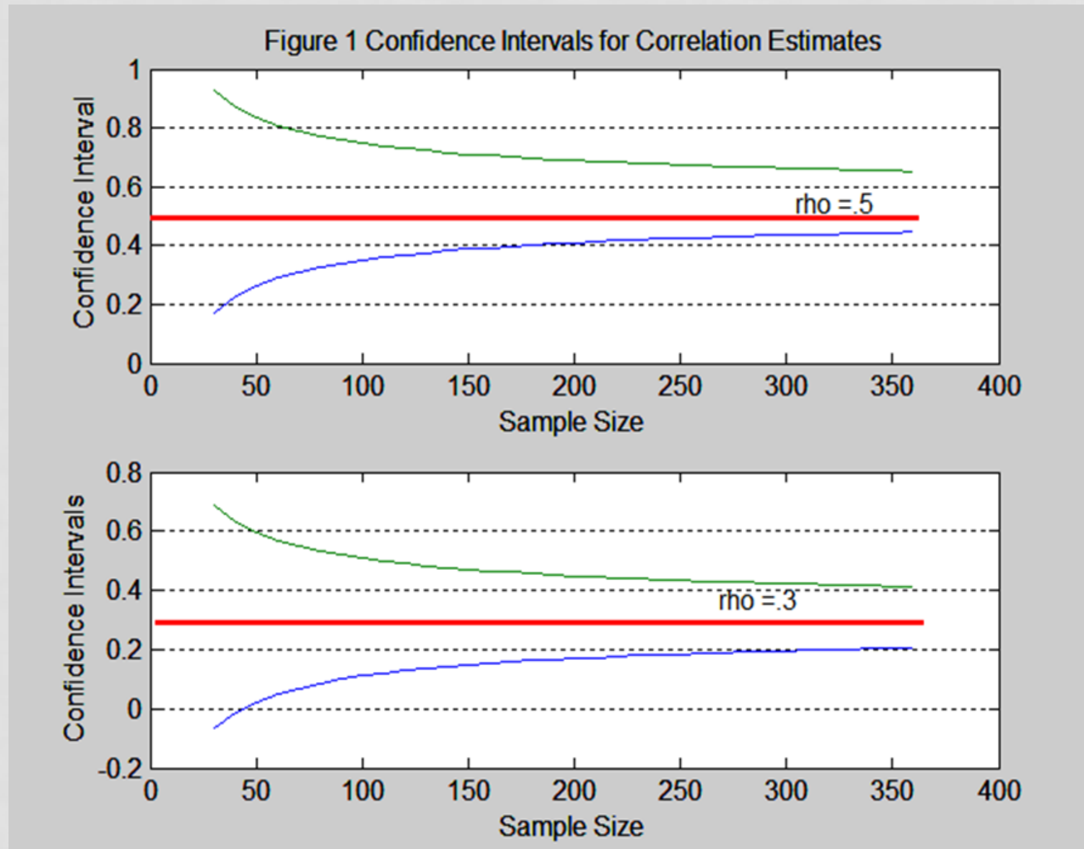
- The Fisher Z-Transformation calculates the bounds; for 95% confidence interval:

$$\rho_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}\,;\; z_L = \hat{z} - \frac{1.96}{\sqrt{N-3}};$$

$$\rho_H = \frac{e^{2z_H}-1}{e^{2z_H}+1}\,;\; z_H = \hat{z} + \frac{1.96}{\sqrt{N-3}};$$

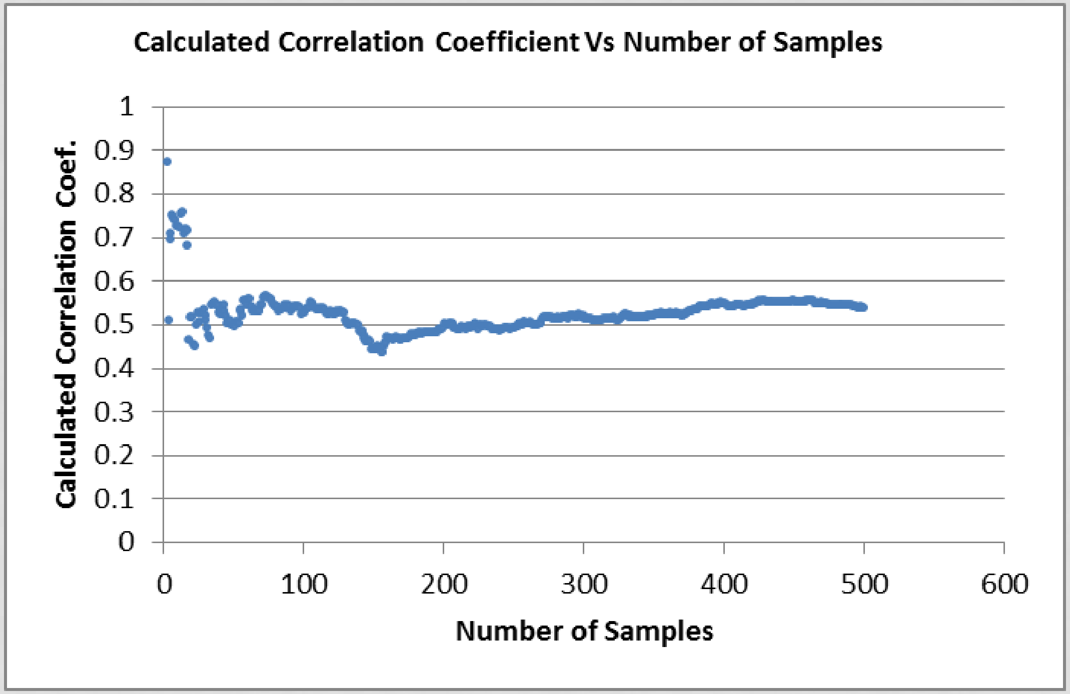
$$\hat{z} = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$

- Most space system/subsystems have far fewer data points than necessary for accurate depiction.



LIMITS ON ACCURACY EXAMPLE

- Would like to check out with my own example
- Use Excel function to generate 2 random uniforms
- Use inverse function to generate 2 N(0,1), random normal (x_1, x_2)
- Create 2 correlated random normal by using the
- $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \rho = 0.5$
- Use CORREL function to generate correlation coefficients between (y_1, y_2), as a function of number of samples
- At less than 20 samples, the deviation is substantial



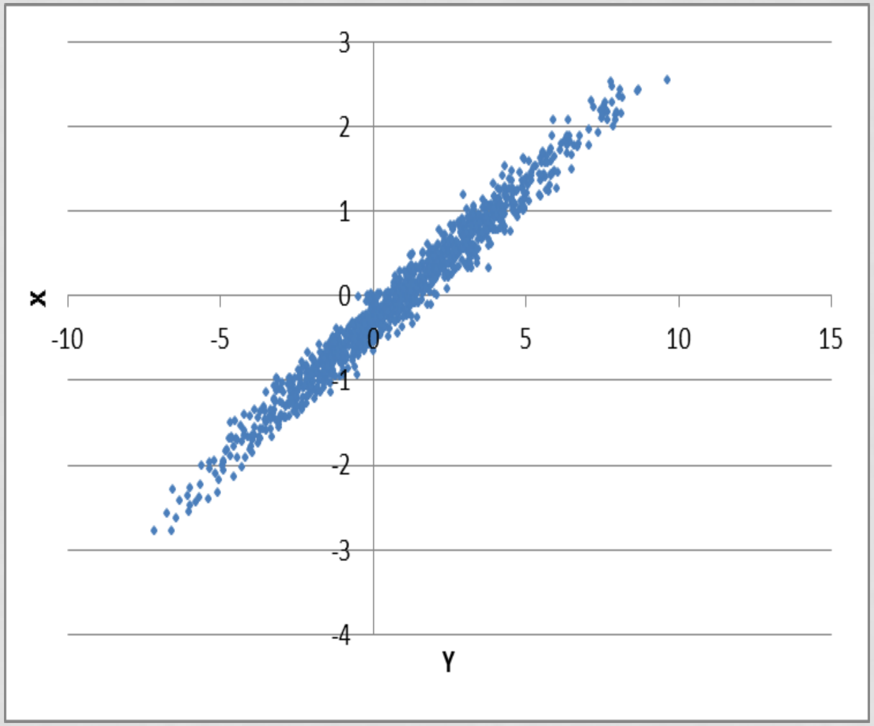
CORRELATION AND LINEAR REGRESSION

- A linear regression model is an estimation tool and it has the following generalized form:
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Where
- β_0 is the intercept
- β_1 is the slope of the regression line
- ϵ_i are assumed to be N(0, σ^2), and σ^2 = VAR(Y)

- It can be shown that
- $\widehat{\beta_1}$ = COR(Y,X)$\frac{SD(Y)}{SD(X)}$ =$\rho_{yx}\frac{\sigma_Y}{\sigma_X}$; $\widehat{\beta_0} = \mu_Y - \widehat{\beta_1}\,\mu_X$
- And that
- $R^2 = \rho_{x,y}^2$
- R is actually the correlation coefficient between Y and X

LINEAR REGRESSION EXAMPLE

- A scatter plot of the equation
- $Y = 1 + 3 * x + \epsilon$
- Where
- β_0 is 1
- β_1 is 3
- ϵ_i are assumed to be N(0, σ^2), and σ^2 = VAR(Y)=.5
- Calculations:
- ρ_{yx}= .9854; σ_Y = 2.928; $\sigma_x = .966$
- μ_X = -.00015; $\mu_y = 1.028$
- $\widehat{\beta_1} = \rho_{yx}\frac{\sigma_Y}{\sigma_X} = 2.9868$
- $\widehat{\beta_0} = \mu_Y - \widehat{\beta_1}\,\mu_X$= 1.028



CORRELATION MATRIX

- When more than 2 random variables are modeled, a correlation coefficient matrix is necessary to represent the inter-relationship.
- A correlation matrix must be consistent, or defined as positive semi definite.
  - A test of positive semi definite is that all Eigenvalues are greater than or equal to 0.
- A portfolio of standard deviation can be written in matrix form as:

$$\sigma_p = \sqrt{\sigma C \sigma'}$$

$$\sigma = [\sigma_1, \sigma_2, \ldots\ldots\ldots \sigma_n]$$

- σ is a row vector of individual standard deviation and C is the correlation coefficient matrix.
- It is obvious that σ_p can not be negative, therefore, the requirement that C must be positive semi-definite

CORRELATION MATRIX CONT'D

- The matrix C must be positive definite if we require $\sigma_p > 0$, which will be the case for all real-life cases.
  - Correlation matrix calculated from raw data is guaranteed to be consistent.
  - However, most correlation in practice are either arbitrarily assigned or a subjective guess.
- The importance of a consistent matrix is 2-fold:
  - In calculating a correct portfolio standard deviation, and
  - A necessary condition in generating correlated random variables for Monte Carlo Simulations
    - Most simulation tools will give you warning when the consistency criterion is not met.
    - There are tools to repair inconsistent correlation matrix

CORRELATION MATRIX EXAMPLE

- When correlation matrix is calculated from sample data, it is guaranteed to be consistent, in practice however, most are subjectively assigned, for example:
  - Original matrix C_1 is consistent
  - Wished to change C_1 to a more desired correlation of C_2.
  - Now C_2, however, is inconsistent.
- By adjusting some minor changes to C_1, $C_3$is consistent.
- Note how small the differences between $C_1$and C_3

$$C_1 = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.3 \\ 0.7 & 0.3 & 1 \end{bmatrix}$$

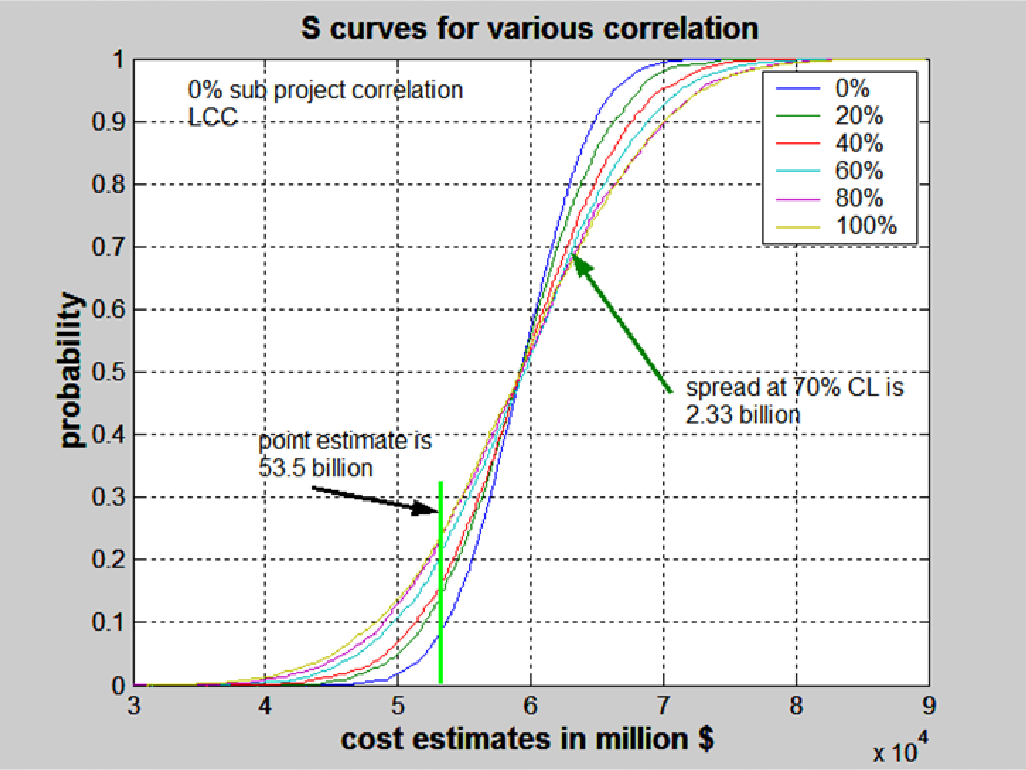
$$C_3 = \begin{bmatrix} 1 & 0.894 & 0.696 \\ 0.894 & 1 & 0.301 \\ 0.696 & 0.301 & 1 \end{bmatrix}$$

EFFECT OF CORRELATION IN COST RISK ANALYSIS

- The effect of correlation on cost estimates and cost risk analysis can best be described from a portfolio perspective.
- A cost estimate for an system can be thought of as a portfolio of sub element costs, each with its own mean cost and standard deviation.
- $\mu_p = \sum_{i=1}^{n} \mu_i$, $\sigma_p = \sqrt{\sigma C \sigma'}$, note that $\sigma_p \leq \sum_{i=1}^{n} \sigma_i$
- This property states that the portfolio standard deviation is always less than the sum of its constituent's standard deviation when the correlation between these elements are less than 1.
- Since the steepness of the cost S-curve, and therefore the confidence level, is determined by the standard deviation, the impact of correlation will ultimately be reflected in the confidence level as well.

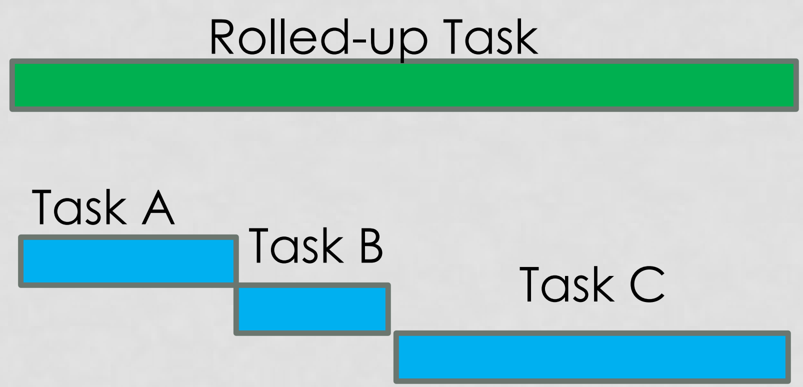
CORRELATION AND COST ESTIMATE- AN EXAMPLE

- From the previous equations, correlation does not change the expected costs or point estimates.
- Correlation only changes the portfolio standard deviation, which relates to the steepness of the S-Curve, and therefore, the confidence level.
- Higher correlation among the sub-elements tend to increase the portfolio standard deviation, and therefore a wide spread of slope.
- Counter intuitive:
  - Higher correlation increases point estimate confidence level.
  - It also increases budget required for the 70% confidence level.
  - So, in general, if the point estimate is below the expected value, correlation improves confidence level.
  - If the point estimate is above expected value, then correlation decrease confidence level.



EFFECT OF CORRELATION IN SCHEDULE RISK ANALYSIS

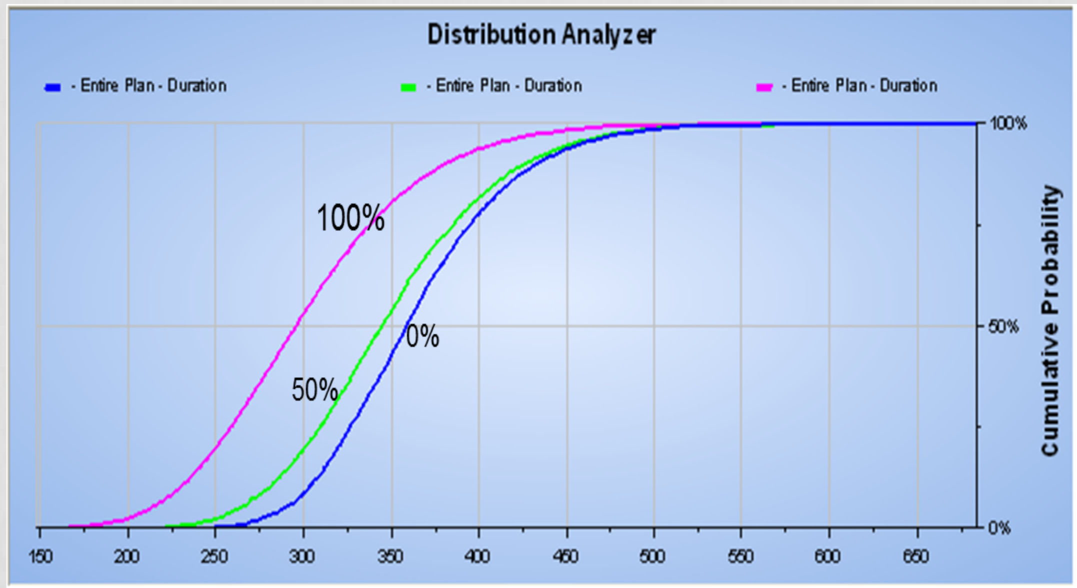
- Correlation effect on schedule risks analysis is more interesting and counter intuitive.
- It has different effect, depending on whether we are modelling rolled-up, parallel or serial tasks.
- The effect of correlation on serial tasks is similar to that of cost. Higher correlation coefficient tends to tilt the S-Curve.
- The variance of rolled-up tasks is dependent on the variances of the subtasks.
- When we used the same variance for the rolled-up tasks and the subtasks, we are implicitly assuming 100% correlation of the subtasks.



Correlation	0	0.2	0.4	0.6	0.8	1
Subtasks SD	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%
Rolled-up Task SD	11.5%	13.7%	15.5%	17.1%	18.7%	20.0%

EFFECT OF CORRELATION IN PARALLEL TASKS

- Example:
  - 4 tasks of equal duration of 300 days and SD of 60 days with Correlation of 0%, 50% and 100%.
  - This results in progressive reduction in mean duration (shift left) but increase in variance.
  - This is because by increasing correlation it means that random samples are more synchronized so that all tasks will converge to the dominant one.



Column1	Deterministic Value	Deterministic Probability	Mean	Standard Deviation
corr=0	300	9%	364	52.4
corr=50	300	20%	350	58.2
corr=100	300	54%	301	60.3

WHERE DO WE GO FROM HERE?

- In this presentation, we have identified some dilemmas regarding the use of correlation in risk analysis.
- The main point is that “ we don’t really know” what the true correlation coefficients are in most of our analysis.
  - Not enough data points
  - Correlation may not be true representation of dependence
- However, to quote Dr. Carl Sagan “absence of evidence is not evidence of absence”. The fact that we don’t know what coefficients are does not mean there is no correlation.
- Therefore, by understanding the impact of correlation on cost/schedule analysis, one can either take conservative or optimistic assumptions, dependent upon the circumstances.
- However, there can be other legitimate strategy as well, based on decision and game theory.

WHAT IS A MINIMAX STRATEGY

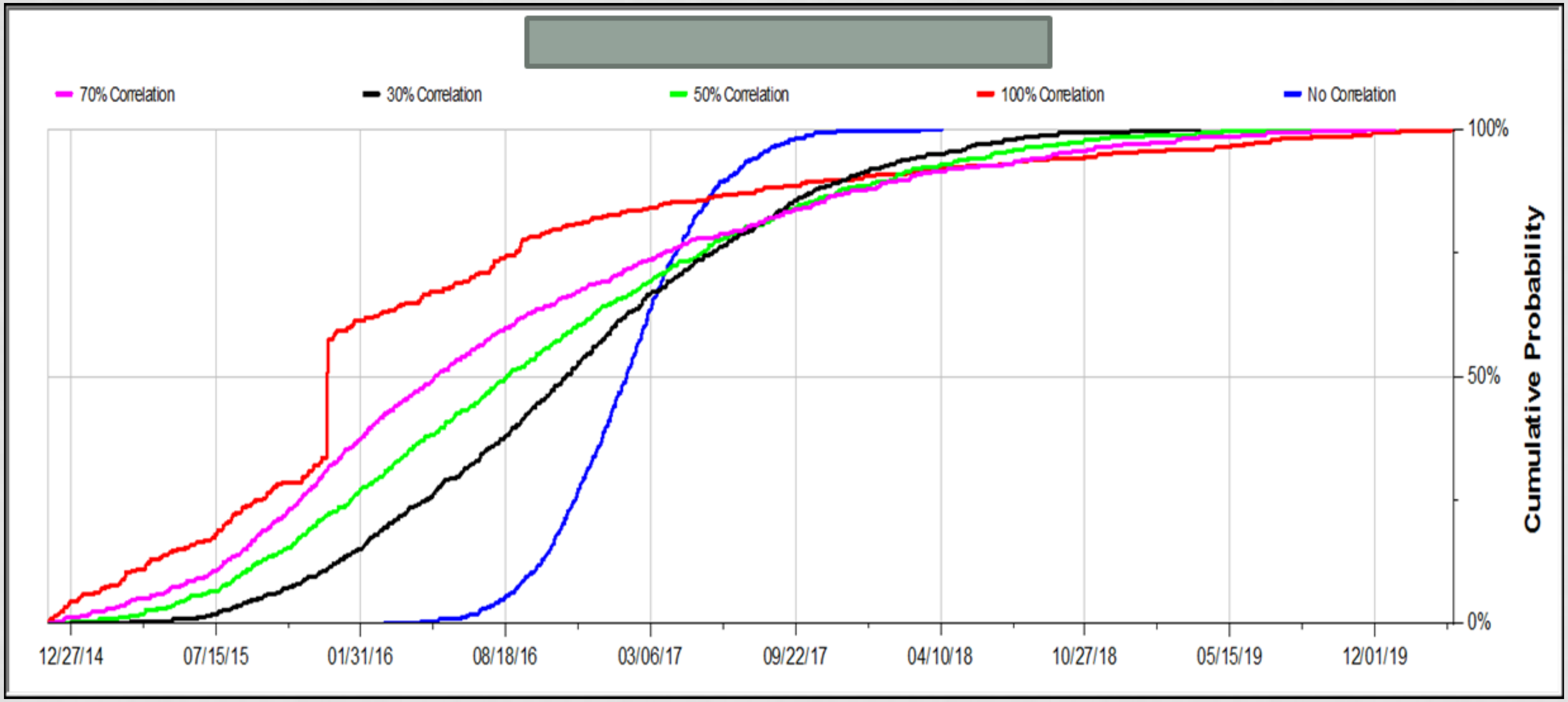
- Minimax is a decision rule used in decision theory, game theory and statistics for *mini*mizing the possible loss for a worst case scenario. I like to call it "Minimum regret" or "Minimum error".
- The idea is very simple: If I used a certain correlation coefficient, and the true correlation is different. What correlation should I use to minimize this error?
- This is an example for the Constellation Program that showed 0.4 is the minimum error. This number is now almost the "de facto" correlation coefficient for cost estimate.
- However, I would suggest to go through the calculation process independently and verify for yourself.

what is the percent error if my correlation assumption is wrong?

Assumed Correlation	True Correlation Coefficients: 0.00	0.20	0.40	0.60	0.80	1.00
0.00		1.76%	3.15%	4.33%	5.36%	6.28%
0.20			1.42%	2.61%	3.66%	4.60%
0.40				1.21%	2.28%	3.23%
0.60					1.08%	2.04%
0.80						0.97%

SIMILARLY FOR SCHEDULE

- Assessed schedule correlation using Minimum Error Method
- 50% correlation produced results with the least error



SUMMARY AND CONCLUSION

- Correlation is an input parameter to most cost/schedule and risk analysis.
- The properties of "correlation", its ranges of applicability as well as its implication on cost/schedule analysis were discussed in this presentation.
- Due to the scarcity of data, correlation coefficient is an unknown quantity in most cost/schedule applications.
- This paper also suggested some strategies in dealing with unknown correlation coefficient.
- Analyst should understand and document the rationale for choosing a particular correlation value, and quantify its impact on the analysis results through sensitivity analysis.