Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

Mathematical Lessons Learned from a Year's Worth of ICEs

Ryan Boulais and Brett Dickey

SCEA – June 2011

Scitor Corporation

Table of Contents

Scitor Corporation

Scitor Corporation

Correlation

The Problem

Scitor Corporation

- **Needed a Monte Carlo simulation tool for a cost estimate conducted on a "stand-alone" system**
  - Did not have access to Monte Carlo SW package
- **Our team created a routine in Excel with Mathematical equations and Visual Basic Code**
- **Used @Risk to validate results**
  - Could have used Crystal Ball, packages are similar
- **Custom developed package appeared to work fine *except for the application of Correlation***
  - Accounting for an acceptable amount of variance due to the "randomness" of Monte Carlo simulation
  - Problem was using Excel's "Correl" function (Pearson) vs. @Risk's method (Spearman) to measure correlation
    - Was not an issue with the results but rather the validation method!

History of Correlation

Scitor Corporation

1888

Sir Francis Galton first introduced the concept of correlation through examining forearm and height measurements

1904

Charles Spearman developed rank correlation, a non-parametric version of the conventional Pearson correlation

2011

Albuquerque, New Mexico- SCEA

< 65 MYA

Dinosaurs ruled the Earth

1896

Karl Pearson introduces "Pearson's product-moment correlation coefficient"

1930s-1940s

WWII

Monte Carlo Simulation

*Photos from Wikipedia

Background

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

- **What is correlation?**
  - A causal, complementary, parallel, or reciprocal relationship, especially a structural, functional, or qualitative correspondence between two comparable entities*
    - A zero correlation indicates that there is no relationship between the variables
    - A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down
    - A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction together
- **Correlation Coefficient Calculations:**
  - Pearson
    - In statistics, the **Pearson product-moment correlation coefficient** is a measure of the linear dependence between two variables *X* and *Y*
  - Spearman
    - **Spearman's Rank Correlation Coefficient** measures the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. It is commonly used either to reduce the amount of calculation or to make the coefficient less sensitive to non-normality in distributions

*www.thefreedictionary.com

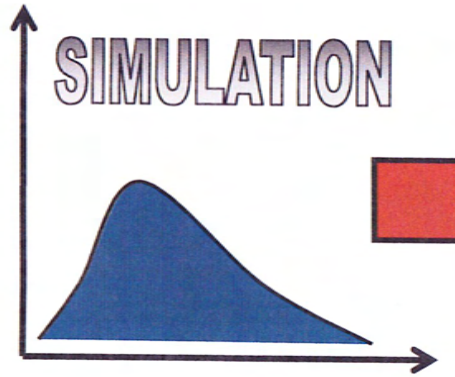
Pearson vs. Spearman in @Risk

Scitor Corporation

- **Example of the difference of Pearson and Spearman in @Risk**

DATASET

	SE Hrs	PM Hrs
Contract 1	1000	500
Contract 2	2000	650
Contract 3	2500	100
Contract 4	3000	900
Contract 5	3500	300
Contract 6	3800	1050
Contract 7	4000	450
Contract 8	5000	550
Contract 9	8000	800
Contract 10	10000	5000



CORRELATION OUTPUTS

Simulation Results	Spearman Input	Output Pearson	Output Spearman
Correlation SE-PM	0.442	0.227	0.440

Calculated correlation of 10,000 SE – PM trials to see if outputs matched the inputs

- Spearman holds
- Pearson does not

Simulation Results	Pearson Input	Output Pearson	Output Spearman
Correlation SE-PM	0.753	0.160	0.738

Pearson	0.753102
Spearman	0.442424

CALCULATIONS

SE Hrs - @Risk	=RiskPearson5(6.1405,30474,RiskShift(-1622.1),RiskName("SE Hrs"))
PM Hrs - @Risk	=RiskLoglogistic(23.126,585,1.9424,RiskName("PM Hrs"))

To provide consistent correlated results from dataset to simulation, it is recommended to calculate the correlation coefficient based on Spearman's method

Consistency in Application of Correlation

Scitor Corporation

- **Example of consistent application of correlation in Simulation Model**

CORRELATE SEED RVs

	Seed X	Seed Y
Uniform Rand()	0.965	0.574
Converted to Std Norm	1.816	0.186
Corr Matrix	X	Y
	1	
	0.5	1
Cholesky	1	0
	0.5	0.866
Transpose Cholesky	1	0.5
	0	0.866
Multiply by Std Norm	X	Y
	1.816	1.069
Convert back to Uniform	X	Y
	0.965	0.857

CONDUCT TRIALS

	Seed X	Seed Y	Result X	Result Y
Trial 1	0.625	0.836	3.318	4.469
Trial 2	0.947	0.844	4.612	4.515
Trial 3	0.010	0.025	0.671	0.072
Trial 4	0.718	0.470	3.576	2.887
Trial 5	0.368	0.286	2.664	2.152
:	:	:	:	:
Trial 45	0.176	0.401	2.070	2.623
Trial 46	0.486	0.541	2.965	3.154

Convert Seeds into Normal Distribution outputs:

- Results X = Normal(3,1)
- Results Y = Normal(3,1.5)

CORRELATE RESULTS

Correlation (X -Y)	Seeds	Results
Pearson	0.545	0.597
Spearman	0.519	0.519

Calculated correlation of Seeds and Results after 46 trials:

- Spearman correlation is consistent from seed to result
- Pearson correlation is not consistent from seed to result

Results

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

- **Spearman Rank Correlation Coefficient:**
  - Used in most simulation packages such as @RISK and Crystal Ball
  - Remains consistent from seeds to results
  - Does not rely on linear relationships
- **Pearson**
  - Is not used in simulation packages, however is found organically in Excel ("Correl" Function)
  - Does not remain consistent from seed to results
  - Statistically derived, but relies solely on presence of linear relationships

When injecting historical correlation into a simulation, it is best to calculate the Spearman rank correlation coefficient as it is consistent from application to results and guarantees the outputs will match the correlated inputs

Spearman User-Defined Function for Excel

Scitor Corporation

Steps for Inserting Spearman User-Defined Function (UDF) into Excel

- *Step 1:* Open Excel File where Spearman will be used
- *Step 2:* Hold "Alt" and Press "F11" to open VBA editor
- *Step 3:* Navigate to Insert > Module
- *Step 4:* Paste code from this slide into new Module
- *Step 5:* Use function in Excel by typing:
  - =Spearman(*Range1*, *Range2*) and pressing Enter

Key Assumptions:

1. Data is arranged in columns
2. There are no duplicate values within a Range

```
Public Function Spearman(Arr1 As Range, Arr2 As Range)

  Dim xiArr() As Integer
  Dim yiArr() As Integer
  Dim SumdiSq
  Dim Rows As Integer

  Rows = Arr1.Rows.Count

  If Arr2.Rows.Count <> Rows Then
    MsgBox "Ranges do not have the same number of rows. Try again."
    Exit Function
  End If

  ReDim xiArr(1 To Rows)
  ReDim yiArr(1 To Rows)

  SumdiSq = 0

  For i = 1 To Rows
    xiArr(i) = Application.WorksheetFunction.Rank(Arr1(i, 1), Arr1)
    yiArr(i) = Application.WorksheetFunction.Rank(Arr2(i, 1), Arr2)
    SumdiSq = SumdiSq + (xiArr(i) - yiArr(i)) ^ 2
  Next

  Spearman = 1 - ((6 * SumdiSq) / (Rows * (Rows ^ 2 - 1)))

End Function
```

Dividing by Average Rates

The Problem (1 of 2)

Scitor Corporation

- **In order to estimate the cost of a SW intensive program, most analysts use some form of the following equation:**
  - SLOC X Productivity Rate X Hourly Rate = $; Where
    - SLOC is some form of code count or distribution;
    - Productivity Rate is viewed as either SLOC/HR or HR/SLOC; and
    - Hourly Rate is viewed as $/HR
  - Leaving 2 main options for the calculation:

$$\frac{SLOC}{\left(\frac{SLOC}{HR}\right)} \times \frac{\$}{HR} = \$ \; or \; SLOC \times \frac{HR}{SLOC} \times \frac{\$}{HR} = \$$$

  - Either equation will work *iff* SLOC/HR is a single data point used as an analogy and therefore:

$$\frac{1}{\frac{SLOC}{HR}} = \frac{HR}{SLOC}$$

  - This is not necessarily the case, however, when using a dataset to determine an ***Average Productivity***

The Problem (2 of 2)

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

- **When using an average productivity some estimates erroneously come forward with the following productivity assumption in their equation:**

$$\frac{1}{\frac{\sum_{i=1}^{n}\frac{SLOC}{Hr}}{n}} = \frac{\sum_{i=1}^{n}\frac{Hr}{SLOC}}{n}$$

- **In reality, when broken down further we see that:**

$$\frac{1}{\frac{\sum_{i=1}^{n}\frac{SLOC}{HR}}{n}} = \frac{n}{\sum_{i=1}^{n}\frac{SLOC}{HR}} \neq \frac{\sum_{i=1}^{n}\frac{HR}{SLOC}}{n}$$

- **As a result, if Average Productivity is your metric of choice, use Hr/SLOC instead of SLOC/Hr:**

$$SLOC \times \frac{\sum_{i=1}^{n}\frac{HR}{SLOC}}{n} \times \frac{S}{HR} = S$$

- **Other alternatives to Average Productivity include using a Weighted Average or calculating the Geometric Mean for the dataset (instead of the Arithmetic Mean)**
  - Both the Wtd. Average and Geometric Mean will provide consistent inverse results for both SLOC/HR and HR/SLOC

Example

Scitor Corporation

1. DATASET

Contract	SLOC	HRS	SLOC/Hr	Hr/SLOC
Contract 1	2000	4000	0.50	2.00
Contract 2	3000	9000	0.33	3.00
Contract 3	5000	2500	2.00	0.50
Contract 4	10000	8500	1.18	0.85
Contract 5	12000	18000	0.67	1.50
Contract 6	4000	8800	0.45	2.20
Contract 7	6000	9600	0.63	1.60
Contract 8	9000	16200	0.56	1.80
Contract 9	9000	9000	1.00	1.00
Contract 10	8000	7600	1.05	0.95
Sum			8.36	15.40
n			10	10
Average			0.84	1.54

2. Myth

$$\frac{1}{\frac{\sum_{i=1}^{n}\frac{SLOC}{Hr}}{n}} = \frac{\sum_{i=1}^{n}\frac{Hr}{SLOC}}{n}$$

$$\frac{1}{\frac{8.4}{10}} = \frac{15.4}{10}$$

or

$$1.19 = 1.54$$

3. Reality

$$\frac{1}{\frac{\sum_{i=1}^{n}\frac{SLOC}{HR}}{n}} = \frac{n}{\sum_{i=1}^{n}\frac{SLOC}{HR}} = \frac{\sum_{i=1}^{n}\frac{HR}{SLOC}}{n}$$

$$\frac{1}{\frac{8.4}{10}} = \frac{10}{8.4} = \frac{15.4}{10}$$

$$\frac{1}{.84} = 1.19 = 1.54$$

When multiplying numbers together to develop a "factor," the inverse of the arithmetic average of a dataset will not equal the arithmetic average of the reciprocal dataset

Scitor Corporation

Dividing by a Distribution in a Simulation Model

The Problem

Scitor Corporation

- **Similar to "Dividing by an Average Rate," with other complications**

DATASET

Contract	SLOC	HRS	SLOC/Hr	Hr/SLOC
Contract 1	2000	4000	0.50	2.00
Contract 2	3000	9000	0.33	3.00
Contract 3	5000	2500	2.00	0.50
Contract 4	10000	8500	1.18	0.85
Contract 5	12000	18000	0.67	1.50
Contract 6	4000	8800	0.45	2.20
Contract 7	6000	9600	0.63	1.60
Contract 8	9000	16200	0.56	1.80
Contract 9	9000	9000	1.00	1.00
Contract 10	8000	7600	1.05	0.95

Different Methodology Options	Mean	St Dev
Option 1A: Lognormal (Sloc/Hr)	0.84	0.5
Option 1B: Lognormal (Hr/Sloc)	1.54	0.75
Option 2A: Lognormal Geo. (Sloc/Hr)	0.73	1.71
Option 2B: Lognormal Geo. (Hr/Sloc)	1.37	1.71
Option 3A:Lognormal Wtd Avg (Sloc/Hr)	0.73	0.46
Option 3B: Lognormal Wtd Avg (Hr/Sloc)	1.37	0.61

1. Parameters:
   1000 SLOC, $150/HR

2. Equation Options (w/ Productivity):

$$\frac{1000\ SLOC}{Lognormal(\mu,\sigma)\ of\ SLOC/HR} \times \frac{\$150}{HR} = \$$$

or

$$1000\ SLOC \times Lognormal(\mu,\sigma)\ of\ \frac{HR}{SLOC} \times \frac{\$150}{HR} = \$$$

3. Based on the equation choices above, and the proof in the previous section, there will be a problem with using the arithmetic mean for the lognormal equations for SLOC/HR and HR/SLOC
   - **1.54 is not the inverse of .84**

Simulation Results

Scitor Corporation

Simulation with lognormal distributions on productivity only

Only the Hr/Sloc Options have Mean Inputs and Mean Outputs match

Simulation Options	Mean Estimate Input				Mean Simulation Results		
Option 1A: Sloc/Hr	SLOC	Sloc/Hr	$/Hr	$	Mean $	St Dev	CoV
Lognormal(0.84, 0.50); CoV = 59%	1000	0.84	150	$ 179,336	$ 242,190	$ 143,174	59%
Option 1B: Hr/Sloc	SLOC	Hr/Sloc	$/Hr	$	Mean	St Dev	CoV
Lognormal(1.54, 0.75); CoV = 49%	1000	1.54	150	$ 231,000	$ 230,985	$ 112,217	49%
Option 2A: Geo. (Sloc/Hr)	SLOC	Sloc/Hr	$/Hr	$	Mean	St Dev	CoV
Lognormal(0.73, 1.71); CoV = 233%	1000	0.73	150	$ 205,262	$ 1,323,217	$3,026,591	229%
Option 2B: Geo. (Hr/Sloc)	SLOC	Hr/Sloc	$/Hr	$	Mean	St Dev	CoV
Lognormal(1.37, 1.71); CoV = 125%	1000	1.37	150	$ 205,262	$ 205,135	$ 252,729	123%
Option 3A: Wtd Avg (Sloc/Hr)	SLOC	Sloc/Hr	$/Hr	$	Mean	St Dev	CoV
Lognormal(0.73, 0.46); CoV = 63%	1000	0.73	150	$ 205,588	$ 286,544	$ 180,128	63%
Option 3B: Wtd Avg (Hr/Sloc)	SLOC	Hr/Sloc	$/Hr	$	Mean	St Dev	CoV
Lognormal(1.37, 0.61); CoV = 45%	1000	1.37	150	$ 205,588	$ 205,584	$ 91,599	45%

CoV's match in all Options

Regardless of Option, if you want the mean estimate and the mean of the simulation to be the same, productivity should be measured in HR/SLOC

Problem with Scaling

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

- **Dividing by the min of a SLOC/HR distribution vs. multiplying by the max of an HR/SLOC distribution yields different results for a "max" output**

Example:

Option 2A: Geo. (Sloc/Hr)	SLOC	Sloc/Hr	$/Hr	$	Mean	St Dev	CoV
Lognormal(0.73, 1.71); CoV = 233%	1000	0.73	150	$ 205,262	$ 1,323,217	$3,026,591	229%

Min. of Lognormal (0.73,1.71) = .0014 ⇒ $\frac{1000\ SLOC}{.0014\ \frac{SLOC}{HR}} \times \frac{\$150}{HR} = \$107M$

Option 2B: Geo. (Hr/Sloc)	SLOC	Hr/Sloc	$/Hr	$	Mean	St Dev	CoV
Lognormal(1.37, 1.71); CoV = 125%	1000	1.37	150	$ 205,262	$ 205,135	$ 252,729	123%

Max. of Lognormal (1.73,1.71) = 33.15 ⇒ $1000\ SLOC \times 33.15\ \frac{HR}{SLOC} \times \frac{\$150}{HR} = \$5M$

Dividing by a distribution (particularly when close to zero) will yield inaccurate simulation results

Overall Conclusions

Scitor Corporation

- **Correlation**
  - *When injecting historical correlation into a simulation, it is best to calculate the Spearman rank correlation coefficient as it is consistent from application to results and guarantees the outputs will match the correlated inputs*
- **Dividing by Average Rates**
  - *When multiplying numbers together to develop a "factor," the inverse of the arithmetic average of a dataset will not equal the arithmetic average of the reciprocal dataset*
    - **Dividing by SLOC/HR is *not always* the same as multiplying by HR/SLOC**
    - **Using a weighted average or the geometric mean alleviates this problem**
- **Dividing by Distributions**
  - *Dividing by distributions can yield inaccurate results in simulations*
    - **Developing a distribution for productivity that is measured in HR/SLOC will yield a mean estimate with the same results as the mean of a simulation**

Common Thread: Reliance on a COTS simulation package should not be a substitute for thorough analysis and model verification/validation

References

Scitor Corporation

- http://en.wikipedia.org/wiki/Karl_Pearson
- http://en.wikipedia.org/wiki/Charles_Spearman
- http://en.wikipedia.org/wiki/Francis_Galton
- http://www.thefreedictionary.com/correlation
- http://www.math.toronto.edu/mathnet/questionCorner/geomean.html
- http://www.mrexcel.com/forum/showthread.php?t=44080

- Robinson, Mitch and Sandi Cole. *"Rank Correlation in Crystal Ball® Simulations,"* June 2002 SCEA Conference