AN ABSTRACT OF THE THESIS OF

Maryam Agahi for the degree of Master of Science
in Industrial Engineering presented on August 9, 2013
Title: Rank and Linear Correlation Differences in Simulation and other Applications

Abstract approved:

---

David S. Kim

Monte Carlo simulation is used to quantify and characterize uncertainty in a variety of applications such as financial/engineering economic analysis, and project management. The dependence or correlation between the random variables modeled can also be simulated to add more accuracy to simulations. However, there exists a difference between how correlation is most often estimated from data (linear correlation), and the correlation that is simulated (rank correlation).

In this research an empirical methodology is developed to estimate the difference between the specified linear correlation between two random variables, and the resulting linear correlation when rank correlation is simulated. It is shown that in some cases there can be relatively large differences. The methodology is based on the shape of the quantile-quantile plot of two distributions, a measure of the linearity of the quantile-quantile plot, and the level of correlation between the two random variables. This methodology also gives a user the ability to estimate the rank correlation that when simulated, generates the desired linear correlation. This methodology enhances the accuracy of simulations with dependent random variables while utilizing existing simulation software tools.

# Rank and Linear Correlation Differences in Simulation and Other Applications

By
Maryam Agahi

A THESIS
Submitted to
Oregon State University

in partial fulfillment of
the requirements for the
degree of
Master of Science

Presented August 9, 2013
Commencement June 2014

Master of Science thesis of Maryam Agahi presented on August 9, 2013

APPROVED:

_____

Major Professor, representing Industrial Engineering

_____

Head of the School of Mechanical, Industrial and Manufacturing Engineering

_____

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

_____

Maryam Agahi, Author

ACKNOWLEDGEMENTS

I would like to express my gratitude to my academic advisor, Dr. David Kim for his guidance, patience, and engagement through this research. Without his assistance and knowledge this thesis would not have been possible.

My special appreciation goes to Dr. Lisa Madsen from Department of Statistics whose knowledge brought a new insight to address this research question.

Furthermore, I would like to thank my beloved friends for supporting me emotionally during course of this study.

Last but not least, I would like to thank my family who accompany me every steps of way even though we were far apart.

# CONTRIBUTION OF AUTHORS

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF APPENDIX FIGURES

Figure Page

# DEDICATION

To My Parents, Behzad and Azar.

# 1. Introduction

Simulation includes a broad collection of methods to imitate and predict the behavior of a real system, product, project, or scenario such as the selection of specific investment options. In this research the focus is on computer simulations that utilize pseudo random variates. These simulations are applied in situations where a performance measure is a function of one or more probability distributions (and perhaps other parameters and system dynamics) and an analytical derivation of the function is not feasible. This includes simulations of discrete event systems where the simulation of time is an integral component, and also "Monte Carlo" simulations where the passage of time is not directly simulated[1] (I.-T. Yang, 2005). Although both simulations utilize pseudo random variates, the results of this research apply more directly to Monte Carlo simulation. Examples of two common application areas of Monte-Carlo simulation are financial/engineering economic analysis, and project management, where simulation is used to quantify and characterize uncertainty and thus provide more complete information for decision makers (Cooper and Chapman, 1986). To add more accuracy to simulations, the dependence or correlation between the random variables utilized may also be simulated. However, there exists a difference between how correlation is most often estimated from data (linear correlation), and the correlation that is simulated (rank correlation) in the most popular Monte-Carlo simulation packages (add-in packages for spreadsheets). This difference was recognized by Garvey (2001), who questioned the validity of simulation results in this situation. This research examines the practical effects of this difference by developing methods to:

- Identify when differences in linear and rank correlation may be of concern,
- Estimate the difference between the rank correlation simulated and the linear correlation estimated,
- Estimate the simulated rank correlation that produces a desired linear correlation.

[1] Monte-Carlo simulation is sometimes used as a name for any simulation that utilizes pseudo random numbers and random variates.

The remainder of this paper is organized as follows. First, common measures of correlation are introduced, followed by an explanation of the methods for simulating correlated random variables. Quantile-Quantile plots are suggested as a basis to facilitate estimating Spearman's rank correlation from Pearson correlation, and the proposed empirical method is discussed. Since this research is relevant to applications of Monte-Carlo simulation, examples of implementing the proposed method on the simulation of sums of random variables, and the maximum of two random variables are presented.

## 2. Background Information

### 2.1. Measures of Correlation

The correlation between two random variables is a dimensionless measure (between -1.0 and 1.0) of the degree of association between their values. Multiple measures of correlation have been defined with two of the most common correlation measures being Pearson's product-moment correlation and Spearman's rank correlation (Yule & Kendal, 1950).

Pearson's product-moment correlation coefficient is a measure of the linear relationship between two random variables. If $X$ and $Y$ are two random variables then linear correlation is defined as:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad (1.1)$$

Where $Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\} = E(XY)-E(X)E(Y)$ is the covariance of $X$ and $Y$, and $Var(X)$ and $Var(Y)$ are the variance of $X$ and $Y$ respectively. If two random variables are linearly dependent, then $|\rho(X,Y)| = 1$, and if two random variables are independent, then $\rho(X,Y) = 0$, although zero correlation does not imply the independence of two random variables (Balakrishnan & Lai, 2009). A positive value of $\rho$ indicates a direct relationship, while a negative sign indicates an inverse relationship. The sample linear correlation coefficient $r$, for a sample of $n$ bivariate observations $(x_1, y_1), \ldots, (x_n, y_n)$ is:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \qquad (1.2)$$

where $\bar{x}$ and $\bar{y}$ are the respective sample means of the $x_i$ and $y_i$ values. *r* is equal to the square root of the coefficient of determination of a linear regression model of *y* as a function *x*, commonly known as "$R^2$" (Neter, Wasserman, & Kutner, 1989).

Spearman's rank correlation coefficient is another bivariate measure of association, showing the strength of the monotone relationship between two random variables. Spearman's rank correlation coefficient $\rho_s$, is based on the ranks of the $x_i$, and the ranks of the $y_i$ . If $(X_i, Y_i)$, $(X_j, Y_j)$ and $(X_k, Y_k)$ are three independent pairs of random variables, then $\rho_s$ is defined to be proportional to the probability of concordance minus the probability of discordance for two pairs $(X_i, Y_i)$, $(X_j, Y_k)$.

$$\rho_s = 3\left\{p\{(X_i - X_j)(Y_i - Y_k) > 0\} - p\{(X_i - X_j)(Y_i - Y_k) < 0\}\right\} \qquad (1.3)$$

It is observed that Spearman's rank correlation coefficient between *X* and *Y* is Pearson's product-moment correlation coefficient between the uniform variates $U \sim Unif(0,1)$ and $V \sim Unif(0,1)$ (Balakrishnan & Lai, 2009).

$$\rho_S = \frac{E(UV) - \frac{1}{4}}{\frac{1}{12}} \qquad (1.4)$$

There are two commonly used estimators (denoted $r_s$) of Spearman's rank correlation coefficient. One estimator is formula 1.2 with the $x_i$ and $y_i$ values replaced by their ranks. The other estimator of Spearman's rank correlation is:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)} \qquad (1.5)$$

where $d_i$ is the differences between the ranks of $x_i$ and $y_i$ in each $(x_i, y_i)$ pair (Sheskin, 2011). The two estimators are the same as long as there are no ties in the ranks of the $x_i$, and no ties in the ranks of the $y_i$ .

Since Spearman's correlation coefficient measures correlation between ranks rather than the *X* and *Y* values, it is unaffected by any monotone transformation of these variables, in contrast to Pearson's correlation coefficient, which is only unaffected by linear transformations of *X* and *Y* (Balakrishnan & Lai, 2009).

If two populations are normal, the relationship between $\rho_s$ and $\rho$ is (Kendall & Gibbons, 1990):

$$\rho = 2sin\frac{1}{6}\pi\rho_s \qquad (1.6)$$

Similar results hold for samples from any elliptically contoured distributions that generalize multivariate normal distribution and inherits its properties (Mildenhall, 2006). Pearson's correlation coefficient is a perfect measure of association for normally distributed variables because it accurately specifies the dependence between the marginal distributions, but it is not always accurate with non-normal variables (Mildenhall, 2006).

Besides the Pearson and Spearman correlation coefficients, there are other methods for describing the strength of association between two random variables such as Kendall's rank correlation coefficient $\tau$, which is a measure of the probabilities of concordance between pairs of observations from two distributions of interest (Kendall & Gibbons 1990 ).

2.2. Correlation in Education and Commonly Used Software

Despite the existence of multiple correlation measures the term "correlation" is typically equated to Pearson's correlation coefficient in undergraduate educational programs. In a selection of 12 introductory statistics and engineering statistics textbooks which included Devore (1999), Hines & Montgomery (1972), Hines, Montgomery, Goldsman, & Borror (2003), and Walpole (2007), correlation is defined as Pearson's linear correlation with little or no mention of other measures. Additionally, the use of computer spreadsheet software has become commonplace in education and practice. In software such as Microsoft Excel there is a single correlation function, which computes Pearson's correlation coefficient.

In contrast, correlation in simulation software typically implements Spearman's correlation. The development of spreadsheet add-in software for generating observations from a wide selection of random variables has made Monte-Carlo simulation much more accessible to analysts and engineers. Two widely used commercial spreadsheet add-in packages for Monte-Carlo simulation are @RISK and Crystal Ball. These add-ins are widely used by engineers (Ali Touran and Suphot 1997, Seila 2001, Yang 2005, and Mildenhall 2006). These software packages have the ability to simulate dependent

random variables by inputting a "correlation coefficient". It is reasonable to believe that most engineers would interpret this correlation to be Pearson's correlation, however the correlation simulated is Spearman's correlation, which is straightforward to simulate. The ramifications of this different interpretation of correlation are examined in this paper.

## 3. Literature Review

Given the prior discussion, it is reasonable to assume that practitioners may estimate correlation as Pearson's correlation and simulate correlation as Spearman's correlation in a simulation model. The literature reviewed falls into one of three categories that are relevant to this situation: 1) The use of Monte-Carlo simulation with correlated random variables, 2) Methods available for simulating correlated random variables, 3) Research addressing or recognizing this situation.

### 3.1. Applications of Monte-Carlo simulation with correlated random variables

While there are many Monte-Carlo simulation application areas, two common applications where the need to simulate dependent random variables has been recognized will be discussed.

Monte-Carlo simulation methods are commonly applied in various financial analysis and project management studies (BadriAmr, 1997; Tummala & Burchett, 1999; Elkjaer, 2000). Discussions about the necessity of considering statistical dependencies to avoid underestimating total cost have also been presented in the literature (Diekmann, 1983; Raftery, 1994; Ranasinghe, 2000). Touran and Wiser (1992) used multivariate lognormal distributions to generate Pearson correlated random variables and calculated the total cost of a construction project from historic data that showed Pearson correlation was an effective means of modeling dependent construction costs. Difficulties in simulating Pearson correlation led Touran and Suphot (1997) to investigate the use of rank correlation in simulating construction cost simulations. Tests showed that the simulation of rank correlation resulted in simulated distributions that were close to the actual distributions. Wall (1997) also chose rank correlation over linear correlation in simulating

dependent costs because it relies on fewer assumptions about the distribution of data (e.g., normality assumptions).

Monte-Carlo simulation has also been used to quantify the uncertainty in project schedules using Critical Path Methods. Rank correlation was used to model the statistical dependency of activities in a project network in Van Dorp and Duffey (1999). Yang (2007) also modeled dependent project task durations using a simulation method proposed by Cario and Nelson (1997) that allows arbitrarily specified marginal distributions for task durations and any desired correlation structure.

### 3.2. Methods for simulating correlated random variables

Multiple techniques are available for generating correlated values from univariate distributions, and other methods work when the particular multivariate distributions are fully specified (Lurie & Goldberg, 1998). Given that joint distribution functions are often difficult to specify, methods that use marginal distributions to simulate dependent data have been developed.

#### 3.2.1 Rank correlation

Iman and Conover (1982) use a normalizing transformation and Cholesky decomposition in a method for simulating a desired rank correlation matrix for multivariate random variables. The Iman and Conover (1982) method operates on random variates generated from the marginal distributions of interest, and generates permutations of these values that approximates a desired rank correlation structure.

Although this distribution-free approach was aimed to generate random variables with a desired rank correlation while preserving the intent of the sampling scheme, it is misinterpreted in some literatures to be able to simulate dependent data for a specific linear correlation (Mildenhall, 2006). Several spreadsheet simulation software packages including Oracle Crystal Ball and @Risk have implemented this method to simulate dependent variables.

Copulas are an alternative method for simulating desired rank correlation that are very similar to Iman and Conover Algorithm. Copulas have been also used for simulating correlated random variables in risk assessment (Hass 1999). Two of the key differences between the normal copula method and Iman and Conover Algorithm are:

1. The normal copula method corresponds to the Iman and Conover method when the latter is computed using normal scores (a normal score for observation $i$ of sample of size $n$ is equal to $\varphi^{-1}(\frac{i}{n+1})$ and rescaled to have standard deviation 1).

2. The Iman and Conover method works on given sample from a marginal distribution, whereas the normal copula method generates samples by inverting the distribution function during the simulation process.

In addition, a general method to incorporate correlations between cost elements is discussed by Yang (2005). The proposed method first checks the feasibility of the correlation matrix (Pearson or Spearman's) and does the necessary adjustment and before starting the simulation phase.

### 3.2.2. Linear correlation

Li and Hammond (1975) developed an approach where marginal distributions are specified, and a method is developed to determine an intermediate correlation matrix from the desired correlation matrix. However, the computations required in this method are extremely time consuming.

Johnson and Ramberg (1978) viewed the marginal distributions as transformations of normal distributions to impose a correlation structure, but the mathematics of their method becomes intractable for distributions other than the lognormal and inverse hyperbolic sine.

Lurie and Goldberg (1998) used the Li and Hammond idea (1975) and developed an approximate method applicable to any sets of continuous distributions.The method simulates random variables based on a marginal distribution and a Pearson correlation matrix using Cholesky decomposition and Gauss-Newton iteration. This algorithm tries to minimize the difference between the correlation matrix of an initial lower triangular

matrix and the original correlation matrix in several iterations. Most simulation software packages implement the Iman and Conover algorithm, but Risk+™ (a schedule analysis add-in to Microsoft@ Project) implements the Lurie and Goldberg algorithm for generating correlated variables.

3.3.Research addressing or recognizing the problem

Garvey (2001) claimed that Spearman's correlation is not appropriate for cost risk analyses because sums of random variables representing costs involve only Pearson's product-moment correlation and not Spearman's rank correlation, and also because Pearson and Spearman's correlation can be very different. If we assume that total cost of a system can be calculated as:

$$Cost_{total} = X_1 + X_2 + X_3 + \ldots + X_n \qquad (2.1)$$

where $X_i$ is the *ith* cost element of the system, the mean and variance of $Cost_{total}$ can be obtained from the following formulas:

$$\mathrm{E}\,(Cost_{total}) = E(X_1) + E(X_2) + E(X_3) + \ldots\ldots + E\,(X_n) \qquad (2.2)$$

$$Var\,(Cost_{total}) = \sum_{i=1}^{n} Var\,(Xi) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov\,(X_i, Y_j)$$

$$= \sum_{i=1}^{n} Var\,(Xi) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \rho\,(X_i, Y_j)\,\sigma_{X_i}\sigma_{Y_j} \qquad (2.3)$$

Since the correlation in equation 2.3 is linear correlation, Garvey (2001) stated that it is not appropriate to use other methods such as rank correlation for estimating dependencies in cost estimation simulation. Price (2002) also believes that product-moment correlation is the appropriate measure for cost-schedule risk analysis since durations and costs are interval and not ordinal measures. In contrast, a study presented in 2004 at the Society of Cost Estimating and Analysis meeting concluded the opposite of Garvey (2001), however the distributions used in this study were all identical triangular distributions and the results were not generalized (Robinson & Cole, 2004).

## 4. Estimating Pearson and Spearman's Correlation Differences

Given two random variables $X$ and $Y$ with distribution functions $F_X$ and $F_Y$ that are linearly correlated with ρ estimated from sample data, an empirical approach was taken to quantify the difference between the Spearman correlation simulated and the Pearson's correlation that is realized. "S-P Diff" will be used to refer to this difference. A Quantile-Quantile plot (Q-Q plot) was used to compare two distributions and features of these plots become the basis of the empirical model to estimate "S-P Diff". An experiment was designed to identify factors (features of the Q-Q plots) that have a significant impact on the value of "S-P Diff", and then these factors were used in a regression model to estimate the "S-P Diff". The data used in this analysis come from simulations run using the Crystal ball Excel Add in. In this section the Quantile-Quantile plot and its properties are introduced, and then the simulation procedure, analysis of experiments, and regression analysis are discussed in detail.

A theoretical Q-Q plot is a plot of the percentage points or quantiles of one distribution against the corresponding percentage points or quantiles of another distribution (Fowlkes, 1987). If $F(y)$ is the cumulative distribution function of random variable $Y$, then quantile of $F$, is equal to $Q_t(p)$ which satisfies the following equation for $p\ (0<p<1)$:

$$Q_t(p)=F^{-1}(p) \qquad (3.1)$$

Examples of Q-Q plots are presented in the Appendix. In a Q-Q plot, $Q_t(p_i)$ of one distribution is plotted versus $Q_t(p_i)$ of another distribution. Q-Q plots can be used to compare the shape of two distributions with each other. When two distributions have similarly shaped density functions, the points on the Q-Q plot will fall near the line $y = x$ (Chambers, Cleveland, Beat, & Tukey, 1983). The shape of the Q-Q plot is invariant under linear transformation of the coordinate axes, which means that two distributions that differ only in scale or location yield a linear Q-Q plot (Fowlkes, 1987). If large departures from a straight line Q-Q plot are observed, it will indicate that the shapes of

two distributions are not similar. The reason for departures from a straight line may fall in one of the following categories:

1. When one distribution has longer or shorter tails than the other. This results in a S-Shaped Q-Q plot. If the left end of the plot is below the $x = y$ line and the right line is above the $x = y$ line, the distribution on the y-axis has a longer tail.

2. Convex or Concave Curvature is related to the asymmetry of the distributions. It shows the distribution on the y-axis is more skewed than the other distribution. For example, an increasing slope from left to right indicates the distribution in the y-axis is more skewed to the right.

In summary, we can conclude that a Q-Q plot is either a straight line, S-shaped, or convex or concave (Chambers et al., 1983). A measure of the straightness of a Q-Q plot is the correlation coefficient (Tsai & Yang, 2005) or coefficient of determination of linear regression.

Q-Q plots are related to correlation coefficients. Several measures of correlation have been defined based on the difference between the probability of concordance and discordance of two random variables. One method calculates this probability using three independent pairs of observation $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$. Using crossed observations $(X_1, Y_1), (X_2, Y_3)$, the probability of concordance is equal to

$$\iota_c = Pr\{(X_1 - X_2)\,(Y_1 - Y_3) > 0\} \qquad (3.2)$$

And the difference between two probabilities is calculated as followings:

$$\iota_c - \iota_d = Pr\{(X_1 - X_2)\,(Y_1 - Y_3) > 0\} - Pr\{(X_1 - X_2)\,(Y_1 - Y_3) < 0\} = 2\iota_c - 1 \quad (3.3)$$

The minimum and maximum values for the above formula are $-\frac{1}{3}$ and $\frac{1}{3}$, which can be transformed to a -1 to 1 scale by multiplying by 3. If each $X_i$ is replaced by $X_i^* = F(X_i)$ where $F$ is the cumulative distribution function of $X$ and $Y_i$ by $Y_i^* = G(Y_i)$ where $G$ is the cumulative distribution function of $Y$, the quantities in 3.2 and 3.3 will not be affected. This correlation between $X_i^*$ and $Y_i^*$ has been called the Grade correlation coefficient

between $X$ and $Y$ and is equal to Spearman's rank correlation coefficient when applied to finite samples (Kruskal, 1958, Kendall & Gibbons, 1990).

When two random variables $X$ and $Y$ with known probability distributions have perfect Spearman rank correlation, pairs of data $(X_i, Y_i)$ generated from the distributions will fall on the line represented by theoretical Q-Q plot of those random variables. The theoretical Q-Q plot contains all pairs of $(X_i, Y_i)$ when $X_i = F^{-1}(X_i^*)$ and $Y_i = F^{-1}(Y_i^*)$. Similarly, if quantiles of one distribution are plotted against the quantile compliments of the other distribution, the graph represents pairs having perfect negative Spearman's correlation. Thus Q-Q plots are able to show perfect Spearman's rank correlation.

To find a possible relationship between a Q-Q plot and "S-P Diff", a designed experiment was conducted to identify characteristics of the Q-Q plot affecting "S-P Diff". The data for this experiment are from a series of simulations with different combinations of distributions, and different levels of correlation. The simulations were implemented using Crystal Ball simulation software. Experiments were created using combinations of the following factors:

1. Shape of their Q-Q plot $(S_j)$ that is categorized as either S-Shaped (1) or Concave or Convex (2).
2. Coefficient of determination of their Q-Q plot $(R_j)$, this value will fall into one of the following categories:

$0 \leq R_j \leq 0.2$

$0.2 < R_j \leq 0.4$

$0.4 < R_j \leq 0.6$

$0.6 < R_j \leq 0.8$

$0.8 < R_j \leq 1$

3. Level of Spearman's correlation simulated. All distribution combinations were simulated using four levels of correlation (0.65,0.75,0.85,0.95 for positive correlation and -0.65, -0.75, -0.85, -0.95 for negative correlation).

4. The distribution combination is considered a random factor in this experiment because each distribution combination was selected randomly from among all possible combinations. This factor is nested within combinations of the first two factors.

The number of paired observations generated in each simulation was 10,000 and each simulation was repeated 10 times. The total experiment size was equal to 2000. After each simulation run, the 10,000 paired values were extracted and the sample Pearson and Spearman's correlation were calculated.

The expected mean square values for a factorial design with a nested random factor was used to test the effect of each factor and their interactions on "S-P Diff". The model for this experiment is:

$$y_{ijkl} = \mu + S_i + R_j + D_{k(ij)} + C_l + SR_{ij} + SC_{il} + RC_{jl} + DC_{k(ij)l} + SRC_{ijl} + e_{ijkl} \quad (3.4)$$

The notation used in this model is:

$S_i$- Shape of QQ-Plot

$R_j$- Coefficient of determination of Q-Q plot

$D_k$-Distribution combination, random factor nested in shape and coefficient of determination

$C_l$-levels of correlation for simulation

$e_{ijkl}$-error term

The factors and their interactions having a significant effect were used in a regression model to estimate "S-P Diff". Regression analysis was conducted using Minitab Software. Both full regression and reduced models were constructed and compared. In order to choose the best fit between several regression models, the significance of each parameter was considered, and also the usability of the model was taken into account.

## 5. Results

The analysis of the nested factorial design results shows a significant effect for all factors and their interactions (p-value < 0.05), which is partly caused by the high degrees

of freedom (1,800) for the error term in the ANOVA. The F-test statistic values were used to obtain the most significant factors in the model. The F-test statistic values for, shape (*S*) and coefficient of determination (*R*), are notably higher than the other test statistics, with values equal to 35,532.97 and 18,715.97 respectively. The third largest F-test statistic value is for the interaction between *S* and *R* ($F$=2415.209), followed by the levels of correlation for the simulation (*C*) with $F$ =799.67. The F-test statistics for distribution combination ($D_k$) is equal to 283.97, and is noticeably smaller than other factors mentioned. Since $D_k$ is a random factor and has a relatively low F-test statistic value, the variance contribution of $D_k$ and its interactions was ignored and considered part of the model error term. Other factors and interactions were included in the regression analysis.

The linear regression analysis was conducted in two ways:

1. Includes only responses with positive correlation levels.
2. Includes only responses with negative correlation levels.

For each regression analysis, full and reduced models were constructed and compared based on their p-values. Table 5.1 summarizes the results:

Table 5.1- Regression analysis significant factors

| Analysis | Significant factors in model |
|---|---|
| 1 | All main effect factors<br>Shape interaction with coefficient of determination<br>Shape interaction with correlation level |
| 2 | All main effect factors<br>Shape interactions with other factors |

The adjusted $R^2$ value for models with significant factors ranged between 79.11 % and 90.16 %, which shows how well the model explains the observed outcomes. For estimating the “S-P Diff” without knowing correlation levels, the models selected are as follows:

Table 5.2- Linear Regression models for *S-P Diff*

*(S-P Diff= Estimated Spearman's correlation – Estimated Pearson correlation from simulation results)*

| Option | Analysis | Regression Model | Adjusted $R^2$ |
|---|---|---|---|
| 1 | 1 | *S-P Diff = 0.317109 - 0.156712 S- 0.698402 R+ 0.466213 C+ 0.269173 S×R- 0.142042 S×C* | 90.54% |
| 2 | 1 | *S-P Diff = 0.316109 - 0.270548 S- 0.697854 R+ 0.466926 S+ 0.269803 S×R* | 90.16% |
| 3 | 1 | *S-P Diff = 0.690305 - 0.269139 S - 0.699916 R + 0.26801 S×R* | 86.02% |
| 4 | 2 | *S-P Diff = -0.27485 + 0.11761 S+ 0.614687 R+ 0.346648 C - 0.269641 S*R - 0.135873 S×C* | 79.49% |
| 5 | 2 | *S-P Diff = -0.274661 + 0.226294 S + 0.614429 R + 0.346719 C - 0.269664 S *R* | 79.11% |
| 6 | 2 | *S-P Diff = -0.551999 + 0.225958 S+ 0.614486 R - 0.269004 S×R* | 76.57% |

In this research, estimating S-P Diff without knowing Spearman's rank correlation is of interest and hence, option 3 and option 6 were selected for estimating S-P Diff when the correlation is positive or negative respectively.

## 6. Model implementation for common applications

Although an empirical model for the relationship between Pearson and Spearman's correlation coefficient has been developed, the impact of simulating Spearman correlation instead of Pearson correlation in models was also examined. The goal is to estimate the effect of simulating Spearman's correlation instead of the Pearson correlation that is estimated.

Two cases when two random variables with known Pearson correlation $(\rho_1)$, coefficient are simulated were examined. In case one $\rho_1$ is input as the desired correlation level in Crystal ball (which simulates Spearman correlation), and in case two $\rho_1' = \rho_1 + SP\ Diff$ is input as the desired correlation level.

Table 6.1- Cases considered when simulating two random variables with known Pearson correlation

| Case | Value input as the desired correlation level | Estimated Pearson correlation simulated |
|---|---|---|
| 1 | $\rho_1$ | $\rho_1 - SP\ Diff$ |
| 2 | $\rho_1' = \rho_1 + SP\ Diff$ | $\rho_1$ |

Simulating case 1 instead of case 2 may result in estimating a mean and standard deviation with less accuracy in a larger simulation with multiple pairs of correlated random variables. The effect on model results will depend on several factors such as: the function simulated, the strength of the relationship between random variables, and the type of distributions involved in the simulation. As mentioned before, financial/engineering economic analysis and project management are two popular applications of Monte-Carlo simulation. The simulated functions in these applications include the summation and maximum of random variables.

Consider two random variables $X_1$ and $X_2$ with known distribution functions, and known means ($\mu_1$ and $\mu_2$) and variances ($\sigma_1^2$ and $\sigma_2^2$). If $Y=X_1+X_2$ and the Pearson correlation coefficient between $X_1$ and $X_2$ is equal to $\rho_1$ ($-1 \leq \rho_1 \leq 1$), then an engineer could mistakenly input $\rho_1$ to define correlation, or instead use one of the regression models represented in table 5.1 and calculate the corresponding Spearman's correlation ($\rho_1'$) to get the desired Pearson correlation ($\rho_1$) and input $\rho_1'$ to define correlation. No matter what levels of correlation input in the software, the value representing the mean of $Y$ would be similar. In contrast, the estimated variance of the value of interest $(Y)$ changes accordingly (Table 6.2).

Table 6.2- Predicted results for simulations of $Y=X_1+X_2$

| Case | Value input in correlation assumption cell in Crystal Ball | Pearson Correlation | *Mean (Y)* | *Variance (Y)* |
|---|---|---|---|---|
| 1 | $\rho_1$ | $\rho_1 - SP\ Diff$ | $\mu_1 + \mu_2$ | $\sigma_1^2 + \sigma_2^2 + 2(\rho_1 - SP\ Diff)\,\sigma_1\,\sigma_2$ |
| 2 | $\rho_1' = \rho_1 + SP\ Diff$ | $\rho_1$ | $\mu_1 + \mu_2$ | $\sigma_1^2 + \sigma_2^2 + 2\rho_1\,\sigma_1\,\sigma_2$ |

The percent difference in the simulated coefficient of variation (CV) of $Y=X_1+X_2$ from case 1 to 2 is:

$$Percentage\ of\ Difference\ in\ CV = \frac{\sqrt{\sigma_1^2+\sigma_2^2+2(\rho_1-SP\ Diff)\,\sigma_1\,\sigma_2}-\sqrt{\sigma_1^2+\sigma_2^2+2\rho_1\,\sigma_1\,\sigma_2}}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho_1\,\sigma_1\,\sigma_2}} \quad (6.1)$$

For one combination of distributions with known $S$ and R, "S-P Diff" could be calculated based on option 3 or 6 from table 5.2, equation 6.1 could be considered as an objective function of a nonlinear optimization problem subject to constraints on $S$, $R$ and $\rho_1$. Using Excel Solver solutions for the maximum of equation 6.1 indicates that the *Percentage of Difference in CV* could be greater than 40 % for some extreme cases. It also shows that for both negative and positive correlation, combinations of distributions with concave or convex Q-Q plots and smaller $R$ values result in extreme differences in the coefficient of variations. Table 6.2 shows two extreme cases with values of equation 6.1.

Table 6.3- Variables and objective function value for maximizing equation 6.1

| Correlation Sign | Shape | $R^2$ of Q-Q Plot | Target Pearson Correlation | Objective function |
|---|---|---|---|---|
| Positive | Concave/Convex | 0.11 | 0.15 | 48 % |
| Negative | Concave/Convex | 0.1 | -0.31 | 41 % |

The maximum values of 6.1 are not realistic for the applications discussed in this paper, but it is advisable to calculate this difference based on real data for specific problems. It can be concluded that lower levels of the coefficient of determination of Q-Q plots will result in less accurate results that are worse with concave or convex plots.

Although there is no general closed formula available for the mean and variance of the maximum of two random variables ($Y=Max\ (X_1,X_2)$), the effect of simulating Spearman's correlation instead of Pearson's correlation can be estimated by simulation. For this matter, combinations of distributions with different Q-Q plot shapes and $R^2$ values were selected and simulated with different levels of correlation categorized in three groups (low, medium and high). For each combination of $X_1$ and $X_2$, $Y$ was simulated for both case1 and case2 from Table 6.1. In order to standardize the response values from the

experiments, frequency of having one variable ($X_1$ or $X_2$) been selected as $Y$ is also simulated. Since, in project task duration simulation, when two tasks reach the same nodes, total project durations varies depending on which task is selected as maximum, if mean of frequency changes significantly, the total project durations might change. Hence, the absolute percentage of difference between mean of frequency for two cases were calculated and used as the response value for the experimental design. The factors in this experiment are summarized in Table 6.3.

Table 6.4. Experimental design factors to test the effect of simulating Spearman's correlation on the maximum of two random variables

| Name | Notation | Number of Levels |
|---|---|---|
| Shape of Q-Q Plot | $S_i$ | 2 (S and Concave/Convex) |
| $R^2$ of Q-Q plot | $R_j$ | 3 (Low, Medium, High) |
| Sign of Target Correlation | $T_k$ | 2 (Positive, Negative) |
| Target Correlation Level | $C_l$ | 3 (Low, Medium, High) |

The model of this experiment is as follow:

$$y_{ijkl} = \mu + S_i + R_j + T_k + C_{l(ijk)} + SR_{ij} + ST_{ik} + RT_{jk} + SRT_{ijk} + e_{ijkl} \quad (6.2)$$

Results from the ANOVA show that shape of Q-Q Plot, $R^2$ of Q-Q plot, sign of target correlation coefficient and the interaction of $T_k$ have significant effect on the response value (P-value <0.05). Sign of target correlation has the highest impact on the response value following by its interaction with shape and also the $R^2$ of Q-Q plot. Observations show larger differences when two random variables have positive correlation, with lower levels of coefficient of determination, and concave or convex Q-Q plot shape. Although the magnitude of the difference varied for each individual observation, the largest difference between mean of the frequency that one variable is selected as $Y$ in this experiment was 44 %.

## 7. Conclusion

Although equation 1.4 shows the relationship between $\rho_s$ and $\rho$ when both random variables having normal distributions, there is no general formula to model the relationship between the two correlation coefficient for non-normal distributions. This research shows that the difference of the two correlation coefficient estimates can be modeled as a function of the shape of the quantile-quantile plot of the two distributions, and the measure of the linearity of the quantile-quantile plot. This model can be beneficial for estimating the Spearman's rank correlation coefficient for a certain combination of distributions and a desired Pearson correlation coefficient. Since the Pearson correlation coefficient is what is normally estimated from data, this methodology will help users estimate the corresponding Spearman's rank correlation to use with simulation software to increase the accuracy of the simulation results by producing the desired Pearson correlation. This method is applicable to any problem that requires the estimation of one coefficient from the other. Also, future research could be conducted to model this relationship analytically rather than empirically.

Although, Garvey claimed that the sums of random variable only includes the linear correlation coefficient and it is not correct to use rank correlation to simulate dependent costs, this research shows that as long as the correlation coefficient captures the dependency characteristics correctly, using rank correlation coefficient to simulate dependent costs results in accurate outcomes with respect to mean and variance. This is possible because of the fact that a certain level of Pearson correlation coefficient for specific combinations of two random variables triggers a certain level of Spearman's rank correlation coefficient. Furthermore, it becomes clear that simulating Spearman's correlation instead of Pearson correlation can lead to large differences, but in most cases it should not have a large effect in the applications of interest.

## 8. Biblography

BadriAmr, M. A. (1997). Effective analysis and planning of R&D stages: a simulation approach. International Journal of Project Management, 15(6), 351–358.

Balakrishnan, N., & Lai, chin-D. (2009). Continuous Bivariate Distributions (2nd ed.). Springer.

Cario, M. C., & Nelson, B. L. (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Department of Industrial Engineering and Management Science, Northwestern University, Evanston, Illinois.

Chambers, J. M., Cleveland, W. S., Beat, K., & Tukey, P. A. (1983). Graphical Methods for Data Analysis. Wadsworth International Group and Duxbury Press.

Cooper, D. F., & Chapman, C. B. (1986). Risk Analysis for Large Projects, Models, Methods and cases. John Wiley & Sons Inc.

Devore, J. L. (1999). Probability and Statistics for Engineering and Sciences (5th ed.).

Diekmann, J. E. (1983). Probabilistic estimating: mathematics and applications. Journal of Construction Engineering and Management, 109(3), 297–308.

Elkjaer, M. (2000). Stochastic budget simulation. International Journal of Project Management, 18(2), 139–147.

Fowlkes, E. B. (1987). A folio of Distributions : A collection of Theoretical Quantile-Quantile Plots (Vol. 78). Department of Statistics, Southern Methodist University, Dallas, Texas: Marcel Dekker, Inc.

Hass, C. N. (1999). On Modeling Correlated Random Variables in Risk Assessment. Risk Analysis, 19, 1205–1214.

Hines, W. W., & Montgomery, D. C. (1972). Probability and statistics in engineering and management science. New York: Ronald Press.

Hines, W. W., Montgomery, D. C., Goldsman, D. M., & Borror, C. M. (2003). Probability and Statistics in Engineering (4th ed.). John Wiley & Sons, Inc.

Iman, R. L., & Conover, W. J. (1982). A Distribution-Free Approach to Inducing Rank Correlation among Input Variables. Communications in Statistics - Simulation and Computation, 11(3), 311–334.

Jhonson, M. ., & Ramberg, J. . (1978). Transformations of the multivariate normal distribution with applications to simulation. International conference on systems sciences, Honolulu, HI, USA.

Kendall, M., & Gibbons, J. D. (1990). Rank Correlation Methods (5th ed.). Oxford University Press.

Kruskal, W. H. (1958). Ordinal measures of association. Journal of the American Statistical Association, 814–861.

Li, S. T., & Hammond, J. L. (1975). Generation of Pseudo-Random Numbers wih Specified Univariate Distributions and Correlation Coefficients. IEEE Transactions on Systems, Management and Cybernetics, (5), 557–561.

Lin, G. D., & Huang, J. S. (2010). A note on the maximum correlation for Baker's bivariate distributions with fixed marginals. Journal of Multivariate Analysis, 101(9), 2227–2233. doi:10.1016/j.jmva.2010.04.005

Lurie, P. M., & Goldberg, M. S. (1998). An approximate method for sampling correlated random variables from partially-specified distributions. Management Science, 44(2), 203–218.

Mildenhall, S. (2006). The Report of the Research Working Party on Correlations and Dependencies Among all Risk Sources, Part1. Casual Actuarial Society Forum: CAS Research Working Party on Correlations and Dependencies Among all Risk Sources.

Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied Linear Regression Models (2nd ed.). Richard D. IRWIN. INC.

Price, J. (2002, June 14). An Implementation of the Lurie-Goldberg Algorithm in Schedule Risk Analysis. SCEA 2002 National Conference Scottsdale, Arizona.

Raftery, J. (1994). Risk Analysis in Project Management. E&FN Spon.

Ranasinghe, M. (2000). Impact of correlation and induced correlation on the estimation of project cost of buildings. Construction Management and Economics, (18), 395–406.

Robinson, M., & Cole, S. (2004, June). Rank Correlation in Crystal Ball simulations. Presented at the The Society of Cost Estimating and Analysis Annual Meeting, Manhattan Beach, Ca.

Seila, A. F. (2001). Spreadsheet Simulation. In Proceedings of the 2001 Winter Simulation Conference.

Sheskin, D. J. (2011). Handbook of Parametric and Nonparametric Statistical Procedures (5th ed.). Chapman and Hall/CRC.

Touran, A., & Wiser, E. P. (1992). Monte Carlo technique with correlated random variables. Journal of Construction Engineering and Management, 118(2), 258–272.

Touran, Ali, & Suphot, L. (1997). Rank Correlation in Simulating Construction Costs. Journal of Construction Engineering and Management, 123(3), 297–301.

Tsai, D.-M., & Yang, C.-H. (2005). A quantile–quantile plot based pattern matching for defect detection. Pattern Recognition Letters, 26(13), 1948–1962. doi:10.1016/j.patrec.2005.02.002

Tummala, V. M., & Burchett, J. F. (1999). Applying a risk management process (RMP) to manage cost risk for an EHV transmission line project. International Journal of Project Management, 17(4), 223–235.

Van Dorp, J. R., & Duffey, M. R. (1999). Statistical dependence in risk analysis for project networks using Monte Carlo methods. International Journal of Production Economics, 58(1), 17–29.

Wall, D. M. (1997). Distributions and Correlations in Monte Carlo Simulation. Construction Management and Economics, (15), 241–258.

Walpole, R. (2007). Probability & statistics for engineers & scientists. Upper Saddle River, NJ: Pearson Prentice Hall.

Yang, I. (2007). Risk modeling of dependence among project task durations. Computer-Aided Civil and Infrastructure Engineering, 22(6), 419–429.

Yang, I. -T. (2005). Simulation-based estimation for correlated cost elements. International Journal of Project Management, 23(4), 275–282.

Yule, G. , & Kendal, M. . (1950). An Introduction to the Theory of Statistics (14th ed.). Charles Griffin & Co.

# 9. Appendix

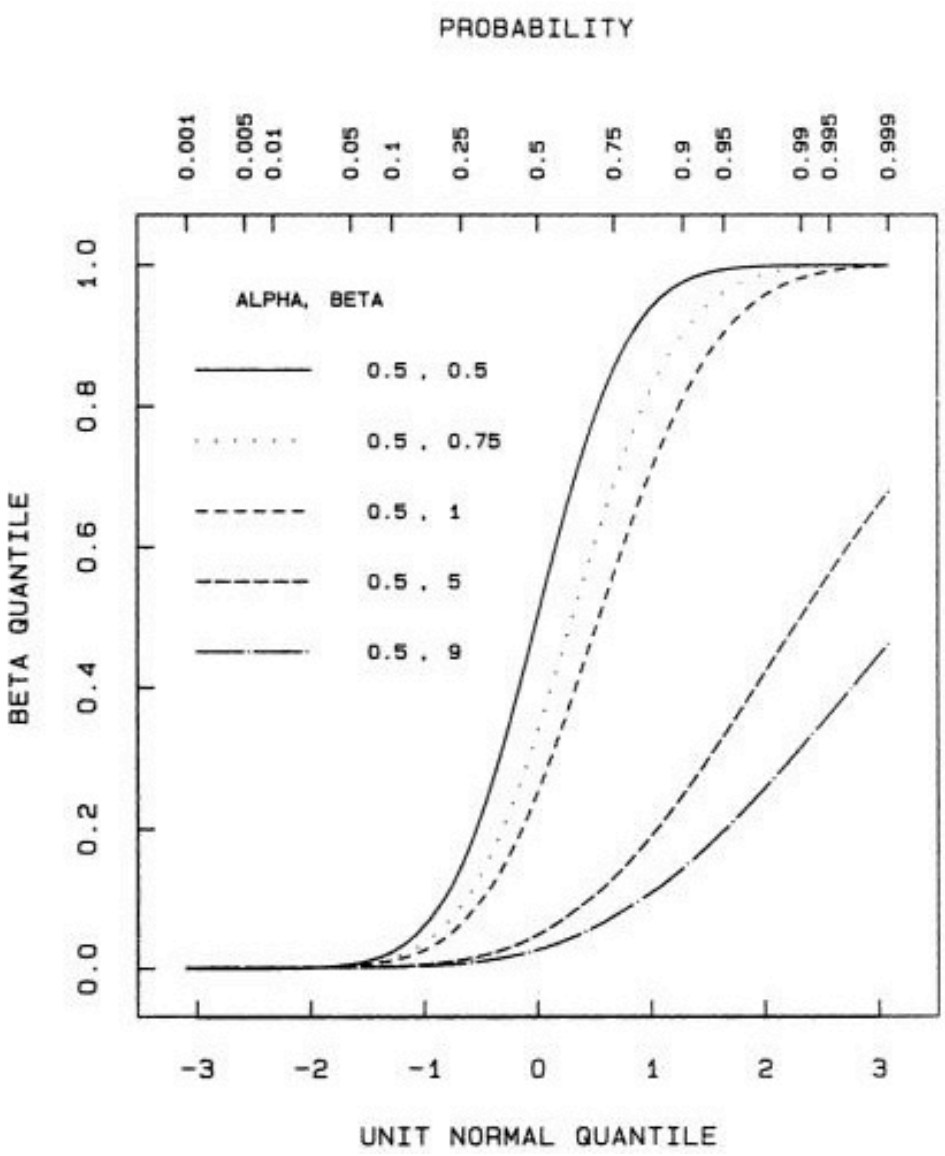


Figure 1-Theoretical plot of normal quantile versus Beta quantile ($\alpha = 0.5$)

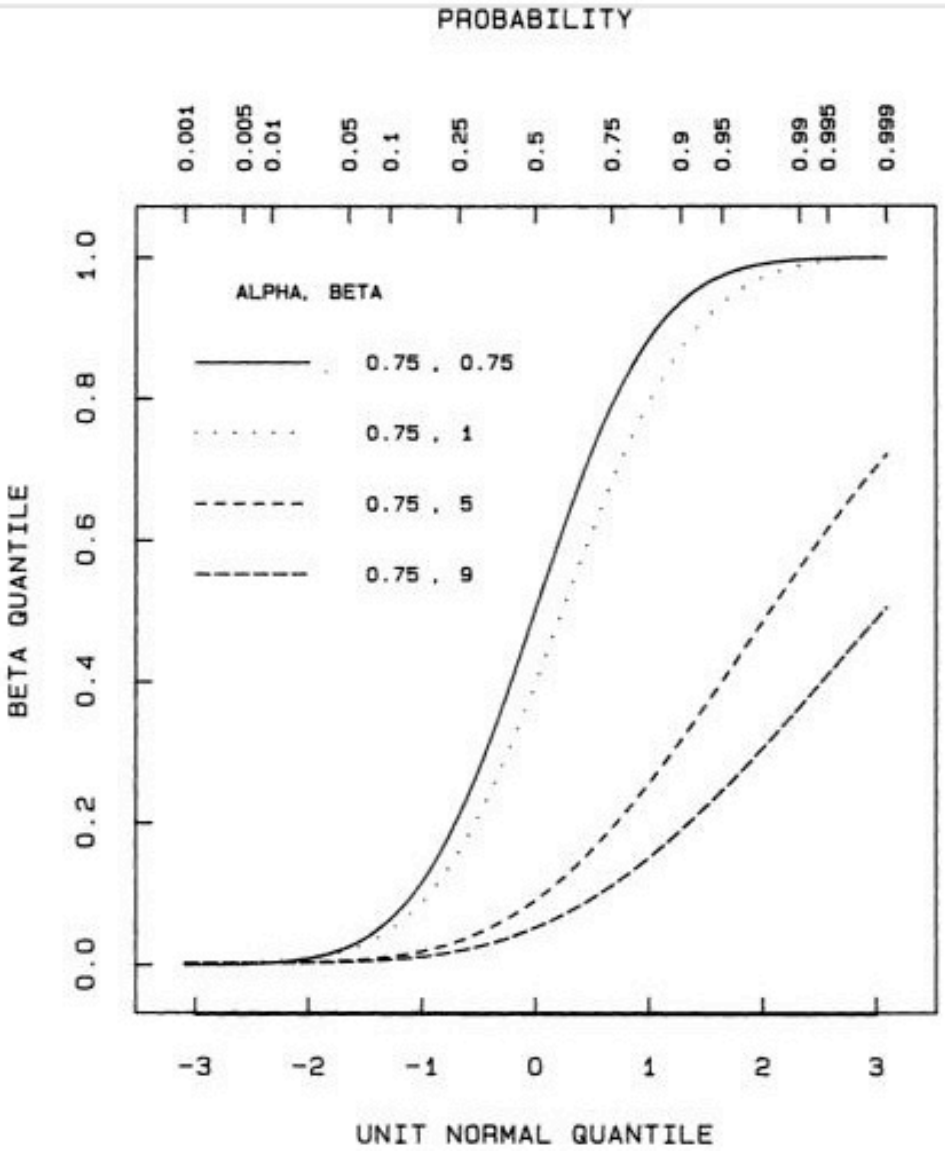


Figure 2- Theoretical Normal vs. Beta Quantile-Quantile ($\alpha = 0.75$)

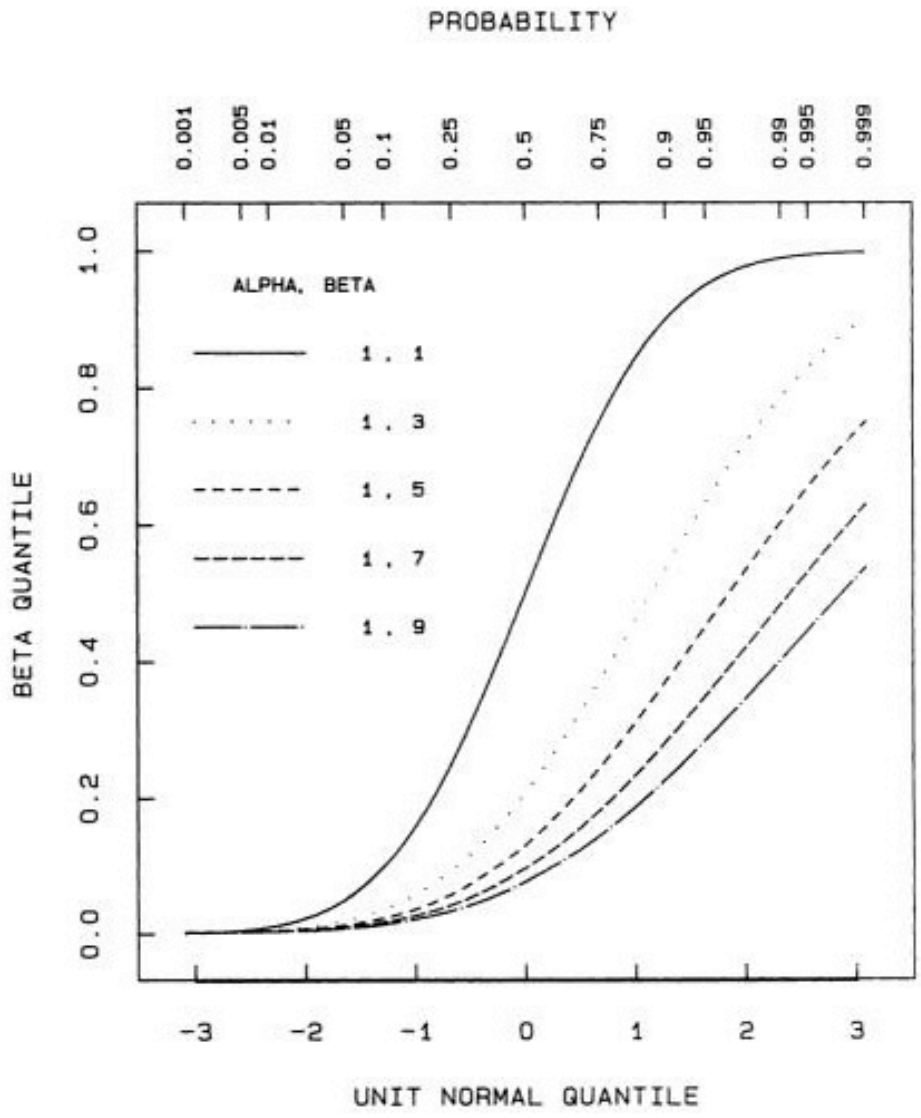


Figure 3- Theoretical Normal vs. Beta Quantile-Quantile ($\alpha = 1$)

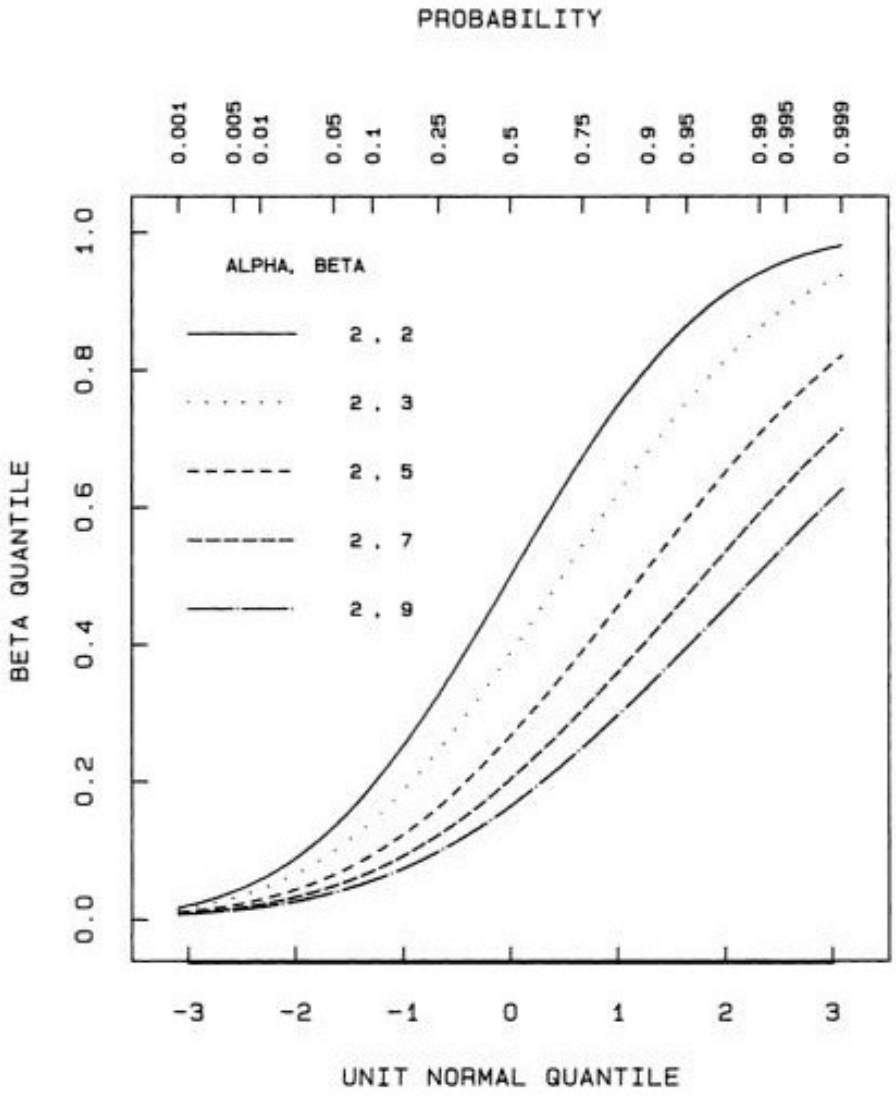


Figure 4- Theoretical Normal vs. Beta Quantile-Quantile ($\alpha = 2$)

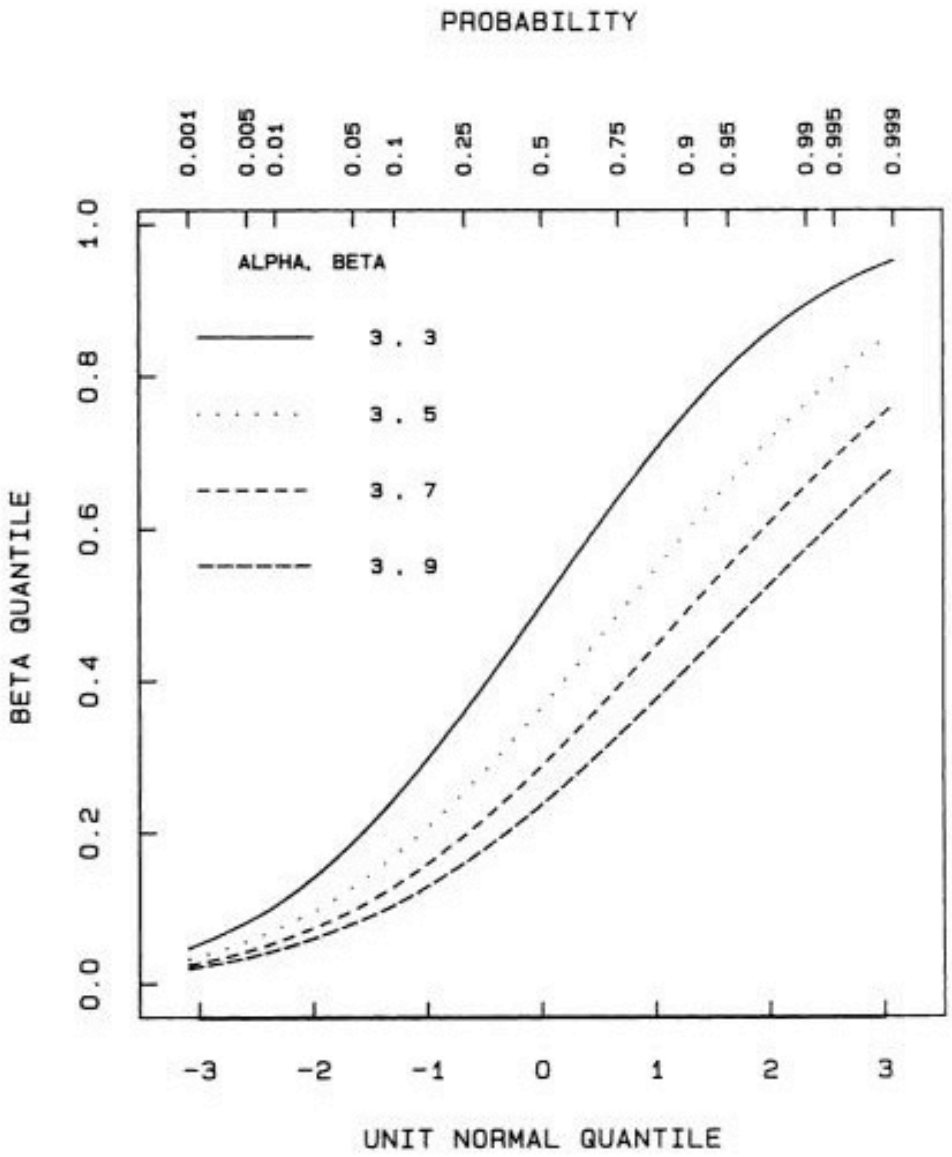


Figure 5- Theoretical Normal vs. Beta Quantile-Quantile ($\alpha = 3$)