Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

A TUTORIAL ON CONCEPTS AND MEASURES OF MANUFACTURING PROCESSES DEPENDENCE

Roberto F. Lu
Shuguang Song

The Boeing Company
P. O. Box 3707
Seattle, WA 98124-2207, USA

ABSTRACT

In a large-scale production system such as building a Boeing 777 airplane, there are more than ten thousand jobs. The interdependence structure among jobs and some jobs' delay's influence on the other jobs in different manufacturing areas are critical information to understand the status of their area. Relationship among groups of jobs may help us find the root cause of problems, to estimate resource planning and to prioritize their tasks. This tutorial is intended to review and introduce commonly used concepts and measures of stochastic dependence. In particular, we focus on concept of positive quadrant dependence, global and local measures of bivariate dependence. Some categorical data analysis techniques and probabilistic models will also be presented. Monte-Carlo simulation techniques will be introduced to provide confidence intervals for estimated measures. Finally, we apply the statistical methods in this tutorial to Boeing 777 production processes as a case study.

1 INTRODUCTION

The concept of stochastic dependence for two random variables X and Y exists widely in many different fields. In reliability analysis, the failure of a component in a system could affects the other components. A manufacturing process in a large-scale production system may affect the status of other manufacturing processes. There are many notions of bivariate and multivariate dependence. In this tutorial, we focus on the concept of positive dependence, in particular, positive quadrant dependence (PQD). PQD basically says X and Y are more likely to be large or small together compared with the independent case. A measure of dependence indicates in some mathematical way the degree or strength of dependence between the variables X and Y. Three global measures of dependence (Pearson's rho, Kendall's tau and Spearman's rho), several local measures of dependence and association measures for categorical data is reviewed in this tutorial. It will be shown that there is a strong relationship between positive dependence and Kendall's tau and Spearman's rho.

The Boeing 777s are the most complicated and customized airplanes in their airplane category. Every one of the Boeing 777 is make-to-order for an airline. The total number of jobs and contents of some jobs are not always the same from one manufacturing area to another in the 777 final assembly. Relationships and job-to-job correlations are not always identical. Since there are more than ten thousand installation jobs for a 777 airplane, there are hundreds of groups of jobs across the 777 final assembly processes. Compositions among group of jobs have established for a very robust 777 final assembly production system that has been producing one Boeing 777 every three working days during the peak production rate. Detailed correlation and their respective confidence levels among group of jobs can be another area that may further improve the effectiveness of the 777 production system. As for any large-scale production system, it needs an automated tool for users (team leaders/managers)

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

to understand the current production status, find the root cause of problems and estimate the effect of current production status on the overall scheduled plan. Thus, understanding the dependence structures and forecasting status of next production process are essential to estimate resource planning and prioritize tasks in manufacturing Boeing 777 airplanes. In this tutorial, we will apply concepts and measures of bivariate dependence to Boeing 777 manufacturing system. Some probabilistic models will also be presented to forecast the future manufacturing process based on the given manufacturing status. Monte-Carlo simulation approaches will be presented to provide corresponding confidence intervals.

In Section 2, we present the concepts of positive dependence, global measures of dependence, local measures of dependence, and association measures for categorical data. Application to Boeing 777 manufacturing system is descried in Section 3. Finally, summary and some discussion are provided in Section 4.

2 CONCEPTS AND MEASURES OF BIVARIATE DEPENDENCE

Concepts of stochastic dependence for a bivariate distribution play an important role in statistics. There are numerous examples of dependence in medical study, economic study, reliability engineering. A complete review on this topic can be found in Joe (1997), Lai and Xie (2006) and Balakrishnan and Lai (2009). In this tutorial, we mainly review positive quadrant dependent (PQD) concept, global measures of dependence, local measures of dependence and some association measures for categorical data.

2.1 Concept of Positive Quadrant Dependence

Two random variables X and Y are positively quadrant dependent (PQD) if $P(X > x, Y > y) \geq P(X > x)P(Y > y)$ for all x and y, or equivalently, $P(X \leq x, Y \leq y) \geq P(X \leq x)P(Y \leq y)$ for all x and y. We say X and Y are negatively quadrant dependent (NQD) if the inequalities are reversed. X and Y are PQD if the probability that they are simultaneously small or large is at least as great as it would be if they were independent. The concept of PQD is widely use statistics like reliability applications (Barlow and Proschan 1981), partial sums (Robbins 1954), order statistics (Robbins 1967), analysis of variance (Kimball 1951) and contingency table (Douglas, Fienberg, Lee, Sampson, and Whitaker 1990). There are families of bivariate distributions that are PQD. Lai and Xie (2006) provides list of well known PQD bivariate distributions that can be used to model stochastic dependence.

PQD is a weaker notion of positive dependence. Positive dependence means that large values of Y accompany large values of X, and small values of Y accompany small values of X. A stronger notion of positive dependence is called totally positive of order 2 (TP$_2$). That is, for all $x_1 < x_2$ and $y_1 < y_2$, the joint density function $f(x,y)$ of (X,Y) satisfies the inequality $f(x_1,y_1)f(x_2,y_2) \geq f(x_1,y_2)f(x_2,y_1)$. It can shown that if f is TP$_2$, then bivariate distribution function $F(x,y)$ and bivariate survival function $S(x,y)$ are also TP$_2$, i.e. $F(x_1,y_1)F(x_2,y_2) \geq F(x_1,y_2)F(x_2,y_1)$ and $S(x_1,y_1)S(x_2,y_2) \geq S(x_1,y_2)S(x_2,y_1)$ for $x_1 < x_2$, $y_1 < y_2$. If F is TP$_2$, then F is PQD.

2.2 Global Measures of Dependence

We often need to quantitively measure the strength or degree of dependence between two random variables X and Y. Such measure can be expressed as a scalar, which is called *global measure* in Drouet-Mari and Kotz (2001). Rényi (1959) proposed a set of seven conditions for global measures of dependence. Lancaster (1982b) modified and enlarged Réyni set of axioms to nine conditions. The main point of those axioms is to make us think about the meaning and measure of stochastic dependence. There are three prominent global measures of dependence: Pearson's product-moment correlation coefficient, Kendall's tau and Spearman's rho.

Pearson's product-moment correlation coefficient is a measure of the strength of the linear relationship between two random variables. It is defined as

$$\rho(X,Y) = \frac{E[(X-EX)(Y-EY)]}{\sqrt{E(X-EX)^2} \cdot \sqrt{E(Y-EY)^2}}.$$

It is clear that $-1 \leq \rho(X,Y) \leq 1$. $|\rho(X,Y)| = 1$ if X and Y are linearly dependent; If X and Y are independent, then $\rho(X,Y) = 0$. However, zero correlation does not imply independence. $\rho(X,Y)$ is symmetric, $\rho(X,Y) = \rho(Y,X)$ and invariant under linear transformations, i.e. $\rho(X, f(Y)) = \rho(X,Y)$ if f is a linear function of Y. However, if f is a nonlinear function, $\rho(f(X), f(Y))$ is generally different from $\rho(X,Y)$. Suppose Y and X has a strong nonlinear relationship $Y = X^2$ and X follows a gamma distribution with parameters (δ, θ), i.e. $f(x) = \theta^{\delta} x^{\delta-1} e^{-\theta x} / \Gamma(\delta)$. It can be shown that ρ is independent of θ and is an increasing function of δ. ρ varies from $\sqrt{2/3}$ when $\delta = 0$ to 1 when $\delta = \infty$. Thus, the Pearson's correlation coefficient can be lower than 1 if the dependence is nonlinear. The correlation coefficient measures only linear association. It is not a good summary of association if the scatter plot has a nonlinear pattern.

The usual formula to estimate the Pearson's correlation coefficient in a sample of n bivariate observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is

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

where $\bar{x}$ and $\bar{y}$ are sample means. A disadvantage of r is that it is very sensitive to outliers in the sample. The sample distribution of r has been thoroughly reviewed in (Johnson, Kotz, and Balakrishnan 1995), Chapter 32. The message about the robustness of r are conflicting. Researchers should be careful of the underlying assumptions of the population before reporting the value of r. It should be kept in mind that different data sets could give the same value of r and the value of r calculated from a small sample may be totally misleading and should be viewed in the context of its likely sampling error. For highly skewed bivariate distribution function, the Person's correlation coefficient is not a very useful measure of association, see Barnett (1985).

Kendall's tau (τ) and Spearman's rho (ρ_S), see Kendall (1938) and Spearman (1904), are the well-known rank correlation coefficients. They are the measures of correlation between rankings, rather than between actual values of X and Y. Thus, they are invariant by any increasing transformation of X and Y; while Pearson's moment-product correlation coefficient (ρ) is invariant only under linear transformations. For a set of bivariate parallel data (x_i, y_i) that are assumed to independently and identically distributed, where $i = 1, \ldots, n$, the Kendall's tau is defined as

$$\tau \equiv E\text{sign}\{(X_1 - X_2)(Y_1 - Y_2)\},$$

where sign(x) is -1 for $x < 0$, 0 for $X = 0$, 1 for $x > 0$. For continuous probability distribution, let p be the probability that the order of the coordinate 1 observations is the same as the order of the coordinate 2 observations

$$p = P[(X_1 - X_2)(Y_1 - Y_2) > 0].$$

Then it can be shown that $\tau = 2p - 1$. It follows that $-1 \leq \tau \leq 1$, and $\tau = 1$ if $p = 1/2$. A conceptual drawback of Kendall's tau is that the interpretation of τ needs two pairs. Consider the times to finish two jobs when building a Boeing 777 airplane. If job A finish time for airplane 1 is longer than job A finish time for airplane 2, Kendall's tau measures if job B finish time for airplane 1 is also longer than the job B finish time for airplane 2. Nelsen (1992) proved that $\tau/2$ is an average measure of total positivity defined as

$$T=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[f(x_2,y_2)f(x_1,y_1)-f(x_2,y_1)f(x_1,y_2)]dx_1dx_2dy_1dy_2$$

for all $x_1 < x_2$ and $y_1 < y_2$, where density function $f(x,y)$ satisfies $f(x_2,y_2)f(x_1,y_1)-f(x_2,y_1)f(x_1,y_2) \geq 0$ for all $x_1 < x_2$ and $y_1 < y_2$. Thus τ is a global measure of a strong dependence concept.

Kendall's tau (τ) can be estimated from bivariate parallel data. Let $a_{im} = 1$ if $X_i > X_m$, 0 if $X_i = X_m$, -1 if $X_i < X_m$ and let $b_{im} = 1$ if $Y_i > Y_m$, 0 if $Y_i = Y_m$, -1 if $Y_i < Y_m$. In the case of complete data without ties:

$$\hat{\tau}=\frac{\sum_{i,m}a_{im}b_{im}}{n(n-1)}.$$

In the case of ties or censored data, the formula is generalized to

$$\hat{\tau}=\frac{\sum_{i,m}a_{im}b_{im}}{[\sum_{i,m}a_{im}^2\cdot\sum_{i,m}b_{im}^2]^{1/2}}.$$

The score for a_{im} can be obtained by

(D_i,D_m)	$X_i > X_m$	$X_i = X_m$	$X_i < X_m$
$(1,1)$	1	0	-1
$(0,1)$	1	1	$a[\hat{S}(X_m)/\hat{S}(X_i)]-1$
$(1,0)$	$1-2\hat{S}(X_i)/\hat{S}(X_m)$	-1	-1
$(0,0)$	$1-\hat{S}(X_i)/\hat{S}(X_m)$	0	$\hat{S}(X_m)/\hat{S}(X_i)-1$

where $\hat{S}(t)$ denotes the Kaplan-Meier estimate of the survival function. The value for b_{im} is similarly obtained. The variability of this estimate was also studied. See Brown, Hollander, and Korwar (1974), Meier and Basu (1980), Oakes (1982) for more details.

Spearman's rho (ρ_S) is a population version of the measure of association. It is a non-parametric measure, independent of marginal transformations. For an arbitrary continuous marginal distributions, it is defined by

$$\rho_S = 12\int_0^1\int_0^1 S(S_1^{-1}(u),S_2^{-1}(v))dudv-3,$$

where $S(u,v)$ is the joint survival function of two random variables. This expression is not simple to integrate and can be handled by numerical integration. However, in survival data, there could be a point mass at ∞. Then Spearman's rho can not be evaluated.

Let $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$ be three independent pairs of random variables with a common distribution function F. ρ_S can also be defined to be proportional to the probability of concordance minus the probability of discordance for the two pairs (X_1,Y_1) and (X_2,Y_3),

$$\rho_S = 3(P[(X_1-X_2)(Y_1-Y_3)>0]-P[(X_1-X_2)(Y_1-Y_3)<0]).$$

It is well know that

$$\rho_S = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F(x,y)-F_X(x)F_Y(y)]dF_x(x)dF_Y(y),$$

see Schweizer and Wolff (1981). Thus, $\rho_S/12$ represents an average measure of quadrant dependence with the average being taken with respect to the marginal distributions of X and Y. Note that we say a pair (X,Y) is positively quadrant dependent (PQD) if $F(x,y)-F_X(x)F_Y(y) \geq 0$ for all x and y, and negatively quadrant dependent (NQD) if the inequality is reversed. It follows that $\rho_S \geq 0$ if X and Y are PQD, and $\rho_S \leq 0$ if X and Y are NQD.

The standard estimate of ρ_S with complete data is based on the marginal ranks (R_{i1}, R_{i2}),

$$\hat{\rho}_S = \frac{\sum_i [R_{i1} - (n+1)/2][R_{i2} - (n+1)/2]}{n(n^2-1)/12}.$$

R_{i1} is the rank of X_i among $X_1, \ldots, X_n$ and R_{i2} is the rank of Y_i among $Y_1, \ldots, Y_n$.

Both Kendall's tau and Spearman's rho are measures of rank correlations. But the values of ρ_S and τ are often quite different. Some explicit relationships between ρ_S and τ have been derived for bivariate normal distribution ($\rho_S = \frac{6}{\pi}\sin^{-1}(\frac{1}{2}\sin\frac{\pi\tau}{2})$), Farlie-Gumbel-Morgenstern bivariate distribution($\rho_S = 3\tau/2$) and Marshall and Olkin's bivariate exponential distribution ($\rho_S = 3\tau/(2+\tau)$). A precise relation between ρ_S and τ does not exit for every bivariate distribution. But we have the following inequalities for general relationships between ρ_S and τ: $-1 \leq 3\tau - 2\rho_S \leq 1$ and $(1 \pm \rho_S)/2 \geq (1 \pm \tau)^2/4$, see Kruskal (1958). When we apply Kendall's tau and Spearman's rho, we should keep in mind that independence of X and Y implies $\tau = \rho_S = 0$, but it does not hold to reverse. τ and ρ_S are less sensitive to outliers compared with sample correlation r. There is very strong relationship between positive dependence and τ, ρ_S. If X and Y are positively quadrant dependent, then $\tau \geq 0$ and $\rho_S \geq 0$.

2.3 Local Measures of Dependence

Pearson's moment-product correlation coefficient, Kendall's tau and Spearman's rho are global measures of dependence. They do not measure the dependence locally. They can be zero when X and Y are not independent. A distribution with high ρ_S may not be a PQD, see Drouet-Mari and Kotz (2001). Thus, the global measures of dependence have some drawbacks. To address the early-late dependence, short-term and long-term dependence, and the time of maximal correlation between two survival variables, we need to define a local measure of dependence.

Let $V(x_0, y_0)$ be an open neighborhood of (x_0, y_0). A distribution $F(x,y)$ is PQD in the neighborhood $V(x_0, y_0)$ if $S(x,y) \geq S_X(x)S_Y(y)$ for all $(x,y) \in V(x_0, y_0)$. In the following, we list local dependence measures.

- The **local correlation coefficient** is defined as, see Bjerve and Doksum (1993),

  $$\rho(x) = \frac{\sigma_X \beta(x)}{(\sigma_X \beta(x))^2 + \sigma(x)^2},$$

  where $\mu(x) = E[Y|X=x], \sigma^2(x) = \mathrm{var}(Y|X=x)$ and $\beta(x) = \frac{\partial \mu(x)}{\partial x}$.
- In an open neighborhood $V(x_0, y_0)$ of (x_0, y_0), **local** ρ_S **and** τ are defined(Drouet-Mari and Kotz 2001),

  $$\rho_{S,(x_0,y_0)} = \frac{12\int\int_{V(x_0,y_0)}(C(u,v) - uv)dudv}{\int\int_{V(x_0,y_0)} dudv},$$

  and

  $$\tau_{(x_0,y_0)} = \frac{4\int\int_{V(x_0,y_0)} C(u,v)dC - 1}{\int\int_{V(x_0,y_0)} dC},$$

  Where C is the copula corresponding to the bivariate distribution function of (X,Y). $\rho_{S,(x_0,y_0)}/12$ can be interpreted as the average of local PQD, while $\tau_{(x_0,y_0)}/2$ can be interpreted as the average of local TP_2.

- Clayton (1978) and Oakes (1989) defined a local association measure as

$$\rho(x,y) = \frac{S(x,y)\frac{\partial^2 S(x,y)}{\partial x \partial y}}{\frac{\partial S(x,y)}{\partial x}\frac{\partial S(x,y)}{\partial y}}.$$

  Here, the local dependence is measured at a single point (x,y). It is shown in Gupta (2003) that $\rho(x,y) > 1$ if and only if $P(X > x, Y > y | X > x', Y > y')$ is increasing in (x', y') for all (x,y). $\rho(x,y) = 1$ if and only if X and Y are independent.
- **A local dependence measure by Holland and Wang:** consider a $r \times s$ contingency table with cell proportions p_{ij}. the cross-product ratios: $\alpha_{ij} = (p_{ij}p_{i+1,j+1})/(p_{i,j+1}p_{i+1,j})$ for $1 \leq i \leq (r-1), 1 \leq j \leq (s-1)$, or $\gamma_{ij} = \log \alpha_{ij}$ measures the association in the 2×2 subtables with pairs of adjacent rows and columns. Motivated by this, Holland and Wang (1987a) and Holland and Wang (1987b) defined a local dependence measure

$$\gamma(x,y) = \lim_{dx,dy \to 0} \frac{1}{dxdy} \log \left(\frac{f(x,y)f(x+dx,y+dy)}{f(x+dx,y)f(x,y+dy)} \right) = \frac{\partial^2}{\partial x \partial y} \log f(x,y),$$

  where $f(x,y)$ is the bivariate density function and its second order partial derivative exists. $\gamma(x,y)$ is shown to be an appropriate local measure of TP_2 dependence. Also, $-\infty < \gamma(x,y) < \infty$ and $\gamma(x,y) = 0$ if and only if X and Y are independent. Note that the three global measures may be zero without X and Y being independent.
- Let $\mu(x) = E(Y|X = x)$ and $\mu(y) = E(X|Y = y)$. Bairamov, Kotz, and Kozubowski (2003) defined a **local linear dependence function** $H(x,y)$ as

$$H(x,y) = \frac{E(X - \mu(y))E(Y - \mu(x))}{\sqrt{E(X - \mu(y))^2 E(Y - \mu(x))^2}}.$$

  Clearly, $H(x,y)$ is obtained from Pearson's correlation coefficient by replacing $E(X)$ and $E(Y)$ by conditional expectations $\mu(x)$ and $\mu(y)$, respectively. The concept of local dependence and measures of local dependence still remain to be fully developed. The current local dependence measures provide us more detailed information about dependence. For application of local dependence measures in survival analysis, see Drouet-Mari and Kotz (2001).

2.4 Association Measures for Categorical Observations

Consider Job A and Job B when building a Boeing 777 airplane. From historical data, we observe the number of delay and on-schedule status of Job A and Job B for n 777 airplanes. Let n_1 be the number of airplanes with delayed Job A and delayed Job B, n_2 be the number of airplanes with delayed Job A and on-schedule Job B; n_3 be the number of airplanes with on-schedule Job A and delayed Job B; n_4 be the number of airplanes with on-schedule Job A and on-schedule Job B. We would like to examine if there is evidence of association between Job A and Job B (e.g. if Job A is delayed, how likely the Job B will also be delayed?) and if so, how strong is it? In this section, we will review three ways to measure the strength of association for categorical data: comparing proportions, odds and odds ratios, concordant and discordant pairs. The two main references for this section are Agresti and Finlay (1997) and Agresti (2002).

- **Comparing proportions:** We treat the samples as independent binomials. Let π_1 be probability that Job B is delayed in the group that Job A is delayed. Let π_2 be probability that Job B is delayed in the group that Job A is on schedule. The estimates of π_1 and π_2 are $\hat{\pi}_1 = n_1/(n_1 + n_2)$ and $\hat{\pi}_2 = n_3/(n_3 + n_4)$, respectively. The difference of sample proportions $\hat{\pi}_1 - \hat{\pi}_2$ must range between -1 and 1. A difference close to one in magnitude indicates a high level of association between Job A and Job B, while a difference close to zero represents little association. Then the 95% Wald confidence interval for $\pi_1 - \pi_2$ is:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{0.025}\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2),$$

where $\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2) = [(\hat{\pi}_1(1-\hat{\pi}_1)/(n_1+n_2) + \hat{\pi}_2(1-\hat{\pi}_2)/(n_3+n_4)]^{1/2}$. If the 95% confidence interval is on the right side of 0 and excludes 0, then there is evidence that if Job A is delayed then Job B will also be delayed. The further the low end of 95% away from 0, the stronger the evidence. If the confidence interval is on the left side of 0 and excludes 0 or include 0, then delay status of Job B is not strongly associated with delay status of Job A.

- **Odds and odds ratios:** For a probability π success, the *odds* are defined to be $\theta = \pi/(1-\pi)$. The odds are nonnegative and $\theta > 1$ when a success is more likely to occur than a failure. Consider a 2×2 table with joint probability $\{\pi_{ij}\}$ where $i = 1,2$, $j = 1,2$. That is, the probability for (X,Y) to be in row i and column j is π_{ij}. The odds of success in each row are $\theta_i = \pi_{i1}/\pi_{i2}$, where $i = 1,2$. The ratio of the odds θ_1 and θ_2

$$\Theta = \frac{\theta_1}{\theta_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

is called the *odds ratio*. The odds ratio is nonnegative. If $\Theta = 1$, i.e. $\theta_1 = \theta_2$, corresponds to independence of X and Y. If $1 < \Theta < \infty$, the subjects in row 1 are more likely to have success than subjects in row 2, i.e. $\pi_1 > \pi_2$. If $0 < \Theta < 1$, then $\pi_1 < \pi_2$. By definition of odds ratio, we have that the odds ratio is invariant to orientation of the table. For observations with cell counts $\{n_{ij}\}$, the sample odds ratio is

$$\hat{\Theta} = (n_{11}n_{22})/(n_{12}n_{21}).$$

The Wald $(1-\alpha) \times 100\%$ confidence interval for $\log\Theta$ is

$$\log\hat{\Theta} \pm z_{\alpha/2}\hat{\sigma}(\log\hat{\Theta}),$$

where $\hat{\sigma}(\log\hat{\Theta}) = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$. For small sample size, a confidence interval for Θ is obtained from inverting the test $H_0 : \Theta = \Theta_0$ conditional observing $n_{11} = k_o$. For $H_a : \Theta > \Theta_0$, the p-value is $P = \sum_{k \geq k_o} f(k; n_{1+}, n_{+1}, n, \theta_o)$. For $H_a : \Theta < \Theta_0$, the p-value is $P = \sum_{k \leq k_o} f(k; n_{1+}, n_{+1}, n, \theta_o)$. Here, $n_{1+} = n_{11} + n_{12}$, $n_{+1} = n_{11} + n_{21}$, $n = n_{11} + n_{12} + n_{21} + n_{22}$, and

$$f(k; n_{1+}, n_{+1}, n, \Theta_o) = \frac{\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}}{\sum_{l=\max(0,n_{1+}+n_{+1}-n)}^{\min(n_{1+},n_{+1})}\binom{n_{1+}}{l}\binom{n-n_{1+}}{n_{+1}-l}\Theta_0^l}.$$

- **Concordant and discordant pairs:** This method is useful only when the categories can be ordered. A pair of observations is *concordant* if the subject who is higher on one variable is also higher on the other variable. A pair of observations is *discordant* if the subject who is higher on one variable is lower on the other. If a pair of observations is in the same category of a variable, then it is neither concordant or discordant and is said to be *tied* on that variable. Consider the following two-way table that categorizes a sample of people in the work force by income level (high or low) and educational level (end after high school or end after college). In Table 1, d people have high education and high income, a people have low education and low income. Thus, there are $C = ad$ concordant pairs. Also, b people have low education high income, c people have high education and low income. Thus, there are $D = bc$ discordant pairs.

  The strength of association between education levels and income levels can be measured by calculating the difference of proportions of concordant ($C/(C+D)$ in this example) and proportion of discordant pairs ($D/(C+D)$ in this example). We will give definition of such

		Income	**Level**
		Low	High
Education	High School	*a*	*b*
Level	College	*c*	*d*

Table 1: An illustrated example of people by income level and educational level.

association measure for $I \times J$ table. Consider two independent observations from a joint distribution $\{\pi_{ij}\}$. For that pair, the probabilities of concordance and discordance are

$$\Pi_c = 2\sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k>j} \pi_{hk} \right), \qquad \Pi_d = 2\sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k<j} \pi_{hk} \right).$$

Given a pair is untied, $\Pi_c/(\Pi_c + \Pi_d)$ is the probability of concordance and $\Pi_d/(\Pi_c + \Pi_d)$ is the probability of discordance. The difference of these two probabilities is called *gamma*,

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

It follows that $-1 \leq \gamma \leq 1$. A positive value of γ indicates a positive association, while a negative value of γ indicates a negative association. If γ is close to 0, then it indicates a very weak association. If γ is close to -1 or 1, then it indicates a very strong association. By definition of γ, we can see that it is not very sensitive to sample size. Note that $\gamma = 1$ if $\Pi_d = 1$ and $\gamma = -1$ if $\Pi_c = 0$. Independence implies $\gamma = 0$, but the converse is not true. In Table 1, The estimate of *gamma* is

$$\hat{\gamma} = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}.$$

In order to have $\hat{\gamma} = 1$, we must have $D = 0$, i.e. $b = 0$ or $c = 0$. Thus, whenever we have observations with b or c close to 0 such that D is very small, the $\hat{\gamma}$ value will be equal to 0.

3 APPLICATION TO BOEING 777 PRODUCTION SYSTEM

Boeing 777 manufacturing system consists of a huge number of jobs. The whole manufacturing process is divided into a number of major assembly areas (MAA). The jobs in each major assembly area are grouped into so-called “milestones” at Boeing. As for any large-scale production system, it needs an automated tool for team leaders and managers to understand the current production status, find the root cause of problems and estimate the effect of current production status on the overall scheduled plan. Thus, understanding the dependence structures and forecasting status of next production process within each major assembly area and across different major assembly areas at both job level and milestone level are essential to estimate resource planning and prioritize tasks in manufacturing Boeing 777 airplanes. In this section, we will apply concepts and measures of bivariate dependence to Boeing 777 manufacturing system. Some probabilistic models will also be presented to forecast the future manufacturing process based on the given manufacturing status. Monte-Carlo simulation approaches will be used to provide corresponding confidence intervals.

3.1 Association of Milestones and Jobs in Boeing 777 Production System

Consider two milestones (denoted by MS_1 and MS_2) that are either within the same Boeing 777 major assembly area or different Boeing 777 major assembly areas. Define the finish time of a milestone as the last finish time of the job in the milestone list. The database recorded the beginning time of each major assembly area for manufactured Boeing 777 airplanes, so the difference in minutes between the milestone finish time and its corresponding starting time is regarded as the time needed to finish the milestone. We say a milestone was delayed if that milestone's finish time was behind its scheduled finish time, on schedule otherwise. After removing 4 outliers due to engineering reasons, the data for MS_1 and MS_2 are given in Table 2. Let π_1 be the true probability that MS_2 is delayed if MS_1 is on schedule; π_2 be the true probability that MS_2 is delayed if MS_1 is delayed. Using the association measure by comparing proportions for categorical data, the 95% confidence interval for $\pi_1 - \pi_2$ is (-0.09, -0.07). Although the 95% confidence interval is to the left side of 0. But it is very close to 0. Thus, there is no strong statistical evidence to support the association between these two milestones.

		MS_2	
		on schedule	delayed
MS_1	on schedule	96	0
	delay	12	1

Table 2: 2×2 contingency table for MS_1 and MS_2.

On the other hand, we use the three global measures of dependence: Pearson's moment-product correlation coefficient, Kendall's tau and Spearman's rho to study the dependence between MS_1 and MS_2. The function "cor.test" in R provides tests for association between paired samples using these three global measures. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS, see <http://www.r-project.org/>. From Table 2, we know that there are 109 samples for MS_1 and MS_2. This sample size is not large enough to invoke the large-sample theory (e.g. the Central Limit Theorem). Also, Plots of 109 times needed to finish MS_1 and MS_2 does not suggest any obvious parameter distribution fitting. Thus, we will apply Bootstrap confidence intervals for estimated global measures of dependence between MS_1 and MS_2. In statistics, bootstrapping is a modern, computation-intensive approach to statistical inference based on resampling methods, see Efron and Tibshirani (1993). A $(1 - alpha) \times 100\%$ bootstrap confidence interval for an estimate denoted by θ (in this case, θ is Pearson's ρ, or Kendall's τ, or Spearman's ρ_S) is obtained by the following procedures:

- Obtain a single sample of size n from the population under study and calculate $\hat{\theta}$. $\hat{\theta}$ is an estimate of θ based on the sample.
- Generate a bootstrap sample of the same size n by resampling with replacement from original sample.
- Calculate $\hat{\theta}^*$ using the generated bootstrap resample.
- Repeat above two steps for a large number N to obtain and order $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_N^*$ from the smallest to the largest.
- The $(1-\alpha) \times 100\%$ bootstrap confidence interval for θ is obtained by taking the $(\alpha/2) \times 100\%$ and $(1-\alpha/2) \times 100\%$ percentiles of the ordered $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_N^*$ as endpoints.

	Pearson's ρ	Kendall's τ	Spearman's ρ_S
Estimates	0.03	0.06	0.08
95% confidence interval	(-0.17, 0.20)	(-0.07, 0.19)	(-0.12, 0.27)

Table 3: Global measures of dependence between MS_1 and MS_2.

Take $\alpha = 0.05$ to obtain 95% confidence interval. Table 3 lists the calculated Pearson's ρ, Kendall's τ and Spearman's ρ_S based on original 109 observed times needed for MS_1 and MS_2, and corresponding 95% bootstrap confidence intervals. The three 95% bootstrap confidence intervals all contain zero. Thus, there is no strong statistical evidence to support the association between MS_1 nd MS_2. This is consistent with the results obtained by comparing proportions. There is no apparent pattern for 109 observed time to finish MS_1 and MS_2, see Figure 1. This further confirms our analysis using measures of dependence.

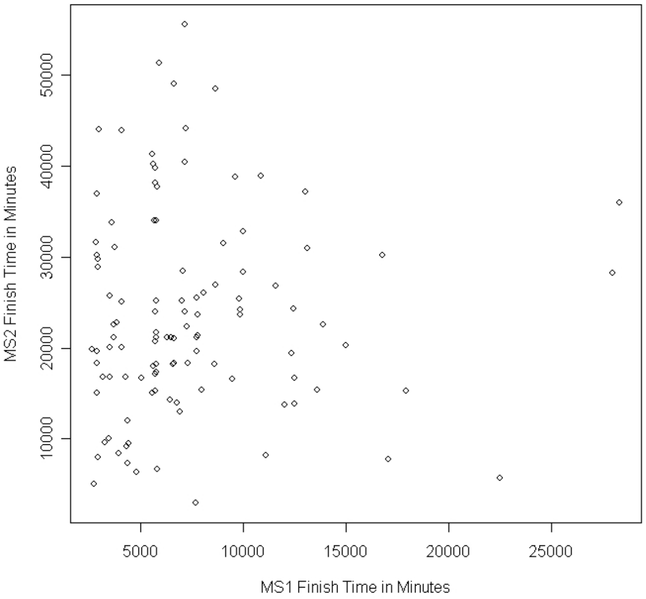

Figure 1: Scatter plot of time needed to finish MS_1 and MS_2.

We apply above study to all milestones and jobs in the Boeing 777 production system to reveal their dependence structure. This will pave the step to probabilistically forecast future status of milestones and jobs for those with interdependence.

3.2 Probabilistic Forecast Models for Boeing 777 Production System

Consider two dependent milestones MS_1 and MS_2 with observed time to finish MS_1 and MS_2: $(T^i_{MS_1}, T^i_{MS_2})$ for $i = 1, 2, \ldots, n$. We can estimate the probability $P(T_{MS_1} < t_1, T_{MS_2} < t_2)$. Then we can address the following questions:

- Estimate $P(T_{MS_2} > T_{MS_1})$ to examine if MS_2 finishes later than MS_1.
- Estimate $P(T_{MS_2} < t_2 | T_{MS_1} < t_1)$. Set t_2 be the scheduled finish time of MS_2. Then we can examine the probability of finishing MS_2 on schedule conditional current status of MS_1.
- Set $P(T_{MS_2} < t_2 | T_{MS_1} < t_1) = 0.95$. Then we can estimate the time needed to finish MS_2 with probability 0.95 conditional on finishing time of MS_1.
- Examine if MS_1 and MS_2 are PQD or NQD by comparing values $P(MS_1 > t_1, MS_2 > t_2)$ and $P)MS_1 > t_1)P(MS_2 > t_2)$. This comparison will shed light on fitting appropriate bivariate distribution functions.

Note that if the two milestone MS_1 and MS_2 are from different major assembly areas MAA_1 and MAA_2 such that MAA_2 can not begin until MAA_1 is finished. It follows that $T_{MS_2} > T_{MS_1}$ by choosing the same starting reference time, say the beginning time of MAA_1. We need to fit bivariate distribution function with this constraint. Since a milestone status could be affected by more than one other milestones - the same for jobs. Assume that milestones MS_j, $j = 1, \ldots, J$, affect a milestone MS independently, then we can estimate the overall effects of milestones MS_j, $j = 1, \ldots, J$, by

$$P(T_{MS} < t | T_{MS_1} < t_1, \ldots, T_{MS_J} < t_J) = \Pi^J_{j=1} P(T_{MS} < t | T_{MS_j} < t_j).$$

If a milestone is shown to be unaffected by any other milestones. We can directly estimate $P(T_{MS} < t)$ from data. Given how long this milestone have been worked, we can estimate $P(T_{MS} < t + s | T_{MS} = s)$ - the probability to finish this milestone with additional time s.

Estimation of above probability distributions can be empirically, parametrically. We can also consider effects of other factors in the modeling if necessary. Figure 2 shows empirically probabilistic forecast of MS_2 conditional on finishing MS_1 in 1000 minutes.

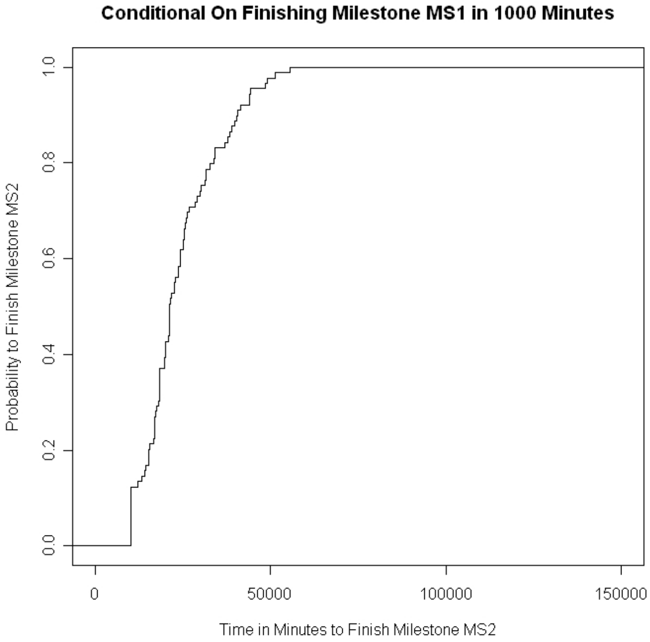

Figure 2: Empirical probabilistic forecast of milestone MS_2 conditional on milestone MS_1.

We can extend the above probabilistic models to forecast jobs. However, there are often many data issues in a large-scale production system. In Boeing 777 production system, there is uncertainty about the starting time for each job, although the records of job finish time are very robust. We know that a job can not start earlier than the starting time of the major assembly area to which the job belongs and should center around the scheduled starting time of the job. This shows that the true starting time of a job follows a triangular distribution. We can then simulate the true starting time for each recoded job N times and obtain the time needed to finish that job by taking the difference between the job's complete time and simulated starting time. Thus, the corresponding probabilistic forecast and its confidence interval can be estimated N times from simulated data.

4 SUMMARY AND DISCUSSION

In this tutorial, we review the concept of positive dependence, global and local measures of dependence and measures of association for categorical data. We then apply these concepts and measures to a large-scale production system - Boeing 777 manufacturing system. Based on the results from dependence study, some probabilistic models and simulation approaches are also provided for this real-world problem. In the following, we present some brief discussion.

- We mainly review some classical concepts and measures for bivariate dependence. There are other dependence concepts like complete dependence, monotone dependence, regional dependence, stochastically increasing. Other measures of dependence are also available. Examples are Gini measure (Nelsen 1999), quadrant test of Blomqvist (1950), measures of dependence by Schweizer and Wolff (1981), and matrix correlation, see Lancaster (1982a) and Lancaster (1982b).

- There is no universal answer to the question of the best measure of dependence. It needs not only mathematical or statistical concepts and measures, but also deep engineering knowledge about the problem to better model the dependence.
- It is often very difficult to describe the dependence between two random variables X and Y. We are essentially to study if there is better design in the case of reliability analysis.
- Independence is still commonly assumed in statistical analysis and correlation is widely used. We should promote the application of other concepts of dependence. For example, PQD or NQD can be relatively easy to verify, as many nonparametric methods have been developed for various bivariate data. The dependence structure will shed light on choosing appropriate bivariate distribution functions with characteristics of such dependence.
- Most concepts and measures of dependence are static in the sense that they are invariant to time and space. However, the degree of dependence could time-indexed (e.g. in survival analysis) or even space-indexed (e.g. in mobile system). Such concepts and measures remain to be fully developed.
- In a large-scale production system, a component could depend on more than one component. Furthermore, the dependence could be dynamic. We need to develop concepts and measures to address this need. We also need to be very careful about the probabilistic modeling of a component's behavior by considering the complicated interdependence. In particular, we need to combine engineering understanding of the system and statistical study of dependence measures to carefully examine various assumptions of independence or conditional independence among components in the system.
- The interdependence structure are becoming more and more complex in modern manufacturing system. There is a need for an automated tool to capture the dependence among components in the system and forecast a component's behavior conditional on the other components' status.

REFERENCES

Agresti, A. 2002. *Categorical data analysis*. 2rd ed. New Jersey: John Wiley & Sons, Inc.

Agresti, A., and B. Finlay. 1997. *Statistical methods for the social sciences*. 3rd ed. Prentice Hall.

Bairamov, I., S. Kotz, and T. J. Kozubowski. 2003. A new measure of linear local dependence. *Statistics* 37:243–258.

Balakrishnan, N., and C.-D. Lai. 2009. *Continuous bivariate distributions*. 2rd ed. New York: Springer.

Barlow, R. E., and F. Proschan. 1981. *Statistical theory of reliability and life testing*. Silver Spring: To Begin With.

Barnett, V. 1985. The bivariate exponential distribution: a review and some new results. *Statistica Neerlandica* 39:343–357.

Bjerve, S., and K. Doksum. 1993. Correlation curves: Measures of association as function of covariates values. *Annals of Statistics* 21:890–902.

Blomqvist, N. 1950. On a measure of dependence between two random variables. *Annals of Mathematical Statistics* 21:593–600.

Brown, W. B., M. Hollander, and R. M. Korwar. 1974. Nonparametric tests of independence for cenosred data with applications to heart transplant studies. In *Reliability and Biometry: Statistical analysis of lifelength*, ed. F. Froschan and R. G. Serfling, 327–354. Philadelphia: SIAM.

Clayton, D. G. 1978. A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65:141–151.

Douglas, R., E. Fienberg, M. L. Lee, A. R. Sampson, and L. R. Whitaker. 1990. Positive dependence concepts for ordinal contingency tables. In *IMS Lecture Notes Monograph Series*, Volume 16, Topics in Statistical Dependence, 189–202. Hayward, California: Institute of Mathematical Statisics.

Drouet-Mari, D., and S. Kotz. 2001. *Correlation and dependence*. London: Imperial College Press.

Efron, B., and R. J. Tibshirani. 1993. *An introduction to the bootstrap*. Boca Raton, Florida: Chapman & Hall/CRC.

Gupta, R. C. 2003. On some association measures in bivariate distributions and their relationships. *Journal of Statistical Planning and Inference* 117:83–98.

Holland, P. W., and Y. J. Wang. 1987a. Dependence function for continuous bivariate densities. *Communication in Statistics – Theory and Methods* 16:863–876.

Holland, P. W., and Y. J. Wang. 1987b. Regional dependence for continuous bivariate densities. *Communication in Statistics – Theory and Methods* 16:193–206.

Joe, H. 1997. *Multivariate models and dependence concepts*. London: Chapman and Hall.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1995. *Continuous univariate distributions*. Vol 2, 2rd ed. New Jersey: John Wiley & Sons, Inc.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30:81–93.

Kimball, A. W. 1951. On dependent tests of significance in analysis of variance. *Annals of Mathematical Statistics* 22:600–602.

Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association* 53:814–861.

Lai, C. D., and M. Xie. 2006. *Stochastic ageing and dependence for reliability*. New York: Springer.

Lancaster, H. O. 1982a. Chi-square distribution. In *Encyclopedia of Statistical Sciences, Volume 1*, ed. S. Kotz and N. L. Johnson, 439–442. New York: John Wiley & Sons.

Lancaster, H. O. 1982b. Dependence, measures and indices. In *Encyclopedia of Statistical Sciences, Volume 2*, ed. S. Kotz and N. L. Johnson, 334–339. New York: John Wiley & Sons.

Meier, D. R., and A. P. Basu. 1980. An investigation of kendall's τ modified for censored data with applications. *J. Statist. Planning Inference* 4:381–390.

Nelsen, R. B. 1992. Measures of association as measures of positive dependence. *Statistics and Probability Letters* 14:269–274.

Nelsen, R. B. 1999. *An introduction to copulas*. New York: Springer-Verlag.

Oakes, D. 1982. A concordance test for independence in the presence of censoring. *Biometrics* 38:451–455.

Oakes, D. 1989. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84:487–493.

Rényi, S. 1959. On measures of dependence. *Acta Mathematica Academia Scientia Hungarica* 10:441–451.

Robbins, H. 1954. A remark on the joint distribution of cumulative sums. *Annals of Mathematical Statistics* 25:614–616.

Robbins, H. 1967. Association of random variables, with applications. *Annals of Mathematical Statistics* 38:1466–1474.

Schweizer, B., and E. F. Wolff. 1981. On nonparametric measure of dependence for random variables. *Annals of Statistics* 9:879–885.

Spearman, C. 1904. The proof and measurement of correlation between two things. *Am. J. Psychiatr* 15:72–101.

AUTHOR BIOGRAPHIES

ROBERTO LU Dr. Roberto F Lu, PE is a Technical Fellow in the Boeing Research and Technology located in Seattle Washington. His first job at Boeing was a manufacturing engineer supporting the Skin and Spar fabrication of the Wing Responsibility Center in Auburn, Washington. He currently supports projects among organizations in Boeing Commercial Airplanes and Boeing Defense, Space & Security business units. He part-time teaches undergraduate and graduate level courses in Industrial and Systems Engineering at the University of Washington as an Affiliate Assistant Professor. He models business and production processes for leaders to optimize impacts of their decisions. He speaks at professional events worldwide, focusing upon discrete event simulation, analytical process optimization, global logistics, large scale production systems, lean manufacturing, robotic applications, and mass customization. His professional experiences prior to Boeing includes machine vision integrated robotic automation, tooling development, geometric dimensional tolerance establishment and inspection, non-destructive testing, metallurgical spectrographic analysis, and manufacturing statistical quality control at the Pilkington North America Libbey-Owens-Ford company and the Intermet-New River Castings company. He has more than 50 combined journal, conference, and patent publications. Roberto received his BS degree in Materials Science from Feng Chia University in Taiwan, first masters degree in Mechanical Engineering from Marquette University in Milwaukee, Wisconsin, second masters degree in Industrial and Systems Engineering from Virginia Tech in Blacksburg, Virginia, and third masters and PhD degrees in Industrial and Systems Engineering from University of Washington in Seattle, Washington. He is a licensed and registered Professional Engineer, a senior member of IIE, a member of INFORMS, SAE, an FE / PE exam committee member in industrial engineering at the National Council of Examiners for Engineering and Surveying, and an inaugural member of the International Institute of Mass Customization. His email address is <roberto.f.lu@boeing.com>.

SHUGUANG SONG Dr. Shuguang Song, has been working for The Boeing Company after obtaining his Ph.D in statistics from University of Washington in 2001. His key technical filed is statistical reliability. He was one of the main contributors to the mathematics of Airplane Health Management project. AHM is a diagnostic and prognostic service designed to increase airplane availability. This work led to a US patent and Boeing Silver Award. He was the key statistician for Maintenance Interval Determination and Optimization Tool (MIDOT) project which helps the Boeing 787 maintenance program achieve target maintenance interval goals. MIDOT was selected for the 2008 Boeing Special Invention Awards and 2009 Boeing Breakthrough Achievement. MIDOT led to the replication and development of a new fleet planning and optimization concept that displays complex reliability data and statistical analysis results to provide timely information for intelligent decision making. This work was selected for 2010 Boeing Technical Replication Award. Dr. Song has 3 U.S. patents pending and more than 25 publications and presentations. His email address is <shuguang.song@boeing.com>.