The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 219-231

Graphical Methods for Correlation and Independence

Chong Sun Hong[1] and Jang Sub Yoon[2]

Abstract

When the correlation of two random variables is weak, the value of one variable can not be used effectively to predict the other. Even when most of the values are overlapped, it is difficult to find a linear relationship. In this paper, we propose two graphical methods of representing the measures of correlation and independence between two random variables. The first method is used to represent their degree of correlation, and the other is used to represent their independence. Both of these methods are based on the cumulative distribution functions defined in this work.

Keywords : Correlation; Distribution function; Independence; Regression.

1. Introduction

A simple metric for describing bivariate data is to measure the degree of correlation between two random variables, X and Y. This can be done graphically using scatter plots, or analytically using various formulas. The most common and well-known statistic is Pearson's correlation coefficient. This measures the degree to which X and Y are linearly related. The correlation coefficient could play a role as a measure of the degree to which X (Y) can be used to predict Y (X). This works for all cases where the relationship between them is monotonic.

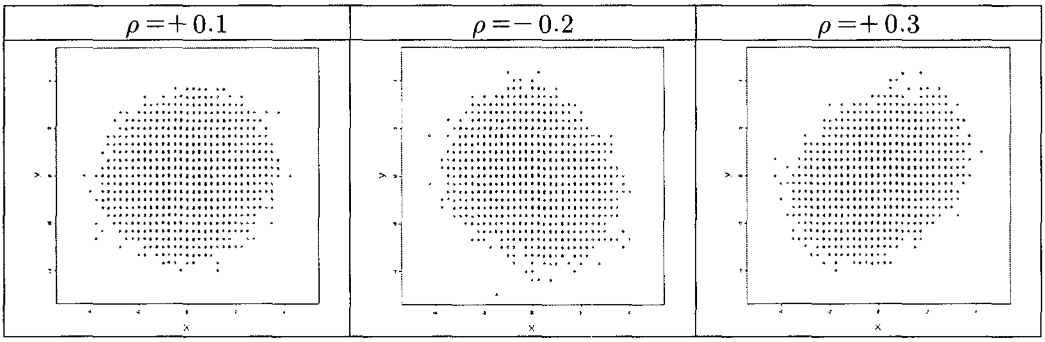
We consider some cases of weak correlation coefficients. The weak correlation means that its correlation coefficient has a value which is close to zero. A weak correlation coefficient might turn out to be significant with large sample size. When sample size is 400, the 5% two-tailed significant value of the correlation coefficient is 0.098. (see Snedecor and Cochran 1989; Myers and Well 1991). Under the situation of weak correlation coefficient, the value of one variable cannot be used effectively to predict the other. While it may be impossible to make individual predictions, it may still be possible to characterize aggregate behavior. This is done by linking the distribution of X and Y. In particular, we are

[1] Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea. Correspondence : cshong@skku.ac.kr

[2] Research Fellow, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.

interested in the case where knowing X would allow us to obtain the distribution of Y, but this is insufficient to make precise predictions regarding the exact value of Y, because the distributions for different values of X have a large amount of overlap.

Some examples are generated by using bivariate normal density function with weak correlation coefficients with sample size $n=10{,}000$. The data shown in <Figure 1.1> are latticed with a small interval (for example, 0.2). The scatter plots in <Figure 1.1> show that most values are overlapped and that linear or monotonic relationships cannot be found between the two variables. Hence, no meaningful relation can be derived from scatter plots and correlation coefficients under weakly correlated situations.



<Figure 1.1> Data with weak correlation coefficients

In this paper, we propose two graphical methods of representing the measures of correlation and/or independence between two random variables. The first method represents the degree of correlation by using the cumulative distribution function. This correlation graph is explained along with its properties in Section 2. The other method, which is described in Section 3, uses the independence theorem explained with the cumulative distribution function. This graphical method can be used to determine the independence of two random variables, so that the independence of two variables can be evaluated. In Section 4, two illustrated examples are given. The results of the two graphical methods are demonstrated in Section 5, in the case of two random variables which are not independent and whose correlation coefficients are close to zero. In Section 6, we derive the properties of the proposed methods and present our conclusions.

2. Correlation Graph

When a cumulative distribution function (CDF) of two random variables X and

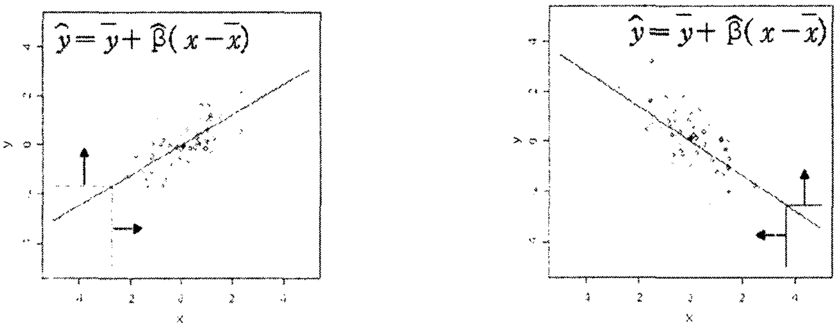
Y, $F_{XY}(x, y)$, is drawn on a two dimensional plane, it is not easy to understand what this three dimensional CDF might tell us regarding their degree of correlation. Hence, we might consider the following distribution function which turns out to be the two dimensional CDF :

$$P(X \leq x,\ Y \leq y_x) = F_{XY}(x, y_x)\,, \tag{2.1}$$

where y_x is the predicted values of the random variable, Y, at $X = x$ obtained from the estimated regression line, i.e. $y_x = \bar{y} + \hat{\beta}(x - \bar{x})$. The CDF in (2.1) can be defined only when the estimated regression coefficient is non-negative. In the case where the estimated regression coefficient is negative, the following probability, which can be interpreted with CDFs, is considered :

$$\begin{aligned} P(X > x,\ Y \leq y_x) &= F_Y(y_x) - F_{XY}(x, y_x) \\ &\equiv F^*_{XY}(x, y_x)\,. \end{aligned} \tag{2.2}$$

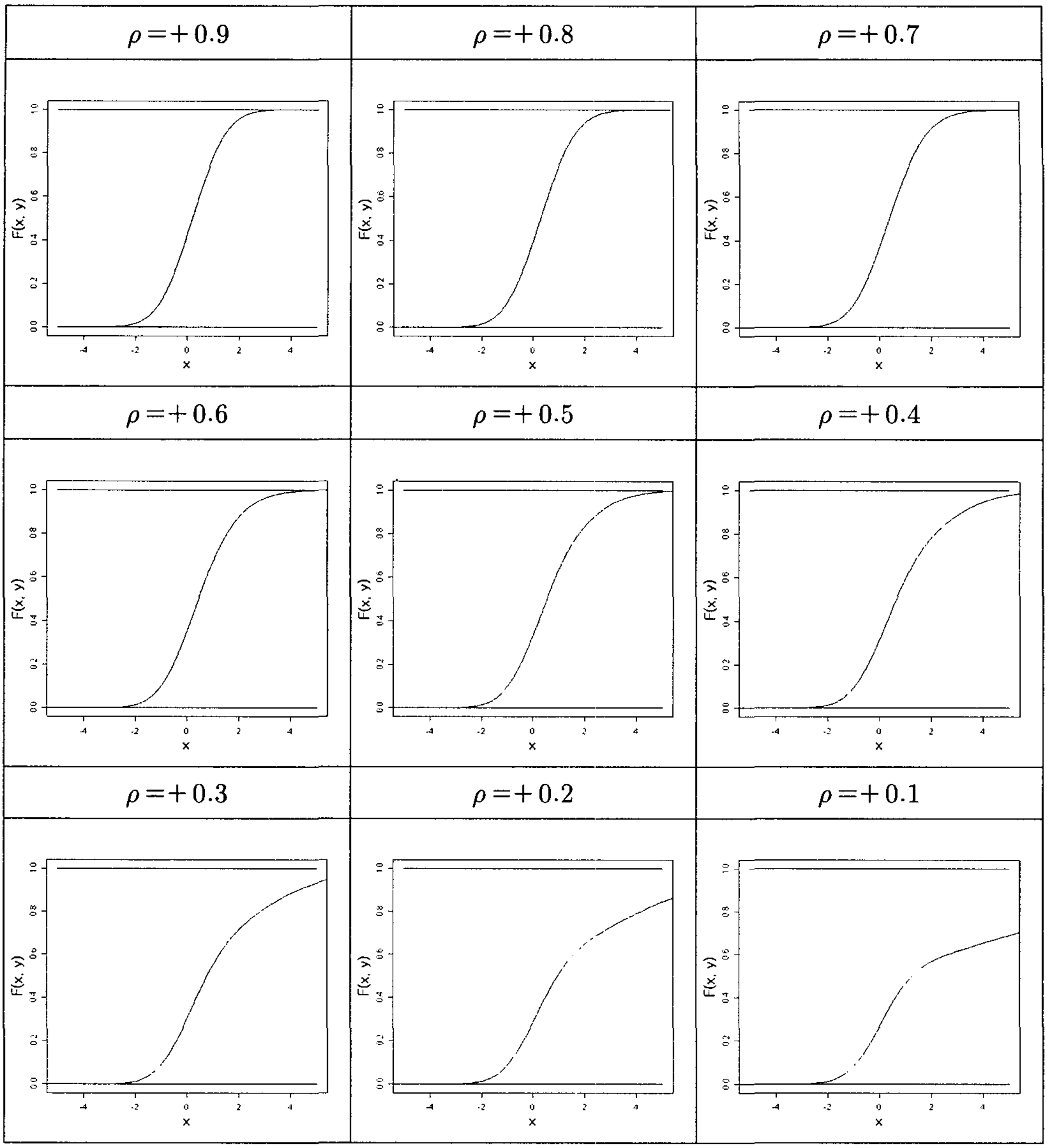
These two probabilities are obtained for all x, as shown in <Figure 2.1>. As x increases, $F_{XY}(x, y_x)$ in (2.1) increases from 0 to 1 for non-negative $\hat{\beta}$, and $F^*_{XY}(x, y_x)$ in (2.2) decreases from 1 to 0 for negative $\hat{\beta}$.



<Figure 2.1> Non-negative and negative correlation

Two probabilities, $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$, are demonstrated in <Figure 2.2> for some generated data following bivariate standard normal densities with various values of the correlation coefficients, ρ. This graphical method is referred to as "Correlation Graph."

From <Figure 2.2.A and B>, we can obtain the following results. The first is that $F_{XY}(x, y_x)$ is symmetric with $F^*_{XY}(x, y_x)$ when the absolute values of the correlation coefficient are equivalent. Secondly, $F_{XY}(x, y_x)$ increases rapidly as ρ goes from 0 to +1 (for non-negative $\hat{\beta}$), whereas $F^*_{XY}(x, y_x)$ decreases rapidly as ρ decreases from 0 to -1 (for negative $\hat{\beta}$).

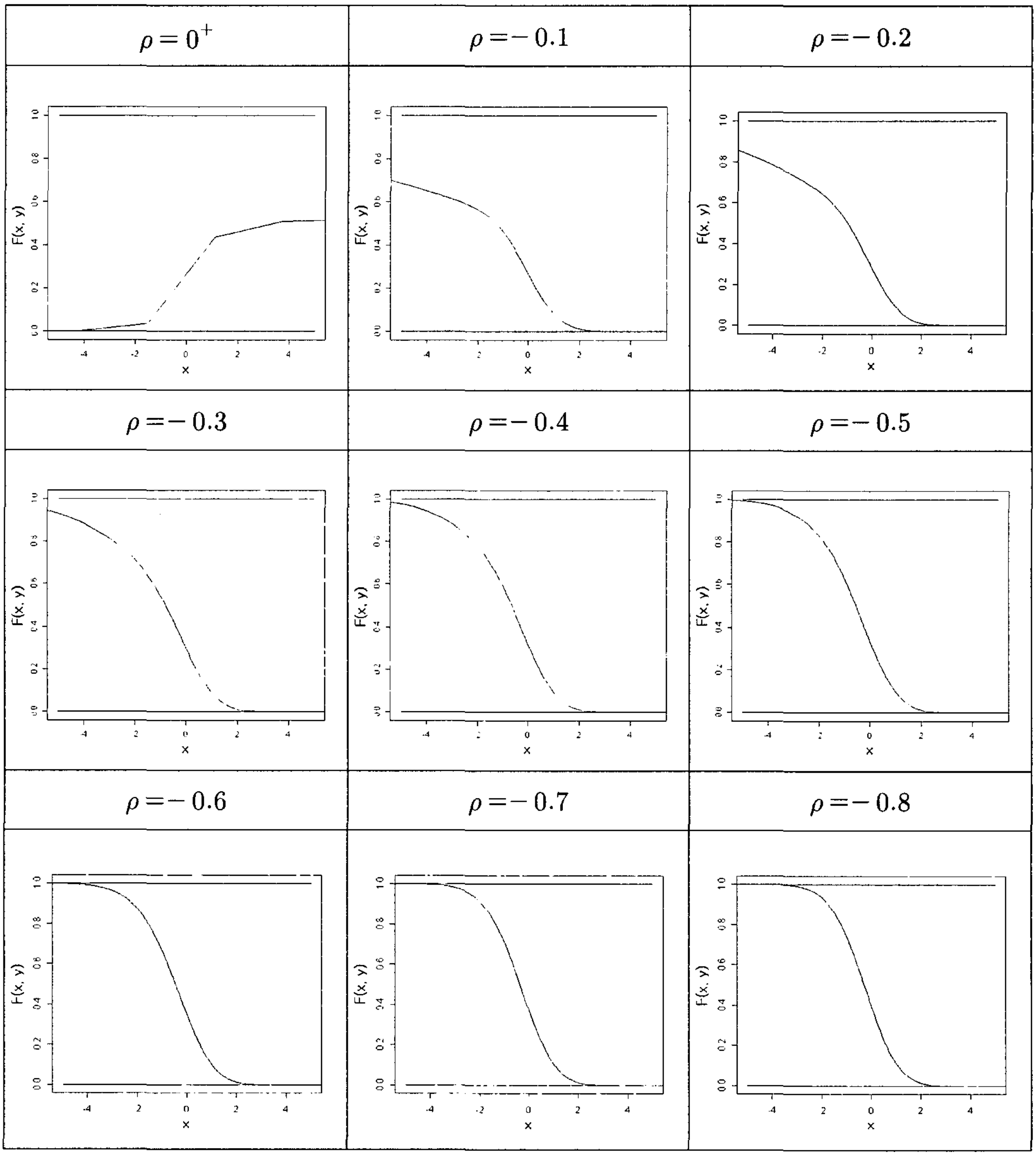

$\rho = +0.9$ $\rho = +0.8$ $\rho = +0.7$

$\rho = +0.6$ $\rho = +0.5$ $\rho = +0.4$

$\rho = +0.3$ $\rho = +0.2$ $\rho = +0.1$

<Figure 2.2.A> Correlation Graphs

These phenomena are also symmetric with respect to the absolute values of ρ. When the non-negative correlation coefficient becomes stronger, $F_{XY}(x, y_x)$ reaches its upper limit (value 1) more quickly. On the other hand, when the negative correlation coefficient becomes stronger, $F^{*}_{XY}(x, y_x)$ decreases more rapidly from its upper limit. When ρ is close to 0, the third result is obtained: both $F_{XY}(x, y_x)$ and $F^{*}_{XY}(x, y_x)$ do not reach their upper limit. In other words, when ρ is non-negative and weak, $F_{XY}(x, y_x)$ increases slowly and does not attain the value of 1 as x increases, while when ρ is negative and weak, $F^*_{XY}(x, y_x)$ has a value of less than 1 for negative values of x and decreases slowly to 0 as x increases.



$\rho = 0^+$

$\rho = -0.1$

$\rho = -0.2$

$\rho = -0.3$

$\rho = -0.4$

$\rho = -0.5$

$\rho = -0.6$

$\rho = -0.7$

$\rho = -0.8$

<Figure 2.2.B> Correlation Graphs

Therefore, both of $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$ in (2.1) and (2.2), respectively, might be used as measures of the degree of correlation between two random variables. If the Pearson correlation coefficient were also included with the correlation graphs based on $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$, one could understand the structure of bivariate data with greater ease.

3. Independence Graph

In the previous section, we found that with the shapes of the correlation graph based on $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$, the degree of correlation between two variables could be explained. Nonetheless, when ρ is weak, it is not easy to predict the correlation coefficients, as shown in <Figure 2.2>, since they do not reach their upper limit.

Now, we derive the following Lemma 1 from the well-known independence theorem:

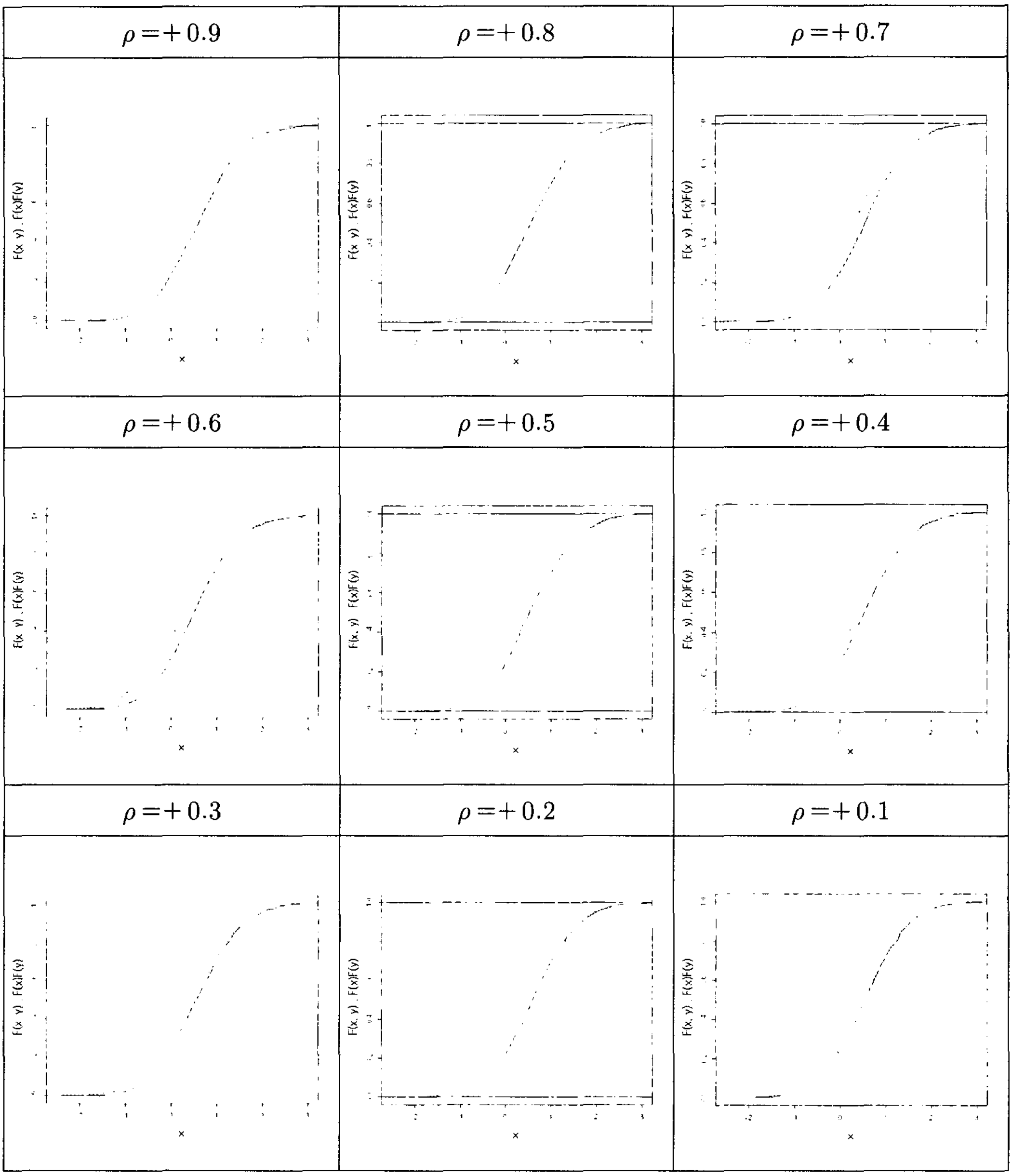
Lemma 1. Two random variables X and Y are defined to be stochastically independent, if and only if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all x and y. Hence, if two random variables are independent, $F_{XY}(x, x) = F_X(x)F_Y(x)$ is satisfied for all $X = x$ and $Y = x$.

In particular, when the estimate of the correlation coefficient approaches zero, y_x could be replaced by $y^*_x = \bar{y} + (x - \bar{x})$ at $F_{XY}(x, y_x)$ in equation (2.1), i.e. $\hat{\beta}$ is replaced by an arbitrary value '1' at $y_x = \bar{y} + \hat{\beta}(x - \bar{x})$ irrespective of whether $\hat{\rho}$ (or $\hat{\beta}$) is non-negative or negative. Then, we calculate and draw $F_{XY}(x, y^*_x)$ for all x, and compare $F_{XY}(x, y^*_x)$ with the product of $F_X(x)$ and $F_Y(y^*_x)$ in order to evaluate the independence of the two random variables. Based on Lemma 1, if $F_{XY}(x, y^*_x)$ and $F_X(x)F_Y(y^*_x)$ are overlapped for most x, we might say that the random variables X and Y are stochastically independent.

For the data used in <Figure 2.2>, we obtain $F_{XY}(x, y^*_x)$, $F_X(x)$, and $F_Y(y^*_x)$ for all x, where $y^*_x = x - \bar{x} + \bar{y}$, and draw the product, $F_X(x)F_Y(y^*_x)$, and $F_{XY}(x, y^*_x)$ on the same plot. These results are summarized in <Figure 3.1>, where $F_X(x)F_Y(y^*_x)$ and $F_{XY}(x, y^*_x)$ are represented by continuous and dotted lines, respectively. This method is referred to as "Independence Graph."

From <Figure 3.1.A and B>, we found that when the values of ρ are close to zero, both of $F_{XY}(x, y^*_x)$ and $F_X(x)F_Y(y^*_x)$ are greatly overlapped for most of x. Moreover, when ρ is positive $F_X(x)F_Y(y^*_x)$ is less than or equal to $F_{XY}(x, y^*_x)$, and when ρ has a negative value $F_{XY}(x, y^*_x)$ is less than or equal to

$F_X(x)\,F_Y(y_x^*)$ for all x.



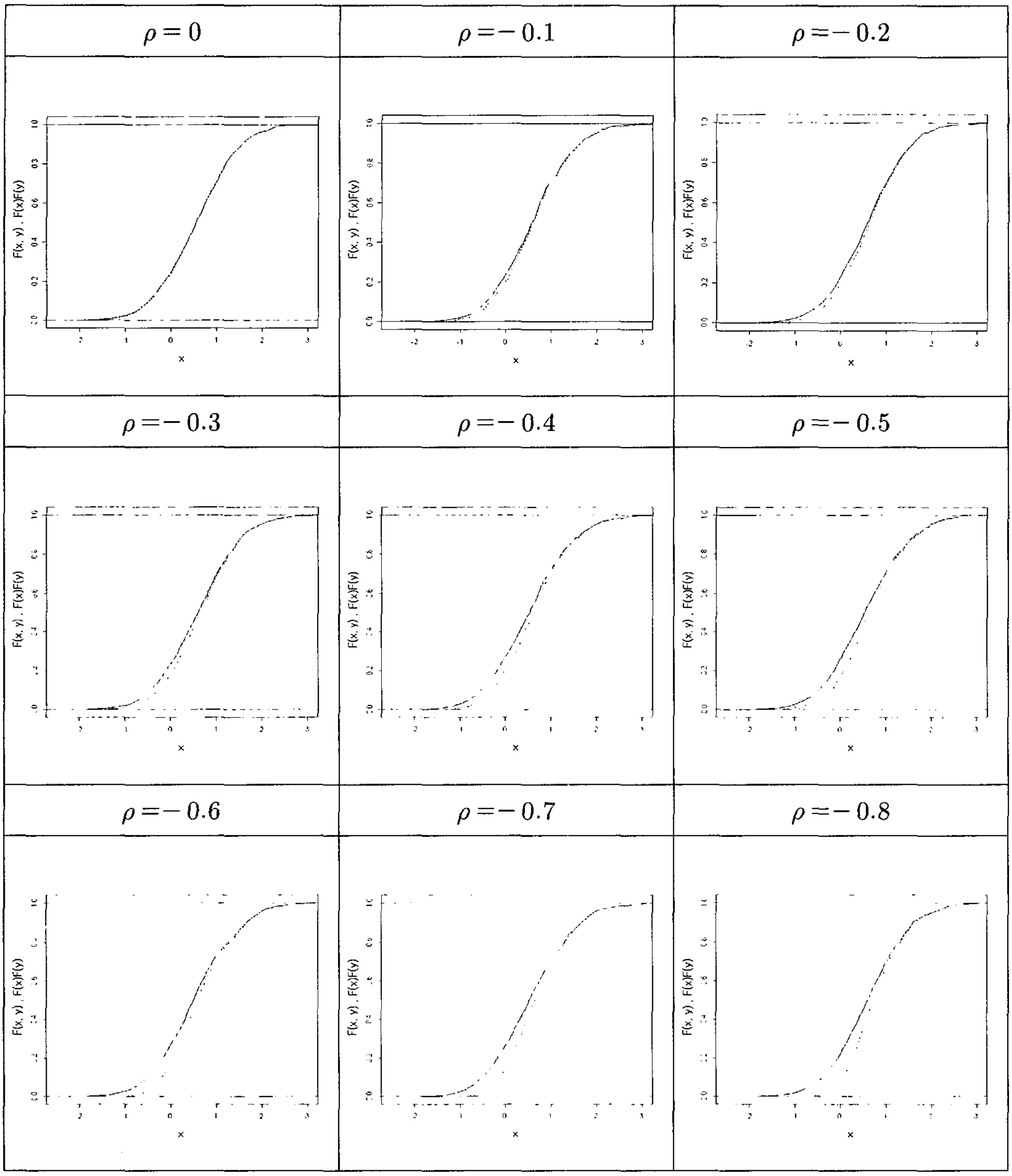
$\rho=+0.9$ $\rho=+0.8$ $\rho=+0.7$

$\rho=+0.6$ $\rho=+0.5$ $\rho=+0.4$

$\rho=+0.3$ $\rho=+0.2$ $\rho=+0.1$

<Figure 3.1.A> Independence Graphs

Therefore by comparing $F_{XY}(x, y_x^*)$ and $F_X(x)F_Y(y_x^*)$ for all x and $y_x^* = x - \overline{x} + \overline{y}$, we are able to determine the degree of independence of the two random variables. Hence, it can be concluded that if the Pearson correlation coefficient were to be used in conjunction with the independence graph as well as the correlation graph, it would be easy to comprehend the structure of bivariate data.


$\rho = 0$ $\rho = -0.1$ $\rho = -0.2$

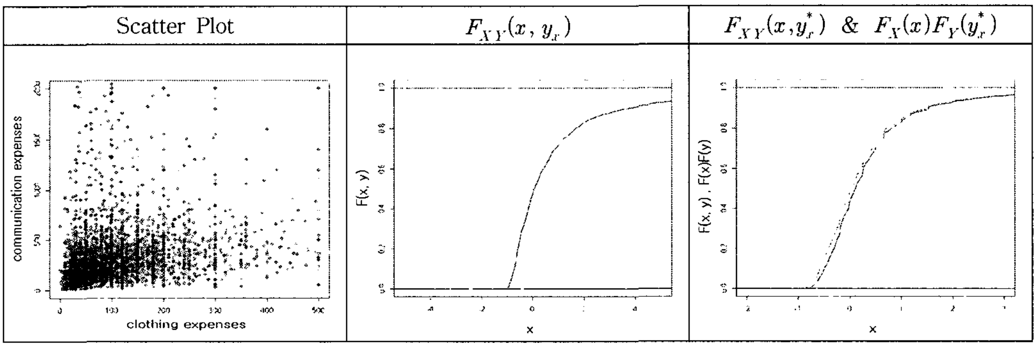
$\rho = -0.3$ $\rho = -0.4$ $\rho = -0.5$

$\rho = -0.6$ $\rho = -0.7$ $\rho = -0.8$

<Figure 3.1.B> Independence Graphs

4. Some Illustrated Examples

Two examples are illustrated in this section. The first data with large sample size N=4,486 is on Household Consumptions from January 1st 2000 to December 31th 2000 which was surveyed by Department of Social Statistics, Korean National Statistical Office during 2000. Korean National Statistical Office has made a survey of korean household consumptions on the whole nation throughout the country every five years in order to comprehend the structure of the korean standard of living and income · expenditures via the depth surveys on household assets including annual income, consumption expenditure, savings and liabilities, etc. Among a lot of variables in the data, we choose two variables: the clothing and communication expenses. And the correlation between the clothing expenses (X) and the communication expenses (Y) is of interest.


Scatter Plot

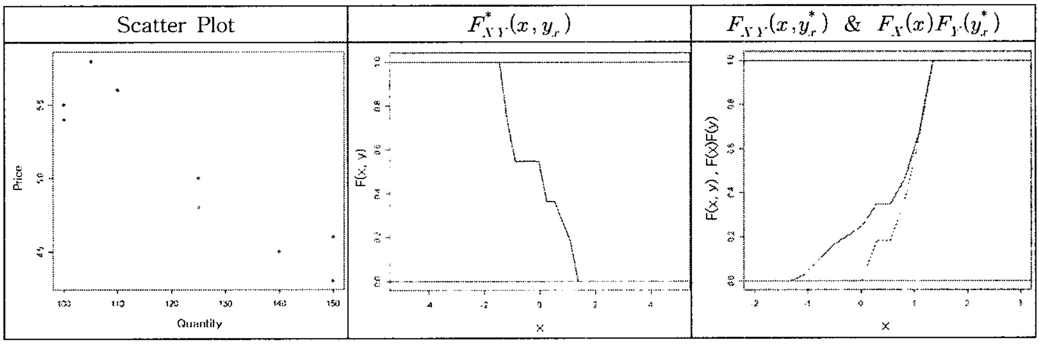
$F_{XY}(x, y_x)$

$F_{XY}(x,y_x^*)$ & $F_X(x)F_Y(y_x^*)$

<Figure 4.1> Distributions for clothing and communication expenses

A scatter plot showing the clothing and communication expenses is shown in the left-hand plot of <Figure 4.1>. Quite obviously, there exists weak linear relationship between these two variables. The Pearson correlation coefficient is found to have a positive and low value, $\hat{\rho}=0.2167$. The bivariate values are standardized to compare with the correlation and independence graphs in <Figure 2.2> and <Figure 3.1>. (Hereafter, the random variables X and Y are regarded as being standardized ones.) Then, we calculates $F_{XY}(x, y_x)$ and both $F_{XY}(x, y_x^*)$ and $F_X(x)F_Y(y_x^*)$, where $y_x = 0.2167x$ and $y_x^* = x$. These probabilities are represented in <Figure 4.1>. From the center plot in <Figure 4.1>, we find that $F_{XY}(x, y_x)$ is increasing and does not reach its upper limit. Also, from the right-hand plot, it is found that $F_{XY}(x, y_x^*)$ and $F_X(x)F_Y(y_x^*)$ are very close, and $F_X(x)F_Y(y_x^*)$ (real line) is slightly less than $F_{XY}(x, y_x^*)$ (dotted line). Hence, it might be concluded that the correlation coefficient of this data is expected to be positive and weak and two random variables are independent based on the correlation and independence graphs in <Figure 4.1>, even though its correlation coefficient is significant with large sample size.

Another set of data is taken from an introductory textbook of Spirer (1975). The producers of yak's milk purchase yak's milk in the market each day at a price determined by a variety of economic factors. These producers take a sample from their records of 11 days' values for the quantity (X) and cost per metric ton (Y) of milk purchased. <Figure 4.2> shows the 11 days' values. One would expect the price to be paid, as soon as the amount to be purchased on the next day is determined. This data set is quite small (N=11), and has a strong relationship ($\hat{\rho}$=-0.9402). All of the values are also standardized in order to compare them with the data in <Figure 2.2> and <Figure 3.1>. Then, we calculate $F^*_{XY}(x, y_x)$ in (2.2) and both $F_{XY}(x, y^*_x)$ and $F_X(x)F_Y(y^*_x)$, where $y_x = -0.9402x$ and $y^*_x = x$. Then these probabilities are represented in <Figure 4.2>. From the center plot in <Figure 4.2>, it can be seen that $F^*_{XY}(x, y_x)$ decreases rapidly from its upper limit. From the right-hand plot in <Figure 4.2>, we find that $F_{XY}(x, y^*_x)$ and $F_X(x)F_Y(y^*_x)$ are not overlapped, and $F_X(x)F_Y(y^*_x)$ is greater than $F_{XY}(x, y^*_x)$ for most x. Hence, we might conclude that the correlation coefficient of this data is strongly negative, and the two random variables are not independent.


Scatter Plot

$F^*_{XY}(x, y_x)$

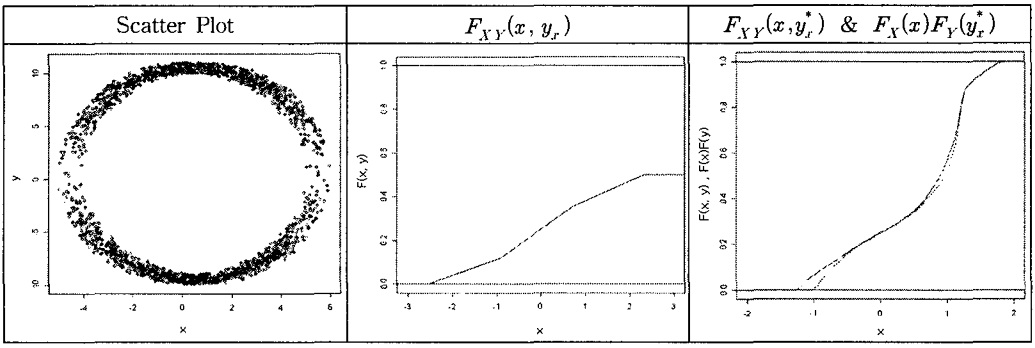
$F_{XY}(x, y^*_x)$ & $F_X(x)F_Y(y^*_x)$

<Figure 4.2> Distributions for yak's milk (price vs. quantity)

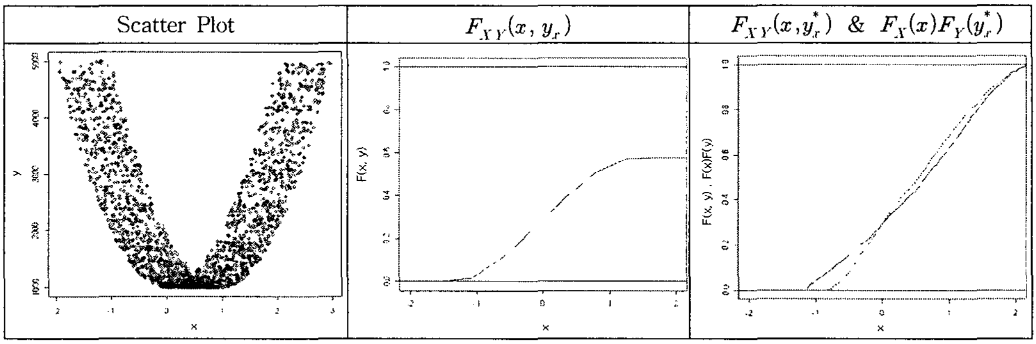
5. Other Uncorrelated Data

We consider some uncorrelated bivariate data, in which there exist more than a linear relationship between two variables, such as the elliptical and quadratic relationships found in the scatter plots in <Figures 5.1 and 5.2>. Two sets of data are generated with the sample size N=2,000, and the probabilities $F_{XY}(x, y_x)$, $F_{XY}(x, y^*_x)$ and $F_X(x)F_Y(y^*_x)$ are then obtained, where $y_x = \hat{\beta}x$ and $y^*_x = x$, and where $\hat{\beta} = \hat{\rho} = 0.0008$ and $\hat{\beta} = \hat{\rho} = 0.0014$ for the first and second set of data, respectively. The correlation and independence graphs are shown in <Figures 5.1

and 5.2>.



<Figure 5.1> Elliptical shaped data



<Figure 5.2> Quadratic shaped data

From the center plots in <Figures 5.1 and 5.2>, both $F_{XY}(x, y_x)$'s are found to increase very slowly and do not reach their upper limit, so that we might conclude that the correlation coefficients of the two random variables in both set of data are non-negative and weak.

Also from the right-hand plots in <Figure 5.1>, it can be seen that $F_{XY}(x, y_x^*)$ and $F_X(x)F_Y(y_x^*)$ are nearly overlapped for most x. Hence one might be conclude that its correlation coefficient is expected to be positive and weak and two random variables are independent based on the correlation and independence graphs in <Figure 5.1>.

Nonetheless, it can be seen that $F_{XY}(x, y_x^*)$ and $F_X(x)F_Y(y_x^*)$ are not overlapped from the right-hand plots in <Figure 5.2>. In particular, $F_X(x)F_Y(y_x^*)$ (real line) is greater than $F_{XY}(x, y_x^*)$ (dotted line) for negative x, but $F_X(x)$ $F_Y(y_x^*)$ is less than $F_{XY}(x, y_x^*)$ for positive x. Hence, from <Figure 5.2> we can conclude that there exists a (not linear) relationship between two random variables, even though their correlation coefficient is close to zero.

Therefore, if the Pearson correlation coefficient were to be included with the correlation graph based on $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$ as well as the independence graph based on $F_{XY}(x, y^*_x)$ and $F_X(x) F_Y(y^*_x)$, we could determine whether two random variables in bivariate data sets are correlated or linearly independent.

6. Conclusion

In this paper, we consider two probabilities for the correlation between two random variables : $F_{XY}(x, y_x) = P(X \leq x,\ Y \leq y_x)$ for a non-negative correlation, and $F^*_{XY}(x, y_x) = F_Y(y_x) - F_{XY}(x, y_x)$ for a negative correlation, where y_x is a predicted value obtained from the estimated regression line at $X = x$, i.e. $y_x = \bar{y} + \hat{\beta}(x - \bar{x})$. The correlation graph based on $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$ plays an important role in evaluating the degree of correlation between two random variables.

Using $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$, the correlation coefficient can be evaluated with ease. However, when the correlation coefficient is weak, $F_{XY}(x, y_x)$ and $F^*_{XY}(x, y_x)$ do not attain a value of 1. Based on the fact that $F_{XY}(x, x) = F_X(x) F_Y(x)$ for all $X = x$ and $Y = x$, we compare $F_{XY}(x, y^*_x)$ with $F_X(x) F_Y(y^*_x)$, where $y^*_x = \bar{y} + (x - \bar{x})$. If both $F_{XY}(x, y^*_r)$ and $F_X(x) F_Y(y^*_x)$ are greatly overlapped for most x, then we could say that the correlation coefficient is close to 0.

Moreover, when the correlation coefficient has a positive value, $F_X(x) F_Y(y^*_x)$ is less than $F_{XY}(x, y^*_x)$, and when the correlation coefficient is negative, $F_{XY}(x, y^*_x)$ is less than $F_X(x) F_Y(y^*_x)$ for all x. Hence, the independence graph based on both $F_{XY}(x, y^*_x)$ and $F_X(x) F_Y(y^*_x)$ can be used as a method to explore the independence of two random variables.

For some uncorrelated bivariate data, in which there exists more than linear relationship between two variables, we could find that $F_{XY}(x, y_x)$ or $F^*_{XY}(x, y_x)$ does not attain the value of 1, and $F_{XY}(x, y^*_x)$ may not be overlapped with $F_X(x) F_Y(y^*_x)$ for most x. In some cases, it can be happened that $F_{XY}(x, y^*_x)$ is crossed over $F_X(x) F_Y(y^*_x)$.

Both the correlation graph and the independence graph proposed in this paper are generally constructed with raw values of the data. In other words, for drawing $F_{XY}(x, y_x)$, $F^*_{XY}(x, y_x)$ and $F_X(x)F_Y(y^*_x)$, the values of y_x and y^*_x are obtained from raw values of the data, where $y_x = \bar{y} + \hat{\beta}(x - \bar{x})$ and $y^*_x = \bar{y} + (x - \bar{x})$. And these graphs are also constructed with standardized values. In particular, in order to compare to the correlation and independence graphs in <Figure 2.2 and 3.1>, the raw data are standardized for drawing these graphs as we did in Section 4 and 5. Then x, $y_x = \bar{y} + \hat{\beta}(x - \bar{x})$ and $y^*_x = \bar{y} + (x - \bar{x})$ might be replaced as (z_x), $(z_y)_x = \hat{\rho}(z_x)$ and $(z_y)^*_x = (z_x)$, respectively, where (z_x) and (z_y) are standardized values of x and y, respectively, and $\hat{\rho}$ is not only the estimated correlation coefficient of x and y but also the estimated regression coefficient of (z_x) and (z_y). Hence the correlation graph and the independence graph constructed with standardized values (z_x) and (z_y) can evaluate the degree of the correlation and independence more precisely by comparing with those in <Figure 2.2 and 3.1>.

Therefore, the two graphical methods proposed in this paper can be helpful to understand the structure of bivariate data if the Pearson correlation coefficient and scatter plots are used simultaneously.

References

[1] Myers, J. and Well, A.D. (1991). *Research Design and Statistical Analysis.* Lawrence Erlbaum Associates, Inc.

[2] Snedecor, George W. and Cochran, William G. (1989). *Statistical Methods,* Eighth Edition. Iowa State University Press.

[3] Spirer, H.F. (1975). *Business Statistics, A Problem-Solving Approach.* Richard D. Irwin, Inc.

[Received December 2005, Accepted April 2006]