PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1162 2007
Applications of Mathematics and Statistics in Economics

Stanisław Heilpern

Wrocław University of Economics, Poland

GRAPHICAL METHODS FOR STUDY OF DEPENDENCE

1. Introduction

Recently, many authors investigate the dependent random variables in the theoretical economic papers. The independence is very comfortable property from theoretical point of view, but in many situations it is not realistic in the practice. In the insurance, finance, mainly in the risk management or in the social science, the variables are often dependent. This paper is devoted to the graphical methods which can be applied in the study of dependence, as an aid of the proper statistical methods.

First, we introduce the theoretical foundations of the dependence problems. We study the correlation coefficients and the copulas. Second section is devoted to the scatterplots. We introduce and show the graphical presentation of the tail dependence concepts. In the last part of the paper we present the chi-plots and Kendall plots, the graphical methods very useful for the investigation of dependence.

All calculations and graphs were made in Excel.

2. Theoretical foundations

Now, we present the basic information connected with the dependence concepts. Let $H(x, y)$ be the continuous joint cumulative distribution function of random variables X and Y, $F(x)$ and $G(y)$ be the margins continuous distributions. The correlation coefficients are the monotone measures of association. The most popular Pearson coefficient

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}$$

is the measure of the linear correlation only. Another well known correlation coefficients: Spearman ρ_S and Kendall τ, are the measures of the rank correlation. They are defined by the following formulas:

$$\rho_S = \rho(F(X), G(Y)),\ \tau = P((X - X')(Y - Y') > 0) - P((X - X')(Y - Y') < 0),$$

where (X', Y') is the independent copy of pair (X, Y).

The Spearman and Kendall coefficients are invariant under the monotone transformations. The Pearson coefficient is not invariant under such transformations. If random variables X and Y are independent, then $\rho = \rho_S = \tau = 0$.

The correlation coefficients describe the dependence by one number. Another way of studying the dependence is based on copulas. They give us more information. The *copula C* is the function which joins the margins with the joint cumulative distributions:

$$H(x, y) = C(F(x), G(y)).$$

We can define it by formula [5; 6]

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)).$$

When the margins distributions F and G are continuous, the copula is univocally determinated [5; 6]. We say that the copula describes the dependence structure. For independent random variables, the copula is the simple, product function:

$$C(u, v) = uv.$$

The families of copulas determined by parameters are often used in practice. The Archimedean families are the most popular and simple families. They are induced by one-dimensional generators φ, which separates the variables, so the *Archimedean* copula takes the form:

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)).$$

We give below the formulas of three popular Archimedean copulas [5]:

$$C(u,v) = \frac{1}{\alpha}\ln\left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1}\right) \qquad \textit{Frank},$$

where parameter $\alpha \neq 0$,

$$C(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha} \qquad \textit{Clayton},$$

where $\alpha > 0$, and

$$C(u, v) = \exp(-((-\ln u)^{\alpha} + (-\ln v)^{\alpha})^{1/\alpha}) \qquad \textit{Gumpel},$$

where $\alpha \geqslant 1$.

Another families of copulas are the elliptical families. They are based on the correlation matrix **R**. Using the normal distribution we obtain the *Gauss* family of copulas [1]:

$$C_{\mathbf{R}}^{G}(u,v)=H\left(\Phi^{-1}(u),\Phi^{-1}(v)\right),$$

where H is a joint distribution function of the standard normal random variables T_1, T_2 with correlation matrix **R** and Φ is a distribution function of standard normal random variable. The *t-Student* family of copulas is based on the t-Student distribution in the similar way (see [1]).

The rank correlation coefficients: Spearman and Kendall, are univocally determined by copulas. We obtain the following formulas:

$$\rho_S(X,Y)=12\int_0^1\int_0^1 C(u,v)dudv-3,\quad \tau(X,Y)=4\int_0^1\int_0^1 C(u,v)dC(u,v)-1.$$

For Archimedean copulas with the generator φ, the Kendall coefficient takes the simpler form

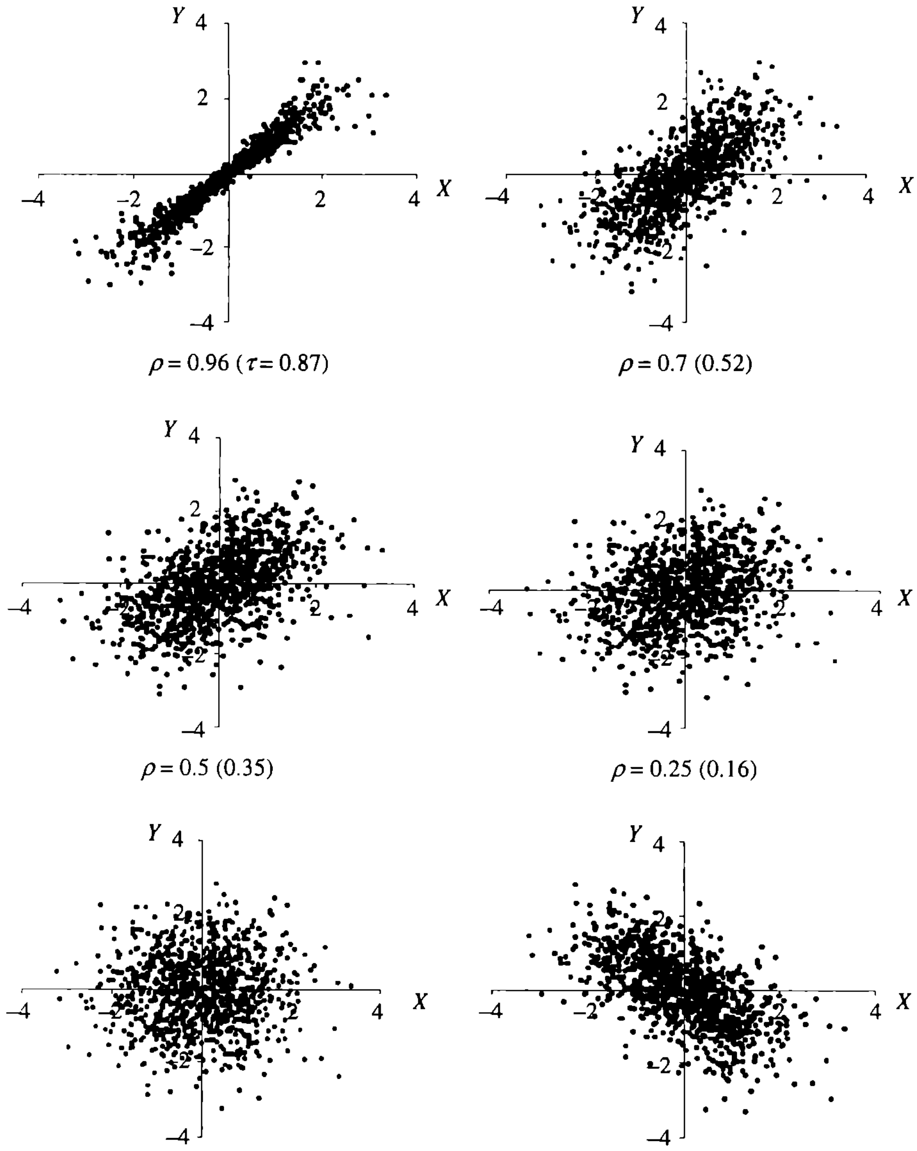
$$\tau(X,Y)=1+4\int_0^1\frac{\varphi(u)}{\varphi'(u)}du.$$

The Pearson correlation coefficient ρ, which describes the linear correlation, depends not only on the copula, but also on the marginal distribution. For instance, for the Archimedean copula we obtain

$$\rho=-\left(\frac{\varphi''(1)}{\varphi'(1)}+1\right)\frac{E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}.$$

3. Scatterplots

The *scatterplots* are the simplest methods, which reflect the dependence by graphical way. Figure 1 shows the scatterplots of the simulated two-dimensional data (n = 1072). The dependence is described by Frank copulas, and different correlation coefficients, and the standard normal margins. We can observe the differences of the shapes of the graphs, when the dependence changes from the almost perfect dependence to the independence and the negative dependence, so we can distinguish the different degrees of dependence. The different copulas give us the different scatterplots for the same correlation coefficient. In Figure 2 we see the scatterplots of the simulated data (n = 1072), when dependence is described by the following copulas: Frank, Clayton, Gumpel, Gauss and t-Student. All datasets are characterized by the same correlation coefficient $\rho = 0.4$ ($\tau \approx 0.26$).



$\rho = 0.96$ ($\tau = 0.87$)

$\rho = 0.7$ (0.52)

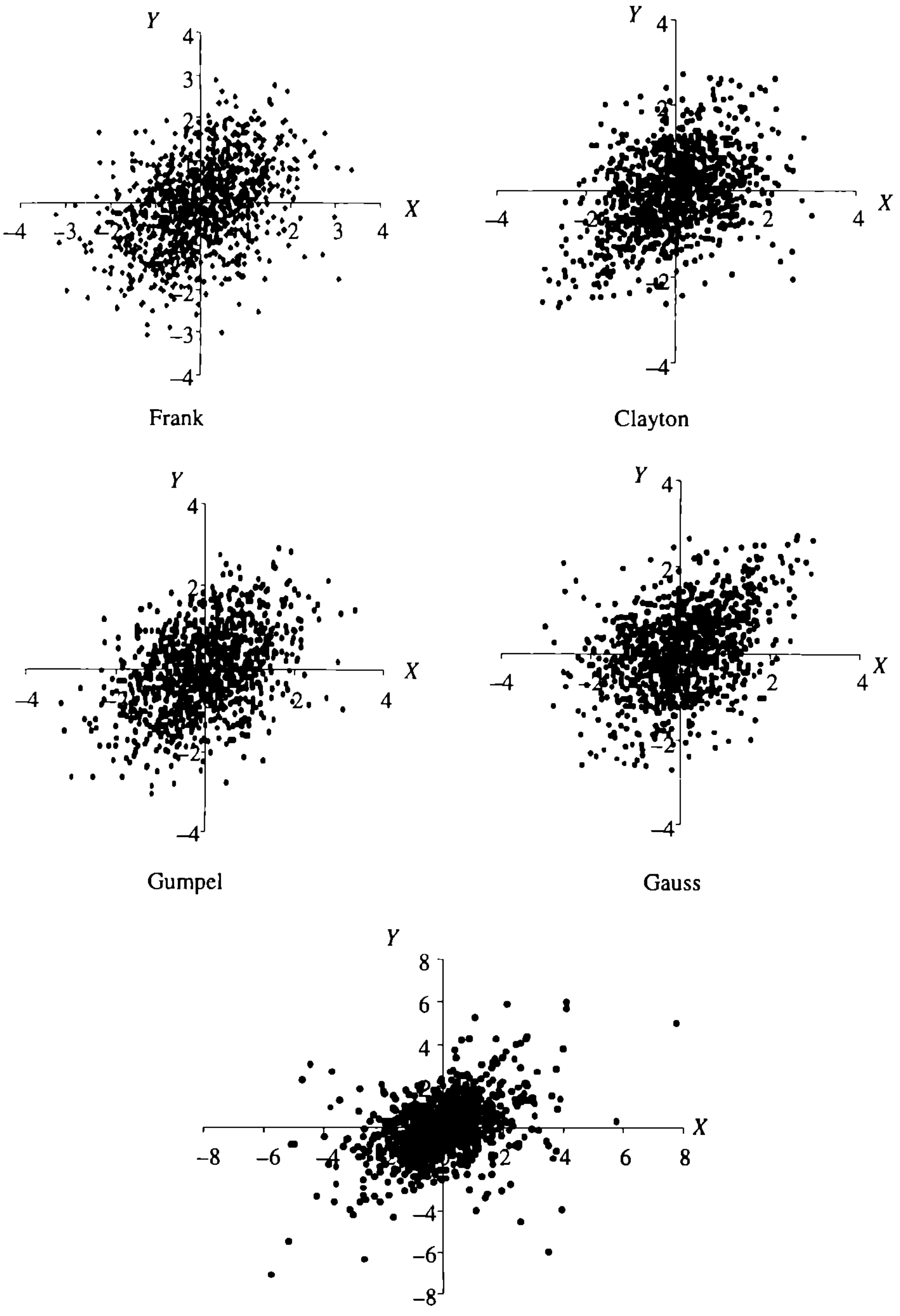
$\rho = 0.5$ (0.35)

$\rho = 0.25$ (0.16)

$\rho = 0$

$\rho = -0.62$ (–0.46)

Fig. 1. Scatterplots of the simulated data:
normal margins, Frank copula, different correlation coefficients



Frank

Clayton

Gumpel

Gauss

Fig. 2. Scatterplots of the simulated data: normal margins, $\rho = 0.4$, different copulas

We can study the asymptotic, *tail dependence*, for big or small values of random variables. We can use the *coefficient of upper* λ_U or *lower* λ_L *tail dependence* to this end. They are defined by the following formulas:

$$\lambda_U = \lim_{u\to 1} P(Y > F^{-1}(u) \mid Y > G^{-1}(u)),\ \lambda_L = \lim_{u\to 0} P(Y > F^{-1}(u) \mid Y > G^{-1}(u)).$$

The coefficients of tail dependence are univocally determined by copulas. We have

$$\lambda_U = \lim_{u\to 1}\frac{1-2u+C(u,u)}{1-u},\ \lambda_L = \lim_{u\to 0}\frac{C(u,u)}{u}.$$

We obtain the upper (lower) tail independence when $\lambda_U = 0$ ($\lambda_L = 0$).

The Frank and Gauss copulas generate tail independence, *t*-Student – copula lower and upper tail dependence, Clayton copula – only lower, and Gumpel copula – upper tail dependence [1; 5]. The spike in the cloud of point in the scatterplot indicates the tail dependence. We can see this fact in the Fig. 3.

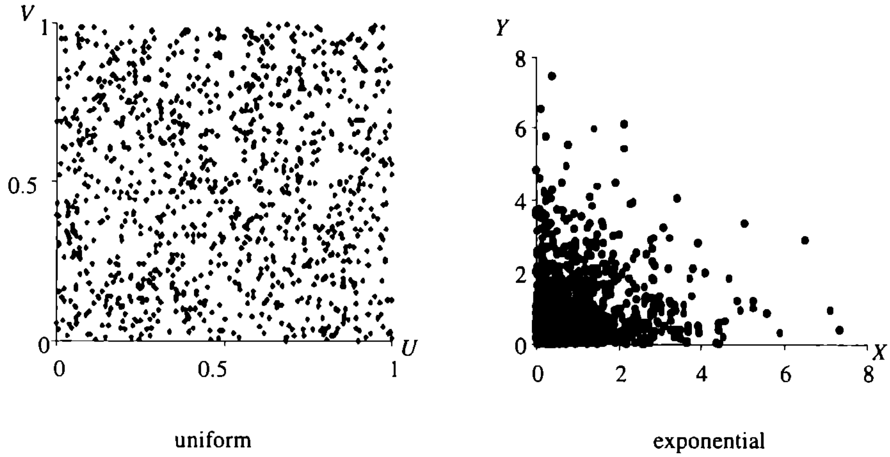


Fig. 3. Scatterplots of the independent simulated data: uniform and exponential margins, Frank copula

The tail dependences are very interesting in the insurance and finance problems, e.g. in the management of catastrophic risks. The existence of tail dependence is more dangerous from insurer point of view, since extreme losses have tendency to occur together.

The scatterplots are very simple and popular, but they depend on the margin distributions. For instance, for the normal, uniform and exponential margins we obtain the fundamentally different scatterplots for independent variables (see Fig. 3). This fact is very important, because the margin distributions are often unknown.

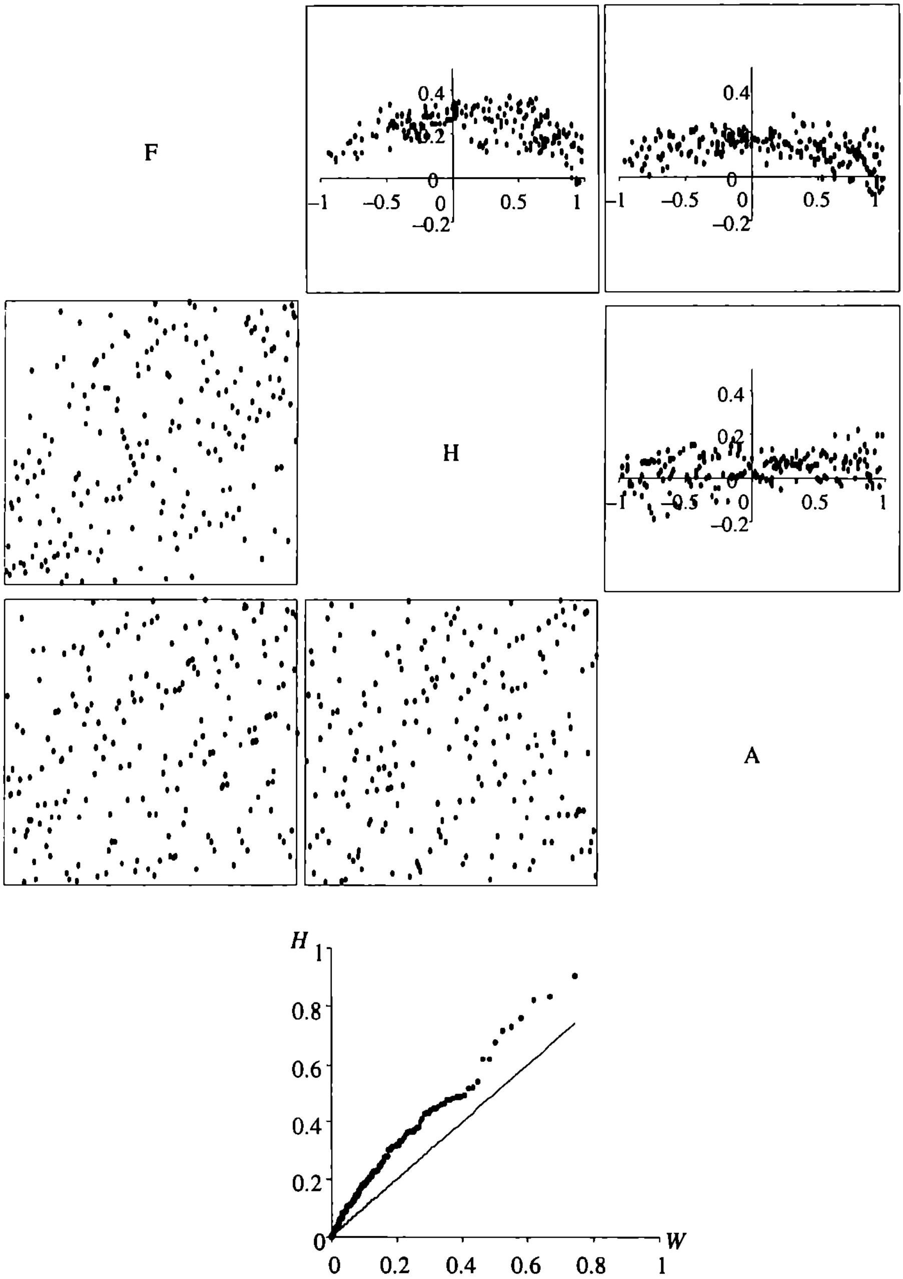


Fig. 4. Example of composite plot and Kendall plot for three-dimensional data

The *rank scatterplots*, scatterplots based on the rank data, have not this fault. They are invariant under monotone transformations. The rank scatterplots for all datasets used in Fig. 3 look like the scatterplot for the uniform margins. They are robust on outliers, but they are not good for the small associations. In such situations, we cannot distinguish obtained graphs (see Fig. 4).

4. Chi-plot

The chi-plot is other plot which can be used to graphical presentation of dependence. It was introduced by Fisher and Switzer [2; 3]. The construction of chi--plot is based on the empirical cumulative distributions: the joint and the margins. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample and $\#A$ denote the number of elements of A, set

$$H_i = \frac{1}{n-1}\#\{j \neq i : X_j \leqslant X_i, Y_j \leqslant Y_i\},$$

$$F_i = \frac{1}{n-1}\#\{j \neq i : X_j \leqslant X_i\},\ G_i = \frac{1}{n-1}\#\{j \neq i : Y_j \leqslant Y_i\}.$$

A *chi-plot* is a scatterplot composed of pairs (λ_i, χ_i), where

$$\lambda_i = 4\text{sign}((F_i - 0.5)(G_i - 0.5))\max\{(F_i - 0.5)^2, (G_i - 0.5)^2\},$$

$$\chi_i = \frac{H_i - F_iG_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}.$$

The λ_i is a signed distance of the pair (X_i, Y_i) from the median of dataset. Fisher and Switzer [2] recommended that only points for which $|\lambda_i| < 4\left(4/(n-1)-0.5\right)^2$ ought to be plotted, so chi-plot is robust on outliers.

The second coordinate χ_i is a signed measure of departure from independence. The authors introduced the control limits $\chi = \pm\ c_p/\sqrt{n}$, where $c_p \approx 1.78$ is determined by Monte Carlo simulations [2]. The approximately 95% of the pairs lie between these limits in the independence case. The big distance of the points of the chi-plot from the horizontal axis indicates the big dependence and the positive values of this plot indicate the positive dependence and negative values the negative dependence. We can observe these tendencies in the Fig. 5, which shows the chi--plots of the 200 simulated data, where the dependences are described by Frank copulas with the standard normal margins and with the different correlation coefficients as in Fig. 1. At the bottom of Fig. 5 we see the cases of the perfect dependences: $\rho = \pm 1$, connected with the commonotinicity.

The chi-plot is rank-based plot connected with copula and rank correlation coefficients: Spearman and Kendall. We obtain the following formula [3]:

$$\sum_{i=1}^{n}(H_i - F_iG_i) = \frac{n}{12}\left(3\tau_n - \frac{n+1}{n-1}\rho_n\right),$$

which links the numerator of the definition of χ with the empirical rank correlation coefficients: Spearman ρ_n and Kendall τ_n.

5. Kendall plots

The Kendall plot is the easier plot which reflects the dependence, too. It is introduced by Genest and Boies [4] and it is the generalization of Q-Q plot based on the order statistics of the empirical joint distribution H_i. A *Kendall plot* is a scatterplot composed of pairs $(W_{i:n}, H_{(i)})$, where

$$H_{(1)} \leqslant \ldots \leqslant H_{(n)},$$

and $W_{i:n} = E(H_{(i)})$ under the null hypothesis of independence.

The expectation of the i-th order statistics $W_{i:n}$ can be computed by formula

$$W_{i:n} = n\binom{n-1}{i-1}\int_0^1 w(K_0(w))^{i-1}(1 - K_0(w))^{n-i}\, dK_0(w),$$

where K_0 is distribution function of H_i under null hypothesis of independence. This distribution function is equal

$$K_0(w) = P(UV \leqslant w) = w - w\ln(w),$$

where U, V are the independent, uniform random variables determined on the interval [0, 1].

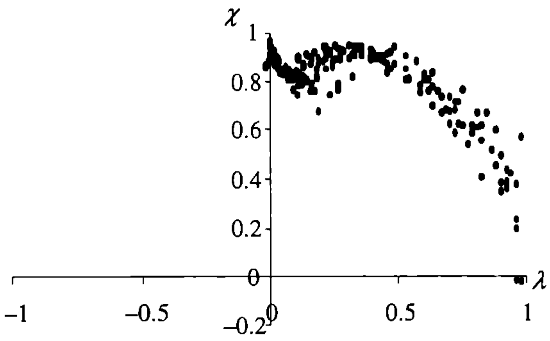
The Figure 6 shows the Kendall plots of the dataset used in Fig. 2. The departure from the diagonal line indicates the degree of dependence.

The Kendall plot is strictly connected with Kendall coefficient and copula. We can calculate the empirical Kendall coefficient τ_n using the empirical joint distribution:

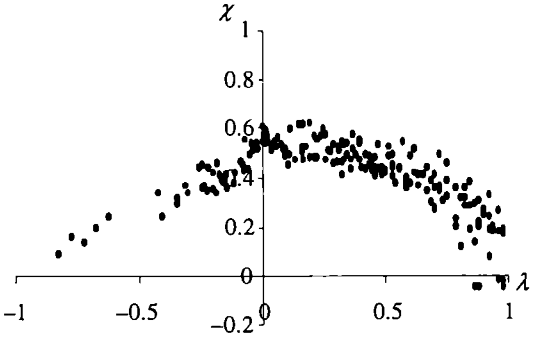
$$\tau_n = 4\left(\frac{H_1 + \ldots + H_n}{n}\right) - 1.$$

Let us investigate the cumulative distribution function K of random variable $H(X, Y)$, i.e.

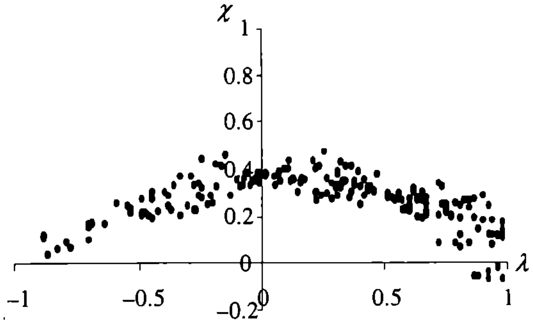
$$K(w) = P(H(X, Y) \leqslant w).$$



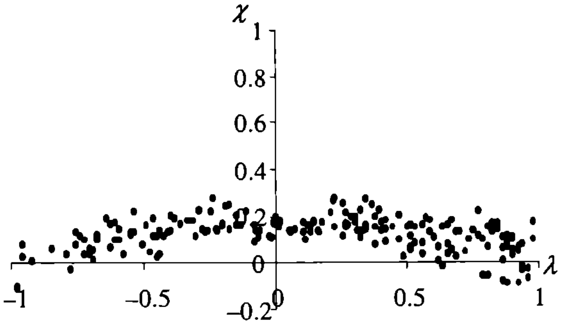
$\rho = 0.96\ (0.87)$



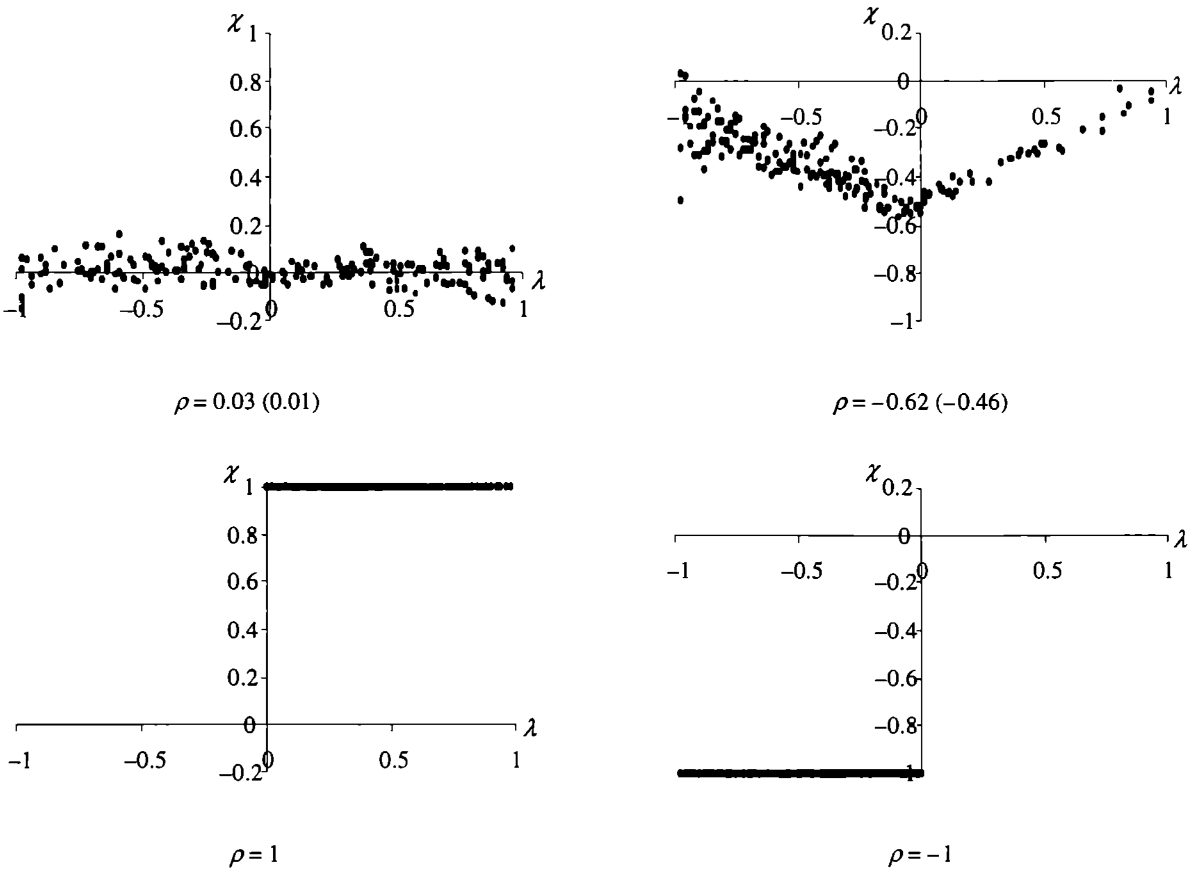
$\rho = 0.7\ (0.52)$



$\rho = 0.5\ (0.35)$



$\rho = 0.25\ (0.16)$



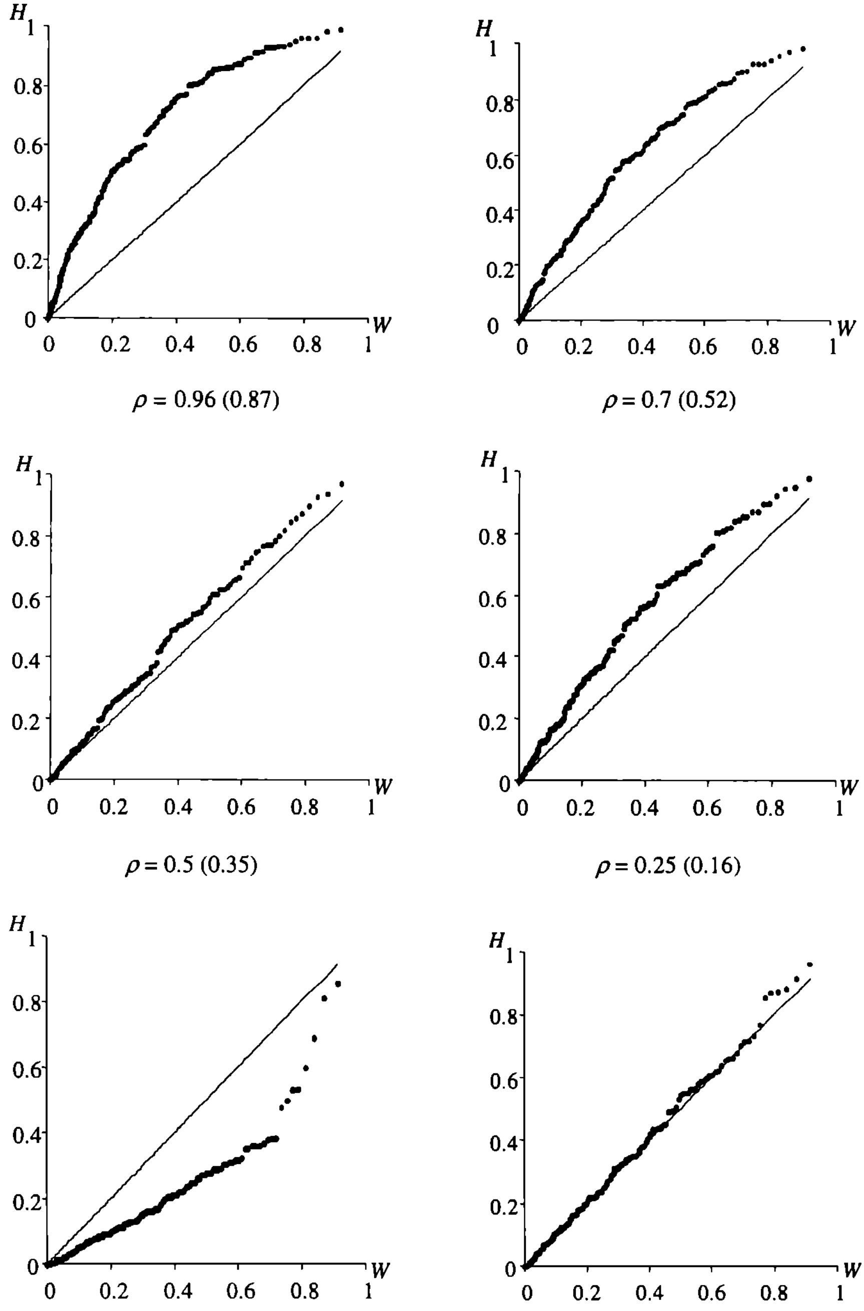
$\rho = 0.03\ (0.01)$

$\rho = -0.62\ (-0.46)$

$\rho = 1$

$\rho = -1$

Fig. 5. Chi-plots of the simulated data: Frank copula, normal margins, different ρ



$\rho = 0.96\ (0.87)$

$\rho = 0.7\ (0.52)$

$\rho = 0.5\ (0.35)$

$\rho = 0.25\ (0.16)$

$\rho = 0.03$

$\rho = -0.62\ (-0.46)$

Fig. 6. Kendall plots of the simulated data: Frank copula, normal margins, different ρ

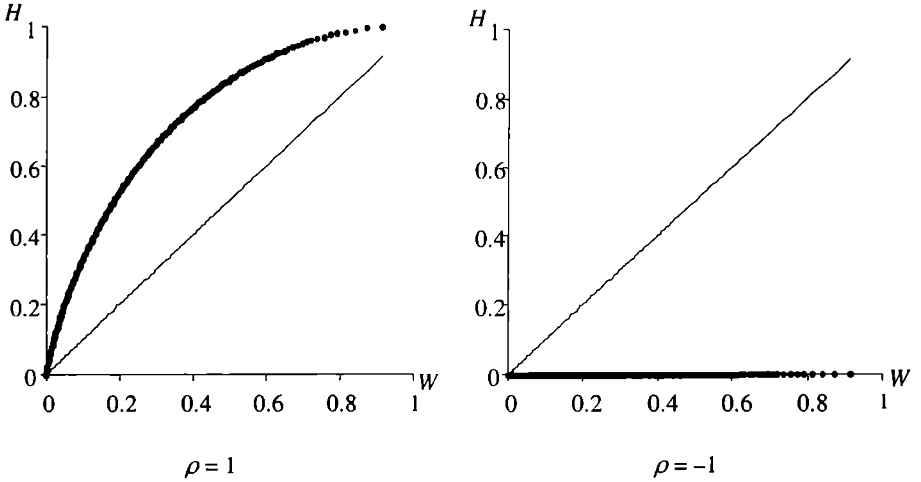
This distribution function depends only on the copula associated with dependence between random variables X and Y [4]. We obtain

$$K(w) = P(C(F(X),G(Y)) \leqslant w).$$

The empirical distribution function K_n of the pseudo-observations $H_1, \ldots, H_n$ is an asymptotically Gaussian, consistent estimator of $K(w)$ [4]. When the random variables X and Y are independent, then $K = K_0$. The function K determines the Kendall coefficient, because we have [4]

$$\tau = 3 - 4\int_0^1 K(w)dw.$$

The distribution function K lets us determine the asymptotical graph of Kendall plot. For large n, the points $(W_{i:n}, H_{(i)})$ lie approximately on the curve $(w, K^{-1}(K_0(w)))$. When the random variables X and Y are the commonotonic, i.e. perfect dependent: $Y = G^{-1}(F(X))$ or $Y = G^{-1}(1 - F(X))$, then $K(w) = w$ for $\tau = 1$ and $K(w) = 1$ for $\tau = -1$. We obtain the following curves in these cases: $(w, w - w\ln(w))$ for $\tau = 1$ and $(w, 0)$ for $\tau = -1$ [4] (see Fig. 7).



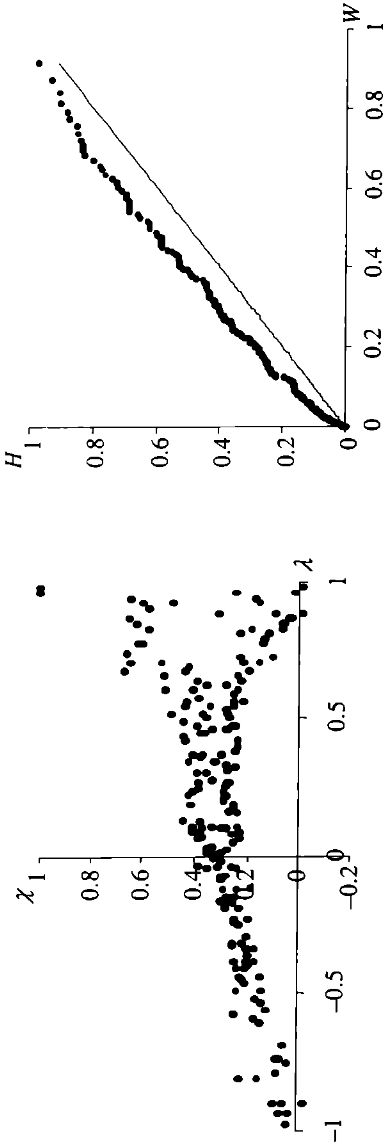
$\rho = 1$

$\rho = -1$

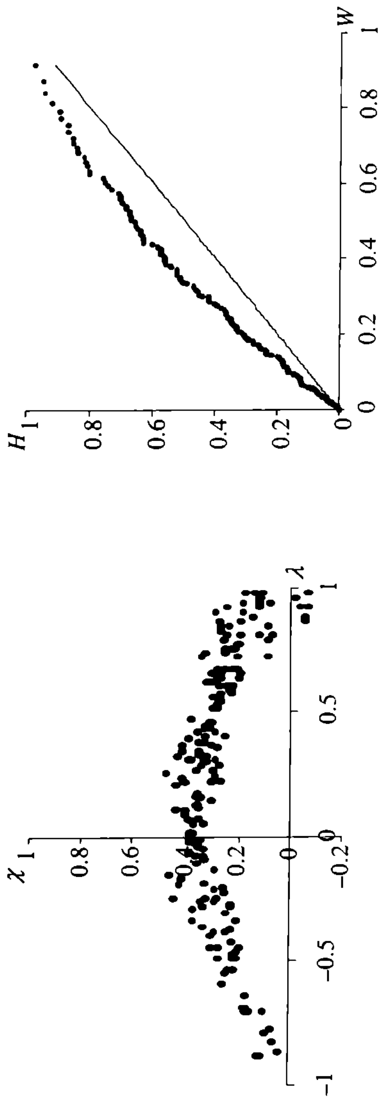
Fig. 7. Kendall plots of the commonotonic simulated data

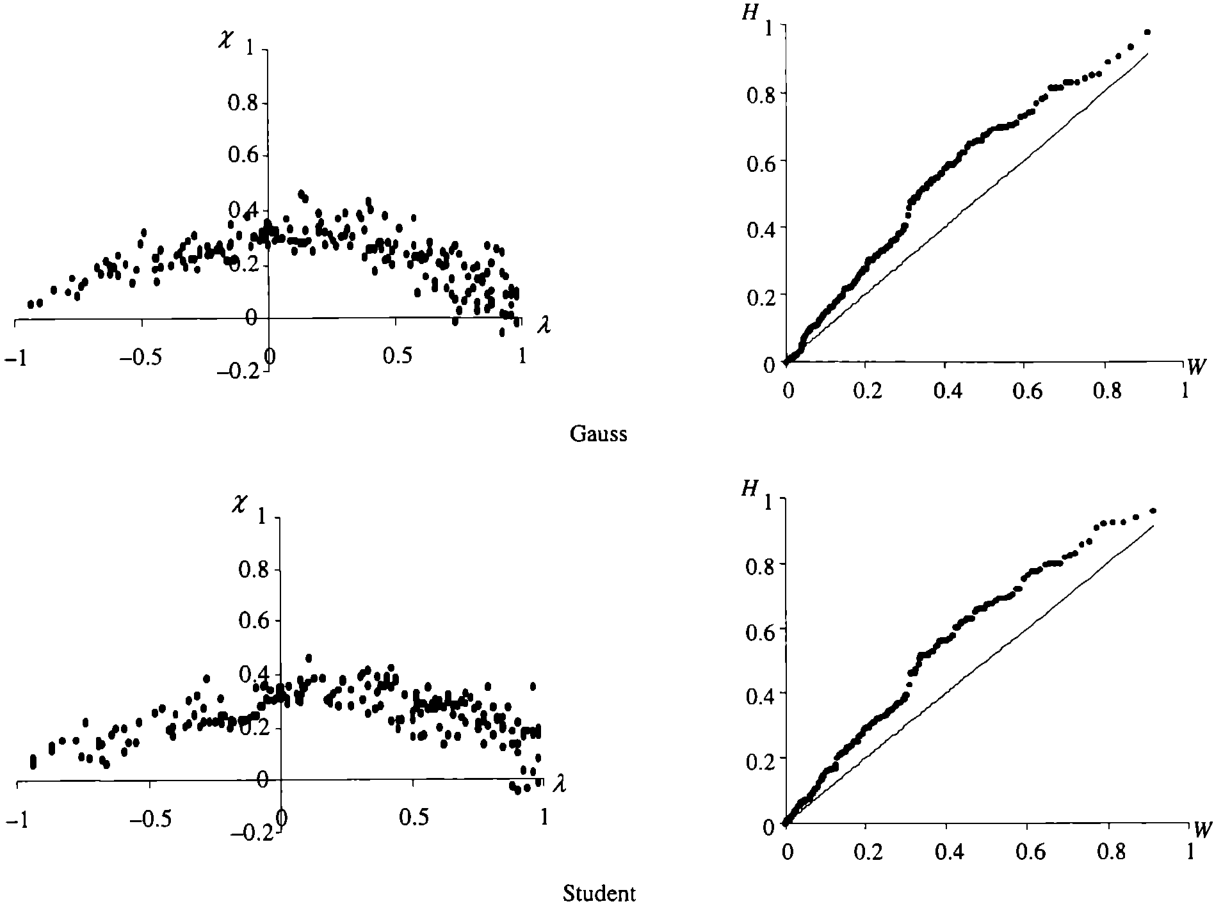
Now we investigate the case of p-dimensional data. For the multi-dimensional data, when dimension $p > 2$, Fisher and Switzer recommended to use the composition plot [2; 3]. This is the composition connected with the rank scatterplot and the chi-plot for every pair of variables. We can also generalize the Kendall plot for the

Clayton



Frank





Gauss

Student

Fig. 8. Chi-plots and Kendall plots for simulated data: normal margins, $\rho = 0.4$, different copulas

multi-dimensional data using the following formula for distribution K_0 of H_i under the independence:

$$K_0(w) = w + w\sum_{k=1}^{p-1}\frac{1}{k!}\log^k\left(\frac{1}{w}\right).$$

Figure 4 presents the composition plot and Kendall plot of the real, 3-dimensional data describing the expenditure of food (F) and health (H) and the size of house (A) of 200 random Polish households in 1993 (RAD Project "Poverty and Targeting of Social Assistance in Eastern Europe and Former Soviet Union"). We see that chi-plot better than rank scatterplot distinguishes the small dependencies. The dependence between the food and health expenditure is greater than dependence between the food expenditure and size of house. We see also that the health expenditure and size of house are almost independent.

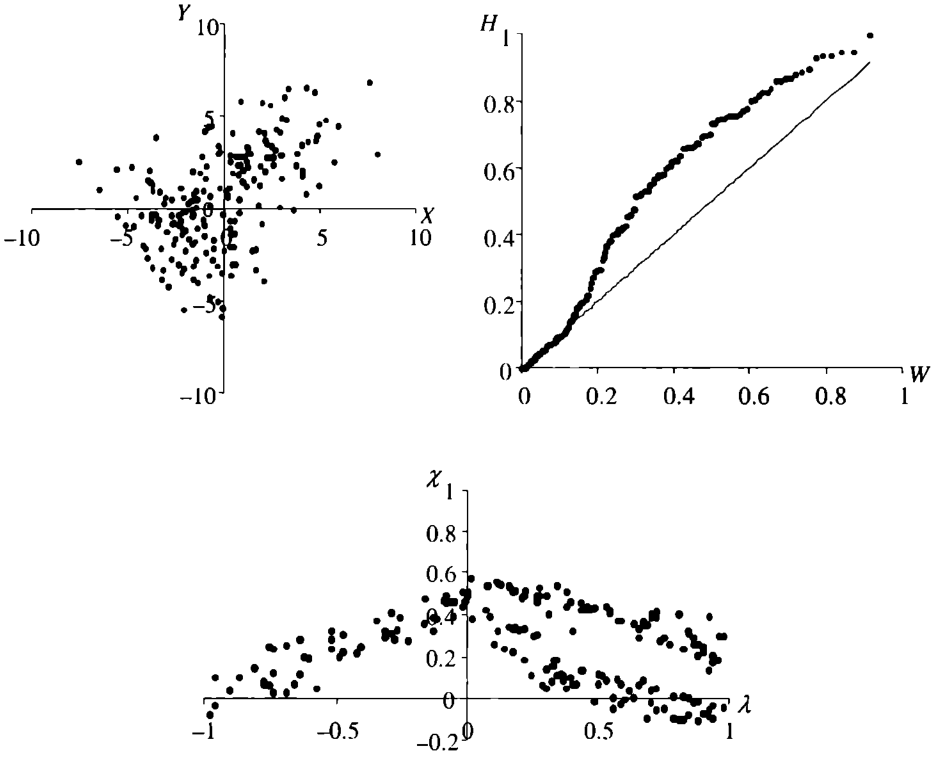


Fig. 9. Scatterplot, Kendall plot and chi-plot for mixture

The chi-plots and Kendall plots depend on the copulas. Figure 8 shows these plots for the datasets, when the dependences are described by the different copulas: Frank, Clayton, Gauss and *t*-Student and by the same values of the correlation coefficient ($\tau = 0.26$). We see that the simpler method – Kendall plot – better distinguishes the copulas.

Now we investigate the mixture of two simulated datasets. Let $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ denote the bivariate normal distribution with means μ_1, μ_2, standard deviations σ_1, σ_2 and correlation coefficient ρ. First 100 pairs come from the $N(-2, -2, 2, 2, -0.4)$ and second 100 pairs from $N(2, 2, 2, 2, 0.4)$. Figure 9 presents scatterplot, chi-plot and Kendall plot of this dataset. We see the two path of the points in the right side of the chi-plot. This pattern can signalize the existence of the mixture of populations. The scatterplot does not indicate the mixture.

References

[1] Embrechts P., Lindskog F., McNeil A., *Modelling Dependence with Copulas and Applications to Risk Management*, ETH Zürich, preprint, 2001.

[2] Fisher N.I., Switzer P., "Chi-plot for Assessing Dependence", *Biometrika* 1985, 72, 253-65.

[3] Fisher N.I., Switzer P., "Graphical Assessment of Dependence: Is a Picture Worth 100 Tests?", *The American Statistician* 2001, 55, 233-239.

[4] Genest C., Boies J.-C., "Detecting Dependence With Kendall Plots", *The American Statistician* 2003, 57, 275-284.

[5] Nelsen R.B., *An Introduction to Copulas*, Springer, New York 1999.

[6] Schweitzer B., Sklar A., *Probabilistic Metric Spaces*, North-Holland, New York 1981.

GRAFICZNE METODY BADANIA ZALEŻNOŚCI

Streszczenie

Artykuł jest poświęcony badaniu zależności za pomocą metod graficznych. Jest podzielony na dwie części: w pierwszej jest przedstawiona graficzna ilustracja pewnych rodzajów zależności, które są modelowane za pomocą funkcji *copula*. W przykładach badana jest głównie zależność w ogonach rozkładów.

Wykresy *chi* i Kendalla są przedstawione w drugiej części artykułu. Są one narzędziami graficznymi opartymi na rangach, które potrafią wykrywać istnienie zależności w próbie losowej z pewnego dwuwymiarowego rozkładu ciągłego. Wykres *chi* jest wykresem rozrzutu punktów (λ_i, χ_i), gdzie λ_i jest odległością punktu danych (x_i, y_i) od środka zbioru danych, a

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}},$$

gdzie $H_i = \#\{j \neq i: X_i \leqslant X_j, Y_i \leqslant Y_j\}/(n-1)$, $F_i = \#\{j \neq i: X_i \leqslant X_j\}/(n-1)$ i $G_i = \#\{j \neq i: Y_i \leqslant Y_j\}/(n-1)$.

Wykres Kendalla jest zwykłym wykresem rozrzutu, składającym się z par $(W_{i:n}, H_{(i)})$, gdzie $W_{i:n} = E(H_{(i)})$.