Article citation info:
HRYNIEWICZ O, KARPIŃSKI J. Prediction of reliability – the pitfalls of using Pearson's correlation. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2014; 16 (3): 472–483.

Olgierd HRYNIEWICZ
Janusz KARPIŃSKI

PREDICTION OF RELIABILITY – THE PITFALLS OF USING PEARSON'S CORRELATION

PROGNOZOWANIE NIEZAWODNOŚCI – PUŁAPKI ZWIĄZANE Z UŻYWANIEM WSPÓŁCZYNNIKA KORELACJI PEARSONA

Pearson's coefficient of linear correlation r is the measure of dependence which is the most popular among practitioners. In the paper we have shown, using comprehensive computer simulations, that its application is very limited when we search for informative variables that can be used for the prediction of reliability. We have shown that Kendall's coefficient of association τ is much better for this purpose.

Keywords: *measures of dependence, Kendall's τ, Pearson's r, Spearman's ρ, prediction of reliability.*

Współczynnik korelacji liniowej r Pearsona jest najbardziej popularną wśród praktyków miarą zależności statystycznej. W artykule na podstawie wyników wyczerpujących symulacji komputerowych pokazano, że w przypadku poszukiwania zmiennych mogących służyć do prognozowania niezawodności zakres jego stosowalności jest bardzo ograniczony. Wyniki badań symulacyjnych pokazują, że temu celowi lepiej służy współczynnik asocjacji τ Kendalla.

Słowa kluczowe: *miary zależności, współczynnik τ Kendalla, współczynnik ρ korelacji r Pearsona, współczynnik korelacji ρ Spearmana, prognozowanie niezawodności.*

1. Introduction

Statistical regression models are widely used in the analysis of reliability data. In the recent overview paper by Elsayed [2] these methods have been indicated as the principal tools in areas such as reliability prediction and accelerated life tests. In a similar overview dedicated to the problem of warranty data analysis Wu [18] gives examples of the applications of regression methods in this area. As statistical data coming from long-lasting life tests are seldom available, many attempts have been made to build mathematical models for the prediction of reliability basing on easily observed (or measured) characteristics. For example, prediction models presented in the Military Handbook MIL-217F [15] link the most popular reliability characteristic, namely the hazard rate λ, with many factors describing the object itself, the condition of its usage, etc. These models are based on the statistical analysis of large sets of reliability data collected over years by organizations such as the U.S. Army. The mathematical models that are used for prediction purposes in MIL-217F and other similar documents are usually obtained using classical regression methods. Consider, for example, the prediction of the base hazard rate of a travelling wave tube. In the Notice 2 of the Military Handbook MIL-217F [16] the following formula is given for the calculation of the basic failure rate of such device $\lambda_b = 11 \cdot (1{,}00001)^P \cdot (1{,}1)^F$, where F is the operating frequency (in GHz), and P is the rated power (in Watts). When we take logarithms of both sides of this formula we arrive at a classical linear regression model that links the basic reliability characteristic with the parameters of the considered device. The parameters of the models presented in [15] and [16] are somewhat obsolete because they were computed using data collected more than twenty years ago. However, the general formulae used in MIL 217F for the prediction purposes are still used (see, e.g., the recent papers by Lee and Lee [9] or by Thaduri *et al.* [14]).

The second important area of the theory and practice of reliability in which regression models are widely used is accelerated life testing. The two most important classes of models used for the description of the accelerated life tests, namely the accelerated failure time models (AFT) and the proportional hazard models (PH), belong to the class of regression models (see [2] for a short overview). Regression models are also used in other areas of reliability and risk analysis. For example, Schneidewind [12] proposed a regression model for the prediction of risk in software engineering.

In order to build prediction models it is necessary to evaluate the strength of statistical dependence between the characteristic of interest and its best predictors. It is obvious that the values of good predictors should be strongly associated with the values of the characteristic of interest. In mathematical statistics many measures of statistical dependence exist, but Pearson's coefficient of correlation r is the most popular among practitioners. The reason of this stems from the fact that in nearly all popular software tools, such as spreadsheets or basic versions of statistical packages, Pearson's coefficient of correlation r is the main measure used for the evaluation of regression models.

Pearson's coefficient of correlation r (usually called simply "the correlation") measures the strength of *linear correlation* between random variables. In all statistical textbooks, readers are warned against the usage of this measure of dependence when the dependence between random variables is nonlinear. For example, in the case of two random variables X and $Y=X^2$ defined on the whole space of real numbers, their linear correlation coefficient will be equal to zero despite the strongest possible (deterministic) relation. In practice however, one cannot easily recognize to what extent random variables are linearly dependent, even if the type of their bivariate probability distribution is known. It is well known from the theory of mathematical statistics that such linear dependence exists when the random variables are jointly distributed according to the multivariate normal

(Gaussian) distribution. When the assumption about the multivariate normality is not fulfilled one needs to use other measures of statistical dependence, such as Kendall's coefficient of association τ or Spearman's coefficient of rank correlation ρ. Unfortunately, a general theory that explains the links between Pearson's coefficient of correlation r and nonparametric measures of dependence, such as Kendall's τ or Spearman's ρ does not exist. Therefore, the relationship between these measures of dependence is usually investigated in particular context. For example, Xu *et. al.* [19] consider the problem of the measurement of correlation in signal processing when measurements are described by contaminated normal models. A very interesting analysis is presented in the paper by Vořechovský [17] who considered the problem of the Monte Carlo simulation of interdependent random vectors.

Regression models can be built for practically all types of statistical data. However, their statistical properties as calculated by popular software or described in the majority of statistical textbooks are valid only for the data described by the normal distribution. When lifetime data are analyzed this assumption is fulfilled only in very few practical cases, as lifetimes are seldom distributed according to the normal distribution. The situation is even worse when we build a regression model for the prediction of the hazard rate λ. In this case, the probability distribution of the predicted variable is *never* distributed according to the normal distribution. Probability distributions encountered in reliability testing, such as the exponential, Weibull, gamma or log-normal distributions, are skewed, and the multivariate (bivariate in practice) normal (Gaussian) distribution *should not* be used for the modeling of statistical dependence between the characteristic of interest and its predictors. Therefore, there is a need to investigate the behavior of Pearson's correlation coefficient r when the underlying models of dependence are applicable in the context of reliability prediction. This is the main aim of this paper.

The paper has the following structure. In its second section we recall some basic information about the methods for measuring the dependence between random variables. The main aim of this section is to highlight important restrictions for the usage of the coefficient of linear correlation. The third section of the paper is devoted to the analysis of the relations between the values of the coefficient of linear correlation and the values of other popular measures of statistical dependence, such as Kendall's coefficient of associationτ or Spearman's coefficient of rank correlation ρ. Approximate formulae, based on the results of extensive Monte Carlo computer simulation experiments, which link the values of r with the values of other measures of dependence are presented in the fourth section.

2. Measuring of dependence between random variables

Let X and Y be random variables whose joint probability distribution is $H(x,y)$. In this paper we assume that these variables have continuous marginal distributions $F(x)$ and $G(y)$ with finite expected values $E(X)$, $E(Y)$, and variances $V(X)$, $V(Y)$, respectively. Many such distributions have been proposed over the last one hundred years. Sklar [13] published his famous theorem which says that any two-dimensional probability distribution function $H(x,y)$ with marginal distributions $F(x)$ and $G(y)$ is represented using a function C, called a *copula,* in the following way:

$$H(x,y) = C\left(F(x), G(y)\right) \tag{1}$$

for all $x,y \in R$.

Any function defined on a square unit $[0,1]\times[0,1]$ and such that:

$$C(0,x) = C(x,0) = 0,$$

$$C(1,x) = C(x,1) = 1,\ x \in [0,1],\ \text{and}$$

$$C(b,d) - C(a,d) - C(b,c) + C(a,c) \geq 0,\ a,b,c,d \in [0,1],\ a \leq b,\ c \leq d$$

is a copula. Conversely, for any distribution functions F and G and any copula C, the function H defined by (1) is a two-dimensional distribution function with marginals F and G. Moreover, if F and G are continuous, then the copula C is unique.

Let $u=F(x)$, and $v=G(y)$. The simplest copula, the product copula $\Pi(u,v) = uv$, describes *independent* random variables. All other bivariate copulas fulfill the Fréchet-Hoeffding inequalities:

$$W(u,v) = \max(u+v-1,0) \leq C(u,v) \leq \min(u,v) = M(u,v) \tag{2}$$

The left inequality in (2) describes the case of full negative dependence between X and Y, and the right inequality in this formula describes the case of full positive dependence between X and Y.

Sklar's theorem has been generalized to the p-dimensional case, so it is applicable for any p-dimensional probability distribution. Similarly, the Fréchet-Hoeffding inequalities have been also generalized for the p-dimensional case. However, in this more general setting all mathematical formulae describing multidimensional probability distributions become very complicated, and thus have limited usage for practitioners. Therefore, in this paper we restrict ourselves only to the two-dimensional (bivariate) case.

The most popular measure of dependence between two random variables is based on the concept of the *covariance* defined for real valued random variables as:

$$Cov(X,Y) = \iint_{S_{xy}} \left(x - E(X)\right)\left(y - E(Y)\right) f(x,y)\,dxdy \tag{3}$$

where S_{xy} is the area for which the bivariate probability density function $f(x,y)$ is positive. When we divide the covariance by the product of the standard deviations $\sigma(X)$, and $\sigma(Y)$ of X and Y we arrive at the famous Pearson's coefficient of linear correlation:

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{4}$$

described in every textbook on probability and statistics.

Let $(x_i, y_i),\ i = 1,\ldots,n$ be the observed sample of n independent pairs of observations of the random vector (X,Y). The sample version of Pearson's coefficient of linear correlation is given by the well known formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{5}$$

It is a well known that $r(X,Y)$ describes only *linear* dependence between random variables, and thus should not be used for many bivariate probability distributions as the measure of dependence. For example, if X and Y are independent, than $r(X,Y)=0$, but the converse is not true. There exist many examples of highly dependent data for whom we observe no linear correlation ($r(X,Y)$ is equal or very close to zero). It has been proven that Pearson's coefficient of correlation fully describes the dependence structure only in the case of the bivari-

ate normal (Gaussian) distribution. This distribution is the special case (for normal marginal distribution) of the normal copula defined as:

$$C_N(u,v) = \Phi_N\left(\Phi^{-1}(u), \Phi^{-1}(v); r\right) \tag{6}$$

where $\Phi_N(x,y;r)$ is the cumulative distribution function of the bivariate standardized normal distribution with the correlation coefficient r, and $\Phi^{-1}(x)$ is the inverse of the cdf of the univariate standardized normal distribution (the quantile function). Pearson's r may be also used as a measure of dependence for random variables that are jointly elliptically distributed. To this class of probability distributions belong the aforementioned multivariate Gaussian distribution, the multivariate t-distribution, and other distributions whose multivariate characteristic function can be represented as a certain quadratic form. However, even in the case of the elliptical distributions Pearson's r has meaning only for the distributions with finite variances.

Another popular measure of dependence is Spearman's coefficient of rank correlation. Let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ and $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ be the ordered elements of (x_i, y_i), $i = 1,\ldots,n$, and let $R_1 \le R_2 \le \cdots \le R_n$ and $S_1 \le S_2 \le \cdots \le S_n$ be the *ranks* of the original observations $x_1,\ldots x_n$ and $y_1,\ldots y_n$ in this ordering. Spearman's coefficient of rank correlation is the coefficient of linear correlation calculated for these ranks, and is given by the formula:

$$\rho_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \tag{7}$$

where:

$$d_i = R_i - S_i, i = 1,\ldots,n. \tag{8}$$

It has been proved, see Nelsen [10], that the population version of Spearman's ρ can be found for any copula using the following formula:

$$\rho(X,Y) = 12\iint_{[0,1]^2} C(u,v)dudv - 3. \tag{9}$$

Kendall's rank correlation coefficient, known as Kendall's τ, was proposed in 1938, and is based on the concept of concordant and disconcordant pairs of observations. A pair of vector observations (x_i, y_i), and (x_j, y_j) of continuous random variables (X,Y) is *concordant* if the respective ranks of the elements of both vectors agree, i.e either $R_i > R_j$ and $S_i > S_j$ or $R_i < R_j$ and $S_i < S_j$. Otherwise, this pair is *disconcordant*. The sample version of Kendall's τ is defined as:

$$\tau_{xy} = 2\frac{\text{no. of concordant pairs - no. of disconcordant pairs}}{n(n-1)}. \tag{10}$$

A convenient representation of τ has been proposed by Genest and Rivest [7] in the following form:

$$\tau_{xy} = \frac{4}{n}\sum_{i=1}^{n} V_i - 1, \tag{11}$$

where:

$$V_i = card\left\{\left(X_j, Y_j\right): X_j < X_i, Y_j < Y_i\right\} \Big/ (n-1), i = 1,\ldots,n. \tag{12}$$

The population version of Kendall's τ can be found, see Nelsen [10], for any copula using the following formula:

$$\tau(X,Y) = 4\iint_{[0,1]^2} C(u,v)dC(u,v) - 1. \tag{13}$$

Many other measures of dependence exist, described for instance in the book by Nelsen [10] or in the paper by Embrechts *et al*. [3]. Some of these measures are called the measures of *concordance*. Scarsini [11] defines a measure of concordance as a real valued measure of dependence κ between two continuous random variables X and Y whose copula C satisfies the following properties:

1. κ is defined for every pair X; Y of continuous random variables.
2. $-1 \le \kappa_{X,Y} \le 1$, $\kappa_{X,X} = 1$ and $\kappa_{X,-X} = -1$.
3. $\kappa_{X,Y} = \kappa_{Y,X}$.
4. If X and Y are independent, then $\kappa_{X,Y} = 0$.
5. $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$.
6. If C and $\tilde{C}$ are copulas such that $C \le \tilde{C}$, then $\kappa_C \le \kappa_{\tilde{C}}$.
7. If $\{(X_n; Y_n)\}$ is a sequence of continuous random variables with copulas C_n, and if $\{C_n\}$ converges pointwise to C, then $\lim_{n\to\infty} \kappa_{C_n} = \kappa_C$.

Spearmans ρ and Kendall's τ are measures of concordance (the proof can be found in the book by Nelsen [10]), but Pearson's r is not (as it is shown in the paper by Embrechts et al. [3]). It does not fulfill the condition 2., and the range of possible values of r depends upon the type of marginal distributions of dependent random variables X and Y. Below, we show some important properties of Pearson's r regarding this property.

Let us consider two continuous random variables X and Y described by the probability density functions $f(x)$ and $g(y)$, respectively. Without loss of generalization let us assume that $E(X) = E(Y) = E$, and $Var(X) = Var(Y) = 1$. Because Pearson's r is invariant with respect to linear transformations, transforming the original random variables to the variables defined above does not change the value of r which in this case is equal to the covariance between X and Y.

Now, let us consider the two limiting cases defined by (2). In the case of full *negative* dependence random variables X and Y are linked functionally in the following way:

$$F(x) = 1 - G(y). \tag{14}$$

where $F(x)$ and $G(x)$ are the respective cumulative probability functions of the random variables X and Y. Hence, the covariance between X and Y is given by:

$$Cov_{neg}(X,Y) = \int_{-\infty}^{\infty}(x-E)\left(\left\{G^{-1}\left[1-F(x)\right]\right\} - E\right)f(x)dx \tag{15}$$

where $G^{-}(x)$ is the inverse (the quantile function) of $G(x)$.

In the case of full *positive* dependence the link is of the form:

$$F(x) = G(y), \tag{16}$$

and a similar formula is given by:

$$Cov_{pos}(X,Y) = \int_{-\infty}^{\infty}(x-E)\left(\left\{G^{-1}\left[F(x)\right]\right\} - E\right)f(x)dx \tag{17}$$

The formulae (15) and (17) can be used for the calculation of the limiting values, r_{min} and r_{max}, of Pearson's r. From the analysis of these formulae we can derive the following properties of Pearson's r.

Property 1: When the probability distributions of X and Y have the same shape, then $r_{max}=1$.

Proof: The proof of this property if straightforward. The same shape of two probability distributions means that after appropriate transformations of scale and location we have $F(x)=G(y)$. Hence, $G^{-1}\left[F(x)\right]=x$ and $Cov_{pos}(X,Y)=Var(X)$, and thus $r(X,Y)=r_{\max}=1$.

Property 2: When probability distributions of X and Y are symmetric around zero (E=0) and have the same shape, then $r_{\min}=-1$.

Proof: For symmetric distributions, with E=0, we have $G^{-1}(-x)=-G^{-1}(x)$. Thus, for distributions with the same shape we $G^{-1}\left(1-F(x)\right)=-G^{-1}\left(F(x)\right)=-x$. Then, we have $Cov_{neg}(X,Y)=-Var(X)$, and thus $r(X,Y)=r_{\min}=-1$.

Property 3: When at least one of the random variables has a symmetric distribution, then $r_{\min}=-r_{\max}$.

Proof: Let Y be the random variable with a symmetric distribution, then $G^{-1}\left(1-F(x)\right)=-G^{-1}\left(F(x)\right)$, Hence, we have $Cov_{pos}(X,Y)=-Cov_{neg}(X,Y)$, and consequently $r_{\min}=-r_{\max}$.

With the exception of cases when Properties 1 and 2 hold, the calculation of $r_{\min}$ and $r_{\max}$ is usually difficult.

Example 1

Consider the case when both X and Y have the same exponential distribution with E=1. Because the variance in the exponential distribution is the same as the expected value we have in the considered case $r(X,Y)=Cov(X,Y)$. Then, the formula (15) takes the following form:

$$r_{\min}=Cov_{neg}(X,Y)=\int_0^{\infty}(x-1)\left(\left\{-\ln\left[1-e^{-x}\right]\right\}-1\right)e^{x}dx=1-\frac{\pi^2}{6}=-0,644934. \quad (18)$$

The integral in (18) has been evaluated using symbolic and numerical calculations provided by the mathematical package Mathematica™.

Example 2

Consider the case when X is distributed according to the exponential distribution with E=1, and Y is uniformly distributed over the interval [-0,5 , 0,5]. The maximal value of the $Cov(X,Y)$ is now:

$$Cov_{pos}(X,Y)=\int_0^{\infty}(x-1)\left(1-e^{-x}\right)e^{x}dx=\frac{1}{4}e^{-2x}\left[\left(4e^{-x}-2\right)x+1\right]\Big|_0^{\infty}=\frac{1}{4}. \quad (19)$$

Hence, $r_{\max}=\sqrt{3}/2=0,866$, and, by the Property 3, $r_{\min}=-0,866$.

When the random variables X and Y are distributed according to the reliability distributions such as the Weibull or the Log-normal, which are so popular in theory and in practice, the calculation of the minimal or maximal values of Pearson's r can be done only numerically or by simulations. However, the numerical integration can, in this case, be very difficult, as the integrated functions may adopt infinite values at zero. For this reason, the Monte Carlo simulations, described in the next section of this paper, seem to be a better way to find these values.

3. Properties of Pearson's r

It is well known that the values of Pearson's r depend upon the type of the marginal distributions of a bivariate random variable. In the previous section we have shown how the range of possible values of r depends upon the shape of these marginals. More questions, important from a practical point of view, could be asked. In this paper we will try to answer some of them, and namely:

a. How the values of r depend upon the type of marginal distributions in the case of distributions used in reliability practice?
b. Do the properties of r depend upon the type of dependence described by some popular copulas?
c. What is the relationship between the values of r and the values of other measures of dependence, such as Kendall's τ or Spearman's ρ?
d. What is the accuracy of the estimation of different measures of dependence?

For these, and many other similar questions, the answers cannot be found using analytical methods. Therefore, we have performed extensive computer simulations and analyzed samples of different size, generated from different copulas with different marginal distributions.

We have considered four types of copulas. The first one, the normal (Gaussian) copula have been already introduced, and defined by (6). The remaining three copulas belong to the family of the Archimedean copulas defined by Genest and McKay [6] in the following way:

$$C(u,v)=\varphi^{-1}\left(\varphi(u)+\varphi(v)\right), \quad (20)$$

where φ^{-1} is a pseudo-inverse of the continuous and strictly decreasing function $\varphi:[0,1]\to[0,\infty]$, called copula's generator, such that $\varphi(1)=0$. From this family we have taken the following three well known copulas:

a. Clayton copula (Clayton [1]), defined as:

$$C(x,y)=\left[F^{-\theta}(x)+G^{-\theta}-1\right]^{-1/\theta},\ \theta\in\left\{(-1,\infty)\setminus\{0\}\right\}, \quad (21)$$

b. Frank copula (Frank [4]), defined as:

$$C(x,y)=-\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta F(x)}-1\right)\left(e^{-\theta G(y)}-1\right)}{e^{-\theta}-1}\right),\ \theta\in\left\{(-\infty,\infty)\setminus\{0\}\right\}, \quad (22)$$

c. Gumbel copula (Gumbel [8]), defined as:

$$C(x,y)=\exp\left(-\left[\left(-\ln F(x)\right)^{\theta}+\left(-\ln G(y)\right)^{\theta}\right]^{1/\theta}\right),\ \theta>0. \quad (23)$$

One of the reasons for using these particular copulas is the relative ease of the computer simulation of samples from these copulas for the given strength of dependence defined by Kendall's τ. For the normal

copula, popular algorithms can be used for this purpose for the simulation of samples from a classical bivariate normal distribution, and for the remaining three copulas we used a general algorithm proposed by Genest and McKay [6] for the Archimedean copulas.

For the measure of dependence in the simulated samples we use Kendall's τ. For this measure of dependence there exist formulae that link the value of τ with the parameters of the copulas. These links depend only on the type of copula, and because of a non-parametric character of Kendall's τ do not depend upon the type of marginals. For the normal (Gaussian) copula the following relation holds:

$$\tau = \arcsin(r)/(\pi/2). \tag{24}$$

For the chosen Archimedean copulas we have the following formulae:

a. Clayton copula

$$\tau = \frac{\theta}{\theta+2}, \tag{25}$$

b. Frank copula

$$\tau = 1 + 4\left(\frac{1}{\theta}\int_0^\theta \frac{t}{e^t-1}dt - 1\right)\Big/\theta, \tag{26}$$

c. Gumbel copula

$$\tau = \frac{\theta}{\theta+1}. \tag{27}$$

We see that except for the case of Frank copula, when we have to solve for θ a very complicated equation, the dependence parameter of a given copula is straightforwardly related to the value of Kendall's τ. Such simple relationships do not exist for Spearman's ρ, so we have chosen Kendall's τ as the measure of dependence in the simulated samples.

In order to investigate the influence of the type of the marginal distribution on the value of Pearson's r we considered two cases. In the first one we assumed that both variables X and Y have the same marginal distribution: normal, exponential and Weibull (with different parameters of shape δ). In the second case, that seems to be more appropriate as regards problems of reliability prediction, we have assumed that the predictor X has the normal distribution, and Y is distributed according to different Weibull distributions (the exponential distribution included).

The properties of the considered statistics depend on the sample size n. In our simulation experiments we considered three values of n: n=500, which allow the approximation of the values of the population (theoretical) versions of the measures of dependence, n=100, which represents the case of a relatively accurate estimation of this measure, and n=20, which represents the sample size more appropriate for the analysis of reliability.

We have simulated 1000000 samples in each of the simulation experiments. Therefore, the results of the experiment are very accurate, and the impact of the randomness of the Monte Carlo methodology can be neglected.

The results of experiments have been summarized in respective tables. In this paper we present only few of them, showing the results only for some chosen values of Kendall's τ. Table 1 represents the results of one of the simulation experiments where the Clayton copula with given marginals, normal N(0,1) for X, and Weibull W(1,5) for Y, was used as the mathematical model. In this experiment samples of n=100 elements were generated for 22 different values of τ, and for each value of τ the respective value or r was estimated from the results of simulation. The consecutive columns of this table represent: the assumed value of Kendall's τ, the estimated mean value of Kendall's τ, the estimated mean value of Spearman's ρ, the estimated mean value of Pearson's r, the estimated standard deviation of Kendall's τ, the estimated standard deviation of Spearman's ρ, and the estimated standard deviation of Pearson's r, respectively.

In Table 2 we present the results of the simulation experiment when dependence is described by the normal (Gaussian) copula. Note, that in this case the random vector (X,Y) does not have a bivariate normal distribution, as its second component (Y) is distributed according to the Weibull distribution with the shape parameter δ=1,5.

Table 1. X – N(0,1), Y – Weibull (1,5), Clayton copula, n=100

TAU	TAU-EST	RHO-SP	R-PEARS	SIG-TAU	SIG-RHO	SIG-R
1	1	1	0,966172	0	0	0,0084
0,9	0,899987	0,981913	0,919786	0,015427	0,005713	0,022431
0,7	0,699994	0,868551	0,780478	0,038352	0,032555	0,043463
0,5	0,500009	0,676824	0,602098	0,054615	0,063348	0,065445
0,3	0,300006	0,430006	0,386769	0,064561	0,087601	0,085434
0,1	0,099982	0,147826	0,135595	0,0683	0,099877	0,098458
0	0,000009	0,000018	0,000000	0,067926	0,100639	0,100626
-0,1	-0,10004	-0,14776	-0,13675	0,066267	0,097949	0,099825
-0,3	-0,30002	-0,42179	-0,38989	0,061608	0,086718	0,094286
-0,5	-0,5	-0,64362	-0,59994	0,057632	0,072373	0,084198
-0,7	-0,69998	-0,81401	-0,76949	0,051404	0,054946	0,0679
-0,9	-0,89995	-0,94535	-0,907	0,03438	0,030933	0,040685
-1	-1	-1	-0,966157	0	0	0,008387

Table 2. X – N(0,1), Y – Weibull (1,5), Normal copula, n=100

TAU	TAU-EST	RHO-SP	R-PEARS	SIG-TAU	SIG-RHO	SIG-R
1	1	1	0,966172	0	0	0,0084
0,9	0,899992	0,983882	0,954148	0,012531	0,004014	0,008852
0,7	0,699955	0,876410	0,859881	0,032894	0,026777	0,022927
0,5	0,499934	0,684427	0,681280	0,049546	0,057634	0,051623
0,3	0,299910	0,433129	0,436664	0,061163	0,083980	0,080477
0,1	0,099936	0,147973	0,150245	0,067111	0,098637	0,098170
0	0,000009	0,000018	0,000000	0,067926	0,100639	0,100626
-0,1	-0,100051	-0,148130	-0,150423	0,067128	0,098671	0,098221
-0,3	-0,300033	-0,433290	-0,436767	0,061215	0,084050	0,080566
-0,5	-0,500011	-0,684503	-0,681312	0,049594	0,057687	0,051656
-0,7	-0,700006	-0,876442	-0,859885	0,032900	0,026783	0,022923
-0,9	-0,899994	-0,983882	-0,954148	0,012528	0,004013	0,008851
-1	-1	-1	-0,966157	0	0	0,008387

Table 3. X – N(0,1), Y – Weibull (2,0), Clayton copula, n=100

TAU	TAU-EST	RHO-SP	R-PEARS	SIG-TAU	SIG-RHO	SIG-R
1	1	1	0,986855	0	0	0,004596
0,9	0,899987	0,981913	0,947121	0,015431	0,005713	0,015718
0,7	0,699994	0,868551	0,816918	0,038351	0,032555	0,038158
0,5	0,500009	0,676824	0,639403	0,054615	0,063348	0,062774
0,3	0,300006	0,430006	0,415361	0,064561	0,087601	0,085563
0,1	0,099982	0,147826	0,146231	0,0683	0,099877	0,099399
0	0,000009	0,000018	-0,000009	0,067926	0,100639	0,100629
-0,1	-0,10004	-0,14776	-0,14625	0,066267	0,097949	0,098495
-0,3	-0,30002	-0,42179	-0,4108	0,061608	0,086718	0,090795
-0,5	-0,5	-0,64362	-0,62437	0,057632	0,072373	0,079514
-0,7	-0,69998	-0,81401	-0,79356	0,051405	0,054946	0,063075
-0,9	-0,89995	-0,94535	-0,929	0,034378	0,030933	0,037035
-1	-1	-1	-0,98685	0	0	0,004588

Table 4. X – N(0,1), Y – Weibull (0,5), Frank copula, n=100

TAU	TAU-EST	RHO-SP	R-PEARS	SIG-TAU	SIG-RHO	SIG-R
1	1	1	0,719517	0	0	0,050144
0,9	0,899983	0,984952	0,674247	0,010692	0,003174	0,079032
0,7	0,699959	0,881641	0,565369	0,029395	0,023669	0,084575
0,5	0,499958	0,688860	0,434077	0,047261	0,055943	0,086974
0,3	0,299965	0,434539	0,274331	0,060343	0,083666	0,093489
0,1	0,099955	0,148060	0,093864	0,067048	0,098664	0,099509
0	0,000009	0,000018	0,000017	0,067926	0,100639	0,100577
-0,1	-0,100040	-0,148174	-0,093847	0,067086	0,098728	0,099595
-0,3	-0,300027	-0,434610	-0,274362	0,060460	0,083831	0,093634
-0,5	-0,499991	-0,688871	-0,434205	0,047386	0,056102	0,087096
-0,7	-0,699970	-0,881636	-0,565504	0,029461	0,023723	0,084793
-0,9	-0,899982	-0,984951	-0,674239	0,010700	0,003183	0,079138
-1	-1	-1	-0,71946	0	0	0,050233

Table 5. X – N(0,1), Y – Exponential, Gumbel copula, n=100

TAU	TAU-EST	RHO-SP	R-PEARS	SIG-TAU	SIG-RHO	SIG-R
1	1	1	0,909318	0	0	0,017728
0,9	0,899655	0,98295	0,901015	0,013871	0,004802	0,017174
0,7	0,69967	0,870859	0,829142	0,036312	0,030757	0,026007
0,5	0,499642	0,676797	0,682615	0,053541	0,062524	0,054827
0,3	0,299587	0,428082	0,461994	0,064564	0,087684	0,090058
0,1	0,09965	0,14718	0,170754	0,068657	0,100141	0,109652
0	0,000009	0,000018	0,000011	0,067926	0,100639	0,100616

From the first two columns of Table 1 and Table 2 one can have the impression that the estimates of Kendall's τ obtained from the generated samples are unbiased. Their average values (over 1000000 simulation runs) are practically the same as their assumed values. This has been confirmed in all simulation experiments, also for small samples of n=20 elements. This could serve as the proof that the algorithms used for the generation of data from different copulas work correctly.

The comparison of the third and the fourth columns of Table 1 and Table 2 shows different relation between Kendall's τ , Spearman's ρ, and Pearson's r. This reflects the influence of the type of copula. In the case of the Clayton copula the values of ρ and r are not symmetric with respect to the case of independence, where all dependence measures should have the value of zero. However, for the normal copula this symmetry is visible. The same situation is observed, but with lower intensity, for the observed standard deviations of the considered measures of dependence.

In Table 3 we present the results of simulations from the Clayton copula, but with a different, as compared to the case presented in Table 1, marginal distribution of distribution of Y, namely the Weibull distribution with the shape parameter δ=2,0. The seed of the generator of random numbers was the same in all performed simulations, so it is possible to compare their results directly. The comparison of the second and the third columns of Table 1 and Table 3 shows that the average values of Kendall's τ and Spearman's ρ are, because of non-parametric character of these statistics, exactly the same. However, the values of Pearson's r are slightly different in the both cases. This confirms the well-known fact that the values of r depend upon the type of the marginal distributions. What seems to be important from a practical point of view is the observation that in the cases in which the marginal distributions are not very different with respect to their skewness the values of Pearson's r are not very different.

The properties of Pearson's r are completely different in the case presented in Table 4 where the data were generated from the Frank copula, and the marginal distribution of Y was highly skewed (the Weibull distribution with the shape parameter δ=0,5). The behavior of Spearman's ρ in comparison to the cases presented in Tables 1 – 3 was similar, and the differences observed could be neglected from a practical point of view. However, the behavior of Pearson's r is completely different. First of all, the absolute minimal and maximal values of r are much smaller than 1, as is the case in the bivariate normal distribution. Therefore, they may be completely misleading when this measure of dependence will be used for the analysis of strongly dependent data.

The most important difficulty with the usage of Pearson's r is its dependence upon the type of marginal distributions. One can ask, however, about the practical impact of the distributions used to the problem of reliability prediction on the range of possible values of Pearson's r. In order to investigate this problem we have assumed that the random variable X is distributed according to the normal distribution N(0,1), and the random variable Y, which in the context of reliability prediction describes the life-time, is distributed according to different Weibull distributions, the exponential distribution included. We have searched for the minimal and maximal possible values of r, defined by (15) and (17), respectively. These two values have been evaluated in the Monte Carlo experiments in which samples of 100, 500, and 1000 items each were simulated in 1000000 runs. The results of this experiment are presented in Table 6.

Table 6. Minimal and maximal values of Pearson's r. One distribution symmetric

Distrib. X	Distrib. Y	n=100		n=500		n=1000	
		r_{min}	r_{max}	r_{min}	r_{max}	r_{min}	r_{max}
N(0,1)	Weib(0,2)	-0,4407	0,4407	-0,3297	0,3298	-0,2949	0,2950
N(0,1)	Weib(0,5)	-0,7195	0,7195	-0,6864	0,6864	-0,6796	0,6796
N(0,1)	Exp	-0,9093	0,9093	-0,9045	0,9045	-0,9039	0,9039
N(0,1)	Weib(1,5)	-0,9662	0,9662	-0,9647	0,9647	-0,9646	0,9646
N(0,1)	Weib(2,0)	-0,9869	0,9869	-0,9863	0,9863	-0,9862	0,9862
N(0,1)	Weib(2,5)	0,9951	0,9951	-0,9949	0,9949	-0,9949	0,9949

In this experiment one of the variables has a symmetric distribution, so according to the Property 3 the absolute values of r_{min} and r_{max} are the same. This has been confirmed in our experiments. Moreover, it appears from Table 6 that the range of possible values of r differs from the range expected for good measures of dependence, namely [−1,1] only in cases of highly skewed distributions, such as the Weibull with the parameter of shape equal or smaller than 1 or the exponential distribution. However, in the case of distributions with the increasing hazard rate the range of the possible values of r is close to [−1,1]. This is not unexpected as with the increasing value of the parameter of shape the Weibull distribution tends to the normal distribution for whom Pearson's r is the proper measure of dependence.

The situation becomes different when both dependent random variables have skewed distributions. In Table 7 we present the results of simulation for several such distributions.

The results presented in Table 6 and Table 7 show beyond doubt that the evaluation of the strength of dependence using Pearson's r in the case of skewed distributions may be highly misleading. In extreme cases the absolute values of r may be very small even in the case of very strong dependence. Therefore, in such cases, Pearson's r cannot be used for finding characteristics that can be used as good predictors of life-times. It is extremely important when observed life-times come from highly accelerated life tests (HALT). In these tests early failures of "weak" elements are frequently observed with the consequence of observing highly skewed life-time distributions.

The results of the simulation experiments have shown another unwanted property of Pearson's r. The estimator of r seems to be highly biased even for large sample sizes. In the Table 6 and Table 7 we see this phenomenon for the extreme values of Pearson's r. However, in practice we are more interested in the analysis of this bias for smaller grades of dependence. In Table 8 we present the comparison of the estimated values of r for different copulas, different marginal distributions, and different sample sizes.

4. Approximate relationships between the values of different measures of dependence

We will estimate the unknown relationship between Pearson's r and Kendall's τ from the simulation data using a polynomial:

$$r_a(\tau) = \sum_{i=0}^{k} w_i \tau^i \,, \tag{28}$$

with additional conditions $r_a(-1) = L$, $r_a(0) = 0$, and $r_a(1) = U$. When we take k=6 after some simple algebra we obtain the following regression equation:

$$r_a(\tau) = \sum_{i=0}^{4} a_i f_i(\tau) \,, \tag{29}$$

where a0=1, and:

$$f_0(\tau) = \tau^5 \left[2U\tau + (U-L)(1-\tau) \right] / 2 \,, \tag{30}$$

$$f_1(\tau) = \tau \left(1 - \tau^4\right), \tag{31}$$

$$f_2(\tau) = \tau^2 \left(1 - \tau^4\right), \tag{32}$$

Table 7. Minimal and maximal values of Pearson's r. Both distributions asymmetric

Distrib. X	Distrib. Y	n=100		n=500		n=1000	
		r_{min}	r_{max}	r_{min}	r_{max}	r_{min}	r_{max}
Weib(0,2)	Weib(0,2)	-0,043138	1	-0,018837	1	-0,014069	1
Weib(0,2)	Weib(0,5)	-0,104550	0,876690	-0,063580	0,804192	-0,054084	0,771048
Weib(0,5)	Exp	-0,430280	0,924765	-0,393422	0,905902	-0,386876	0,901267
Exp	Weib(1,5)	-0,773163	0,983631	-0,762072	0,982289	-0,760580	0,982088
Exp	Weib(2,0)	-0,830921	0,960043	-0,821939	0,957323	-0,820714	0,956929
Weib(1,5)	Weib(2,5)	-0,938714	0,985191	-0,935726	0,984522	-0,935335	0,984433

Table 8. Expected values of the estimator of r

Copula	X	Y	τ	n=20	n=100	n=500	n=1000
Clayton	N(0,1)	Weib(0,5)	0,8	0,652802	0,55909	0,52084	0,513599
			0,5	0,414688	0,35028	0,32491	0,32020
			-0,5	-0,44616	-0,39468	-0,37105	-0,36645
			-0,8	-0,66799	-0,60190	-0,57096	-0,56473
Frank	N(0,1)	Weib(0,5)	0,8	0,71270	0,62183	0,58206	0,574378
			0,5	0,50179	0,43408	0,40502	0,399476
			-0,5	-0,50208	-0,43421	-0,40497	-0,39943
			-0,8	-0,71270	-0,62187	-0,58203	-0,57423
Gauss	Exp	Exp	0,8	0,93886	0,94148	0,94205	0,942104
			0,5	0,66083	0,66647	0,66775	0,667828
			-0,5	-0,53688	-0,50238	-0,49300	-0,49175
			-0,8	-0,69748	-0,63947	-0,62491	-0,62299

$$f_3(\tau) = \tau^3\left(1-\tau^2\right), \tag{33}$$

$$f_4(\tau) = \tau^4\left(1-\tau^2\right). \tag{34}$$

Coefficients a_1, a_2, a_3, and a_4 of (29) have been obtained for different copulas, and different marginal distributions using a standard linear regression methodology for simulated samples of n elements. They are presented in Tables 10 – 12 for n=100, and the case of the normal N(0,1) distribution for one random variable, and the Weibull(δ) distribution, where δ is the shape parameter, for the second one.

The approximate relationship between Pearson's r and Kendall's τ enables us to analyze the impact of the type of a marginal distribution on r. Figure 1 presents functions $r(\tau)$ for the Clayton copula when the marginal distribution of the first random variable X is normal N(0,1) and the marginal of the second variable Y are those represented in Table 10.

Table 10. Coefficients of the polynomial approximation. Clayton copula, n=100

Coefficient	Weibull(0,5)	Exponential	Weibull(1,5)	Weibull(2,0)	Weibull(2,5)
a_1	0,7630	1,155342	1,33729	1,420882	1,473745
a_2	-0,1367	-0,0766	-0,02122	0,029073	0,059246
a_3	-0,08325	-0,34399	-0,55061	-0,64906	-0,7292
a_4	0,206633	0,182692	0,105321	0,007765	-0,05032

Table 11. Coefficients of the polynomial approximation. Frank copula, n=100

Coefficient	Weibull(0,5)	Exponential	Weibull(1,5)	Weibull(2,0)	Weibull(2,5)
a_1	0,939852	1,285098	1,387229	1,426354	1,499399
a_2	-0,00049	-0,00041	0,000288	-0,00042	0,130106
a_3	-0,30958	-0,42447	-0,45036	-0,47068	-0,63732
a_4	0,001063	0,000884	-0,001	0,000948	-0,34511

Table 12. Coefficients of the polynomial approximation. Gauss (normal) copula, n=100

Coefficient	Weibull(0,5)	Exponential	Weibull(1,5)	Weibull(2,0)	Weibull(2,5)
a_1	1,123037	1,420301	1,509639	1,542201	1,557253
a_2	0,000281	-0,00013	-0,00031	-0,00035	0,002749
a_3	-0,44679	-0,56673	-0,60338	-0,61684	-0,63445
a_4	-0,00064	0,00031	0,000842	0,00091	-0,01628

Table 13. Coefficients of the polynomial approximation. Gumbel copula, n=100

Coefficient	Weibull(0,5)	Exponential	Weibull(1,5)	Weibull(2,0)	Weibull(2,5)
a_1	1,676811443	1,787641617	1,763614843	1,727108377	1,695654409
a_2	-1,153926307	-0,817938091	-0,642773387	-0,540317764	-0,477612328
a_3	0,378567463	0,054332151	-0,082513834	-0,133359319	-0,151628861
a_4	-0,703818956	-0,360658976	-0,153403961	-0,084205969	-0,063509321
a_5	0,847882638	0,347773779	0,107779315	0,016341698	-0,014938649

For the most skewed distribution (Weibull with the shape parameter equal to 0,5) the relationship is nearly linear. When the marginal become more symmetric this relationship becomes more non-linear, concave for the positive dependence, and convex for the negative one.

Figure 2 presents the same relationship in the case of the Frank copula. The general properties of this relationship are the same as in the case of the Clayton copula. However, in the case of the most skewed marginal the function $r(\tau)$ is slightly more non-linear than in the case of the Clayton copula.

Figure 3 presents the same relationship in the case of the Gauss (normal) copula. In the case of the Gauss (normal) copula the general properties of the function $r(\tau)$ are the same as in the case of previous copulas. However, this function is more non-linear than in the case of other copulas considered.

In the case of the Gumbel copula we have only two restrictions $r_a(1) = U, \quad r_a(0) = 0$. Hence, the regression formula takes the following form:

$$r_a(\tau) = U\tau^6 + \sum_{i=1}^{5} a_i \tau^i \left(1-\tau^{6-i}\right). \tag{35}$$

The coefficients a_1, a_2, a_3, a_4, and a_5, estimated from the simulation data are presented in Table 13.

Function $r(\tau)$ for the case of the Gumbel copula is presented on Figure 4. For all the copulas considered in this case, this function is clearly the most non-linear (concave), even in the case of the most skewed Weibull distribution. The relationship between Pearson's r and Kendall's τ is approximately linear only in the case of weak dependency between considered random variables.

The approximations given by (29) and (35) are very accurate, as their accuracy measured using the R^2 statistic is close to 1. However, the usage of a simple regression technique does not guarantee in every case that the function $r_a(\tau)$ is monotonously increasing, as it should be. Therefore, it is possible to find a better approximation solving the required optimization problem with linear constraints imposed on the values of derivatives. This can be done using specialized optimization software.

The impact of the type of copula on the relationship between Pearson's r and Kendall's τ is presented in Figures 5 and 6. In Figure 5 we present this relationship when one of the two dependent variables is symmetric, N(0,1), and the second one is highly asymmetric, such as the Weibull(0,5). In Figure 6 we present the similar relationship when the second variable is characterized by only weak asymmetry, such as the Weibull(2,5). In both Figures we present the results of 1000000 simulations of the samples of 100 elements.

From these Figures one can see that in the case of highly asymmetric distributions the type of copula plays an important role. For the Clayton copula the function $r(\tau)$ is nearly linear. For the Frank copula it is not so far from being linear. However, for the Gauss (normal) copula, and especially for the Gumbel copula $r(\tau)$ is visibly non-linear. However, in the case of weakly asymmetric distributions this role is visible to a certain rather low degree only in the case of strong negative dependency. For all considered copulas the function $r(\tau)$ is non-linear, but this non-linearity is not very strong.

One of the most important characteristics of any statistical measure is its variability, measured by its variance or standard deviation. When the value of a statistical measure is bounded, the comparison of variability of different measures is not so straightforward, as for the same data, i.e. while the data dependent in the same way, the values of the measures of dependence may be quite different. Because of the

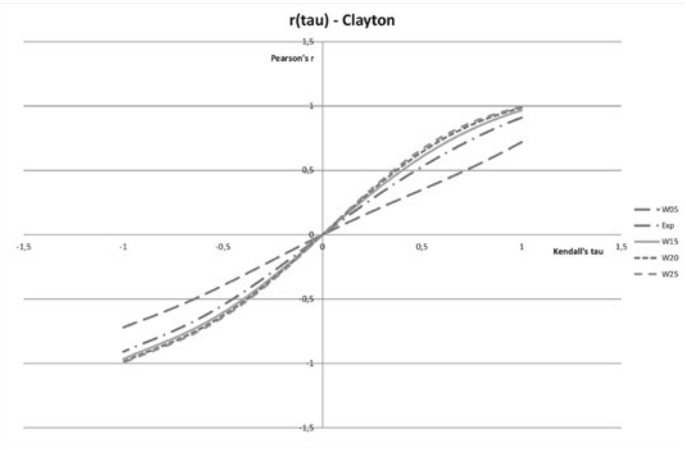

Fig. 1. Approximate relationship r(τ) – Clayton copula

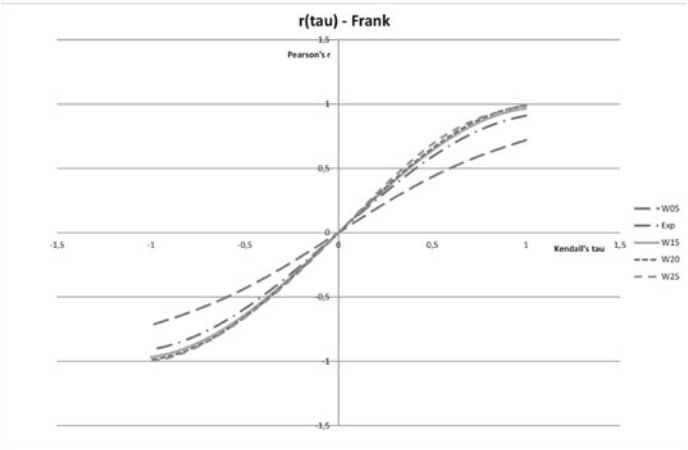

Fig. 2. Approximate relationship r(τ) – Frank copula

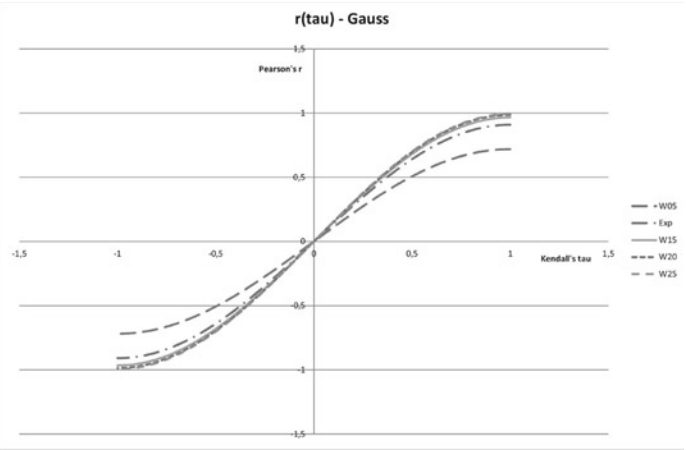

Fig. 3. Approximate relationship r(τ) – Gauss (normal) copula

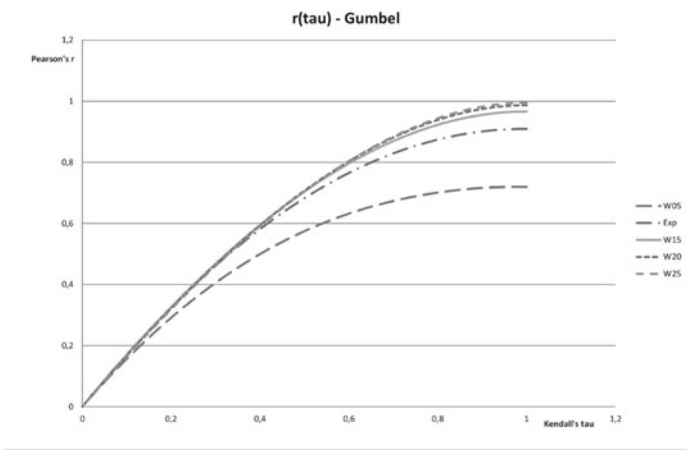

Fig. 4. Approximate relationship r(τ) – Gumbel copula

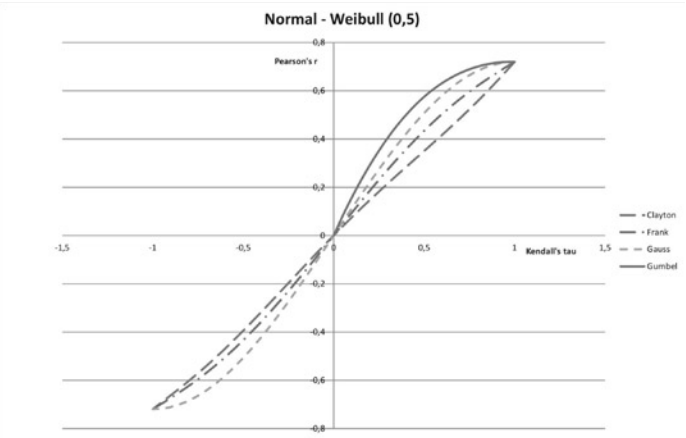

Fig. 5. Approximate relationship r(τ) – X- normal, Y – Weibull(0,5).

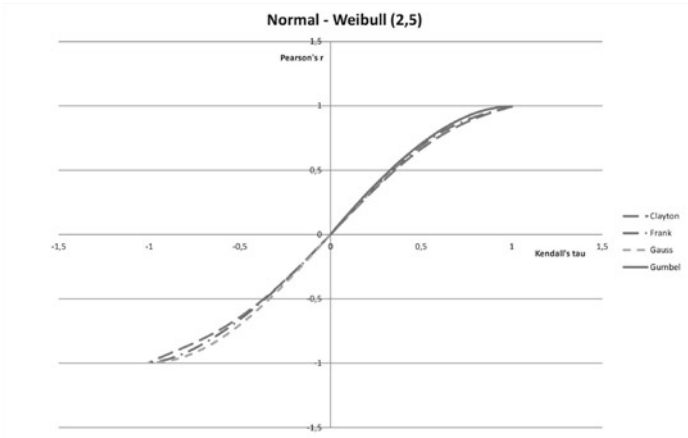

Fig. 6. Approximate relationship r(τ) – X- normal, Y – Weibull(2,5).

Table 14. Comparison of the average values of standard deviations of r and τ (n=100)

Copula	N(0,1) + Weibull (0,5)		N(0,1) + Weibull (2,5)	
	σ_τ	σ_r	σ_τ	σ_r
Clayton	0,022642	0,066985	0,047049	0,061169
Frank	0,040049	0,085868	0,040058	0,051757
Gauss (normal)	0,041423	0,071539	0,041423	0,049027
Gumbel	0,048311	0,07789	0,048311	0,064361

bounds on these values the variance of the measures whose values are closer to the bounds should be smaller. When we analyze the relationship between Pearson's r and Kendall's τ we can see that for highly skewed marginal distributions the absolute values of τ are greater than the values of r estimated from the same sample. Therefore, the observed variability of τ should be smaller than the observed variability of r. However, in the case of the more symmetric marginal distributions the values of r should be greater than the values of τ. Therefore, in

Table 15. Comparison of observed values of standard deviations of r and τ (same values of r and τ)

$r=\tau$	σ_τ	σ_r
0,924983	0,034367	0,028979
0,91753	0,036602	0,029645
0,887326	0,045321	0,038427
0,836815	0,059091	0,061777
0,766665	0,076846	0,095548
0,677651	0,10082	0,134077
0,57147	0,118242	0,172431
0,44923	0,13792	0,206378
0,312261	0,15365	0,231068
0,162145	0,162801	0,241422
0,082468	0,163901	0,238912

Table 16. Coefficients of the polynomial approximation of ρ(τ) for samples of n=100

Coefficient	Clayton	Frank	Gauss
a_1	1,466021	1,496795	1,4891
a_2	0,069209	-0,00018	-0,00051
a_3	-0,60101	-0,48078	-0,48166
a_4	0,001862	0,000489	0,001283

Table 17. Coefficients of the polynomial approximation of ρ(τ) for samples of n=100 (Gumbel copula)

Coefficient	Gumbel
a_1	1,48496
a_2	-0,09709
a_3	-0,33526
a_4	0,14299
a_5	-0,34429

the case of similar variability of both measures of dependence the observed variability of r should be smaller than the observed values of τ. In order to verify this supposition we calculated the average values of standard deviations of the estimated values of τ and r, respectively. In Table 14 we present this comparison for two cases. In the first, the marginal distribution of X is normal, and the marginal distribution of Y is the Weibull distribution with the shape parameter equal to 0,5. This is the case of a highly skewed marginal. In the second case, the marginal distribution of X is also normal, but the marginal distribution of Y is the Weibull distribution with the shape parameter equal to 2,5. Thus, this case represents the situation when both marginals are nearly symmetric. The results presented in Table 14 have been observed for the sample size n=100, and the averages have been calculated for the sets of differently dependent samples.

From Table 14 one can see that the observed variability of Kendall's τ is smaller than the variability of Pearson's r not only, as it has been expected, in the case of highly skewed variables, but also, in contrast to our supposition, in the case of nearly symmetric variables. Therefore, one can say that the variability of Kendall's τ is smaller than the variability of Pearson's r. This finding has been confirmed in another experiment in which standard deviations of both measures of dependence have been calculated from the samples for which the numerical values of both measures were the same. In Table 15 we present the results of such an experiment where samples of n=20 elements were generated from the Gumbel copula with the normal and exponential marginal distributions.

The average value of σ_τ is in this case equal to 0,09905, and the average value of σ_r is equal to 0,13442. Thus, the results presented in Table 15 are in the perfect agreement with our previous findings.

The entire analysis presented so far shows that in the considered cases of the marginal distributions that may be used in the problems of reliability prediction non-parametric measures of dependence, such as Kendall's τ, have better properties than Pearson's coefficient of linear correlation r. This is not only because of the ranges of possible values of r which may be highly misleading for practitioners, but also because of observed smaller variability. However, the question about the choice of the non-parametric statistic that is used for measuring dependence remains open.

The relationship between the most popular measures of dependence, Kendall's τ and Spearman's ρ, , both of which are considered in this paper, have been analyzed by many authors. Some important results, and references to other important papers, can be found in the paper by Fredricks and Nelsen [5]. The authors who considered this problem were interested more by the cases of weak and moderate dependence than in the cases of strong dependence, as they are more important in the context of the problem of reliability prediction. For example, Fredricks and Nelsen [5] proved the assertion previously formulated, in different versions, by other statisticians that Kendall's τ will be about two-thirds of the value of Spearman's ρ when the sample size n is large.

The results of our simulation experiments in which we have calculated not only the values of Kendall's τ, but the values of Spearman's ρ as well, let us analyze both measures in the whole spectrum of their possible values. In order to do so, we can use the same approximation methodology as that described in this section, and to find the approximate relationship $\rho(\tau)$. This relationship does not depend upon the types of the marginal distributions, but only on the type of the copula that describes the dependence. In Table 16 we present the coefficients in the expansion according to (29).

Similar coefficients for the expansion of $\rho(\tau)$ calculated from (35) for the Gumbel copula are presented in Table 17.

The impact of the type of copula on the relationship between Spearman's ρ and Kendall's τ is presented in Figure 7.

From Figure 7 one can see that the function $\rho(\tau)$ is nearly linear, for small and moderate absolute values of τ, and slightly non-linear in the case of strong dependence. The slope of the function $\rho(\tau)$ is for small and moderate values of τ fully determined by the first coefficient a_1 which is very close do 1,5. This gives a numerical confirmation of the theoretical results mentioned above. Moreover, the influence of the type of dependence is visible only in the case of the Clayton copula and the negative dependence of considered random variables. Therefore, in the case of the considered four copulas the function $\rho(\tau)$ is nearly the same.

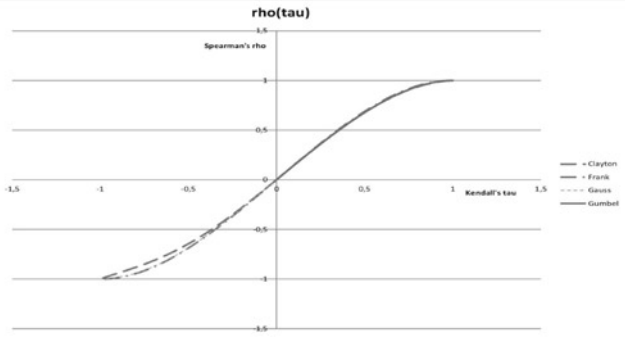

Fig. 7. Approximate relationship ρ(τ)

Because the values of ρ are greater than the respective values of τ one can think, using the same way of inference as it has been already used in this paper, that the observed variability of ρ should be smaller

Table 18. Comparison of the average values of standard deviations of ρ and τ (n=100)

Copula	σ_τ	σ_r
Clayton	0,047049	0,059035
Frank	0,040058	0,05053
Gauss (normal)	0,041423	0,05147
Gumbel	0,048311	0,060327

Table 19. Comparison of observed values of standard deviations of ρ and τ (same values of ρ and τ)

$\rho=\tau$	σ_τ	σ_ρ
0,973269	0,018296	0,017962
0,922716	0,035034	0,045827
0,850858	0,055372	0,079796
0,760838	0,078261	0,115379
0,655911	0,101829	0,149192
0,539128	0,123897	0,178623
0,412939	0,142671	0,202589
0,279648	0,156311	0,219521
0,141375	0,16336	0,228719
0,070883	0,163837	0,230153

than the variability of τ. The results of the analysis presented in Table 18 do not confirm this claim.

In contrast to our supposition the average standard deviations of τ are visible smaller than the standard deviations of ρ. Moreover, it seems that their numerical value does not depend upon the type of the copula. Therefore, one can conclude that the empirical values of Kendall's τ are less variable than the respective values of Sperman's ρ. This is also confirmed in the results of the analysis presented in Table 19 for the case of the Gumbel copula, and the sample size equal to 20. The average value of σ_τ is in this case equal to 0,103887, and the average value of σ_r is equal to 0,146776. Thus, the results presented in Table 18 confirm our claim that Kendall's τ is, from a practical point of view, a more accurate (i.e. less variable) measure of dependence than Spearman's ρ.

5. Conclusions

Pearson's coefficient of linear correlation r is the measure of dependence most popular among practitioners despite the fact that its weaknesses have been known for more than one hundred years. In this paper we have investigated its applicability in the case of reliability prediction. In this particular practical problem the assumptions necessary for a good behavior of Pearson's r are obviously not fulfilled. However, it is not well known how the lack of the fulfillment of these assumptions influences the results of the analysis. Using some simple analytical methods and comprehensive computer simulations we have arrived at the following conclusions:

a) The observed values of Pearson's r may be completely misleading in the evaluation of the strength of dependence when the dependent variables are highly skewed, as is frequently the case in the reliability context;
b) When considered distributions are not very skewed Pearson's r can be used for the evaluation of the strength of dependence;
c) The same values of Pearson's r may describe different levels of strength of dependence depending upon the type of dependence defined by the type of the copula that describes the dependent random variables;
d) Non-parametric measures of dependence such as Spearman's ρ and Kendall's τ are better than Pearson's r when applied to the analysis of dependence of life-times;
e) Kendall's τ is better than a more popular Spearman's ρ, as its variability seems to be lower.

Therefore, in searching for the most informative variables that can be used for the prediction of reliability one should use Kendall's τ as the measure of dependence.

References

1. Clayton GG. A model for Association in Bivariate Life Tables and its Applications in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. Biometrika 1978; 65: 141 – 151.
2. Elsayed EA. Overview of Reliability Testing. IEEE Transactions on Reliability 2012; 61: 282-291.
3. Embrechts P, Lindskog F, McNeil A. Modelling Dependence with Copulas and Applications to Risk Management. In: Handbook of Heavy Tailed Distributions in Finance, S. Rachev (Ed.); 329-384, Amsterdam: Elsevier 2003 (also available as the ETHZ Report, Zurich, 2001).
4. Frank MJ. On the Simultaneous Associativity of F(x,y) and x+y-F(x,y). Æquationes Mathematicæ 1979; 19: 194 – 226.
5. Fredricks GA, Nelsen RB. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. Journal of Statistical Planning and Inference 2007; 137: 2143-2150.
6. Genest C, MacKay J. Copules archimédiennes et families de lois bidimensionelles dont les marges sont données. Canadian Journal of Statistics 1986; 14: 145 – 159.
7. Genest C, Rivest L-P. Statistical Inference Procedures for Bivariate Archimedean Copulas. Journal of the American Statistical Association 1993; 88: 1034 – 1043.
8. Gumbel EJ. Distributions des valeurs extremês en plusieurs dimensions. Publications de l'Institut statstique de l'Universitè de Paris 1960; 9: 171 – 173.
9. Lee SW, Lee HK. Reliability prediction system based on the failure model for electronic components. Journal of Mechanical Science and Technology 2008; 22: 957-964.
10. Nelsen RB. Introduction to Copulas. New York: Springer, 2006.
11. Scarsini M. On measures of concordance. Stochastica 1984; 8: 201-218.
12. Schneidewind N. Predicting risk as a function of risk factors. Innovations in Systems Software Engineering 2005; 1: 63-70.
13. Sklar A. Fonctions de répartitions à n dimensions et leur marges. Publications de l'Institut de statistique de l'Universitè de Paris 1959; 8: 229-231.
14. Thaduri A, Verma AK, Gopika V, Gopinath R, Kumar U. Reliability prediction of semiconductor devices using modified physics of failure approach. International Journal of System Assurance Engineering and Management 2013; 4: 33-47.
15. US MIL-HDBK-217F Reliability Prediction of Electronic Equipment. National Technical Information Service; Springfield, Virginia, 1991.
16. US MIL-HDBK-217F Reliability Prediction of Electronic Equipment. Notice 2. National Technical Information Service; Springfield, Virginia, 1995.

17. Vořechovský M. Correlation control in small sample Monte Carlo type simulations II: Analysis of estimation formulas, random correlation and perfect uncorrelatedness. Probabilistic Engineering Mechanics 2012; 29: 105-120.
18. Wu S. Warranty Data Analysis: A Review. Quality and Reliability Engineering International 2012; 28: 795-805.
19. Xu W, Hou Y, Hung YS, Zou Y. (2013) A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. Signal Processing 2013; 93: 261-276.

Olgierd HRYNIEWICZ
Systems Research Institute
Polish Academy of Sciences
and
Warsaw School of Information Technology
Newelska 6, 01-447 Warszawa, Poland
E-mail: hryniewi@ibspan.waw.pl

Janusz KARPIŃSKI
Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warszawa, Poland
E-mail: Janusz.Karpinski@ibspan.waw.pl