Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 2257-2264
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.6196

Comparison of COVID-19 data analysis between classical dependencies and correlations via copulas

Ahmed AL-Adilee[a], Hiba Abbas AL-Asadi[a,]

[a]*Department of Mathematics, Faculty of Education, University of Kufa, Iraq*

(Communicated by Ehsan Kozegar)

Abstract

In this paper, we analyze the covid-19 data set in two ways, The first one depends on the calculation of correlation coefficient via classical mathematical representation. And the second way of analysis depends on modern technique which is associated with copula function concepts and its relationship to measures of association. Afterwards, we compare the obtained results to decide far which is better in an analysis of the examined dataset.

Keywords: Statistical inferences, Probability concepts, Correlation coefficients, Copula functions, Data analysis

1. Introduction

The data analysis process is very important to identify data from mathematical and statistical perspective and therefore giving results to improve these data or confirm its logical. So, we will study data analysis using classical distributions, data analysis by copula functions, then compare the results and see optimal way of analysis through the quality of the correlation coefficient. For this purpose, we used COVID-19 data collected from ALNajaf Health Directorate to the interval from $2020/6/-13/10/2021$ which includes infections and deaths for all days of this interval. This data follows the normal distribution, so we calculated the correlation coefficient for these data and we also found the linear regression equation to make sure the results are correct, then calculated the correlation coefficient which is associate with copula and compared the results to find out the best result between them.

The word copula was first employed in a mathematical or statistical sense by Sklar [12]. Copulas have recently become popular in financial and insurance applications [Kpanzou, Tchilabalo Abozou].

Email addresses: ahmeda.aladilee@uokufa.edu.iq (Ahmed AL-Adilee), hibaa.alasadi@student.uokufa.edu.iq (Hiba Abbas AL-Asadi)

Received: October 2021 *Accepted:* December 2021

A general form of copula corresponds to the joint distribution function. In general, copula function was firstly presented to describe the dependence structure between bivariate or multivariate random variables. They are useful in the analysis of data, whether the data follows normal distribution or not (elliptical and non-elliptical distributions). In other words, copulas are very effective tools for linear and nonlinear inferences [1].

Copulas are popular in high-dimensional statistical applications as they provide the theoretical framework in which multivariate associations and dependencies can be modeled separately from the univariate distributions of the observed variables [3].

The association between two variables is often of interest in data analysis and methodological research. Pearson's, Spearman's and Kendall's correlation coefficients are the most commonly used measures of monotone association, with the latter two usually suggested for non-normally distributed data. These three correlation coefficients can be represented as the differently weighted averages of the same concordance indicators. The weighting used in the Pearson's correlation coefficient could be preferable for reflecting monotone association in some types of continuous and not necessarily bivariate normal data [5].

The Pearson product–moment correlation coefficient ($r_p; Pearson, 1896$) and the Spearman rank correlation coefficient ($r_s; Spearman, 1904$) were developed over a century ago (for a review see Lovie, [10]). Both coefficients are widely used in psychological research. According to a search of Science Direct, of the 18,419 articles published in psychology in 2014, 24.7% reported [6].

In 2019, X. Zhang, and H. Jiang has studied copula function in order to enhance the ability of China's financial industry. The results show that there is both upper tail correlation and lower tail correlation between the two indexes, and the correlation between the upper tail and the lower tail is high [13].

Moreover, one of the most important statistical terminologies is known by regression. Its usage is very useful to describe the relationships among the connected variables. When the regression is linear between the random variables X, and Y, so this type of regression is called simple regression. In fact, the random variable Y is called the dependent variable, while the random variable X is called the independent random variable. On the other hand, when we have more than one independent random variable then the regression is called multiple linear regression, see [2].

Finally, we refer to the organization of this paper: the next part contains several basic concepts related to copula definition, dependences via copulas, classical correlation coefficients, and view of the main relations, and equations of linear regression. Third part is devoted to present the main results that consist of data analysis within parametric (classical measures of association) and nonparametric ways. Also, we have presented some tables and graphs that explain various situations of data analysis. Eventually, we present some information about the results that we have obtained which can be found within part four.

2. Basic concepts

In this part, we review some basic concepts that our study is concerned with. We begin with the basic definition of bivariate copula, and then we recall some other important properties of such function. Also, we present the forms related to the classical dependences like Person's correlation coefficient.

Definition 2.1. *[11] Let $I = [0, 1]$. A bivariate copula is a function C from $I \times I$ to I with the following properties:*

1. *For every $u, v \in I, C(u, 0) = C(0, v) = 0$;*

2. *For every* $u, v \in I, C(u, 1) = u, C(1, v) = v$;

3. *For every* $u_1, u_2, v_1, v_2 \in I$ *such that* $u_1 \leq u_2$ *and* $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

According to the essential theorem of Sklar, [8], the bivariate copula corresponds to the joint distribution function. Mathematically, suppose that X, and Y are two random variables with the joint distribution function $H(x, y) = Pr(X \leq x, Y \leq y)$, and its marginal distribution function $H_1(x)$, and $H_2(y)$, respectively. Then a bivariate copula C via Sklar's theorem is $C(H_1(x), H_2(y)) = H(x, y)$.

It is important to mention that Sklar's theorem has also shown that when the random variables are continuous then the bivariate copula is unique. There are some other properties related to the inversion formulation of the bivariate copula that has been presented, for example, in [8, 9].

Definition 2.2. *[8] Pearson's correlation coefficient is a measure of the strength of the association between two variables, given by :*

$$\rho(X, Y) = \frac{Cov(X, Y)}{[Var(X)Var(Y)]^{\frac{1}{2}}}$$

When $cov(X, Y)$ *is the covariance between* X *and* Y *which is define as:*

$$cov(X, Y) = E[XY] - E[X]E[X]$$

And $ver(X)$ *is the variance between* X *and* Y *which is define as:*

$$Var(X) = \sum_j (X_j - \mu_x)^2 f_c(X_j)$$

Definition 2.3. *[11] The population version of Kendall's tau is defined as the probability of concordance minus the probability of discordance, define as:*

$$\tau = \tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Definition 2.4. *[11] Spearman's rho is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors* (X_1, X_2) *and* (Y_1, Y_2)*, define as :*

$$\rho_(X, Y) = 3(P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0])$$

Definition 2.5. *[7] Let* X *and* Y *be two continuous random variables with copula* C*, then their Kendall's tau coefficient is given by:*

$$\tau_{X,Y} = \tau_C = 4 \int \int_{l^2} C(u, v) dC(u, v) - 1$$

Definition 2.6. *[7] Let* X *and* Y *be two continuous random variables with copula* C*, then their Spearman's rho coefficient is given by:*

$$\rho_{X,Y} = \rho_C = 12 \int_{l^2} C(u, v) du dv - 3$$

Definition 2.7. *[2] The straight line that connects two variables, one of which is called an independent variable and the other is a dependent variable, is called a simple linear model and is given by the following formula:*

$$Y = \beta_0 + \beta_1 X$$

where β_0 *is known as intercept and* β_1 *is known as the slope. In the standard* x*–*y *Cartesian plane, the intercept is the point on the* y*-axis that is intersected by the line, and the slope is the amount of change in the* y*-axis for a 1 unit change in the* x*-axis.*

3. Data set analysis, results, and discussion

First of all, we refer to the data set that we use to analyze within classical way, and then within measures of association that depend on copula. In fact, the first data set represent the number of daily infections of COVID-19 for the desired period. The second data set is the number of daily deaths at the same date of infections. Before we showing the tables of analyzed data sets, we are obliged to explain that the data has been normalized because one of the main parts of analysis of the desired data depend on copula functions that only deal with numbers belong to the unit interval. The data sets are shown in Table 1, and Table 2, respectively, as follow.

3.1. COVID-19 data analysis by classical correlations

In this part, we analyze the data using the classical correlations coefficients, as well as find the linear regression equation to check the value of the correlation coefficient and the direction of the relationship between the dependent variable (deaths) and the independent variable (infections). For this part, the analyzed data have been processed by using SPSS software, and the results of analysis can be shown by the following tables.

The results in Table 3 show that the Pearson's correlation is positive, but it is not strong or may we say it almost weak since it is less than 0.5. Consequently, we can conclude that the strength of the correlation between infections and deaths is weak. In other words, that the increase in the number of infections does not significantly and clearly effect on the number of daily deaths in Najaf. We may attribute this weakness between the number of infections and death to way of recording the data that we have obtained from the source.

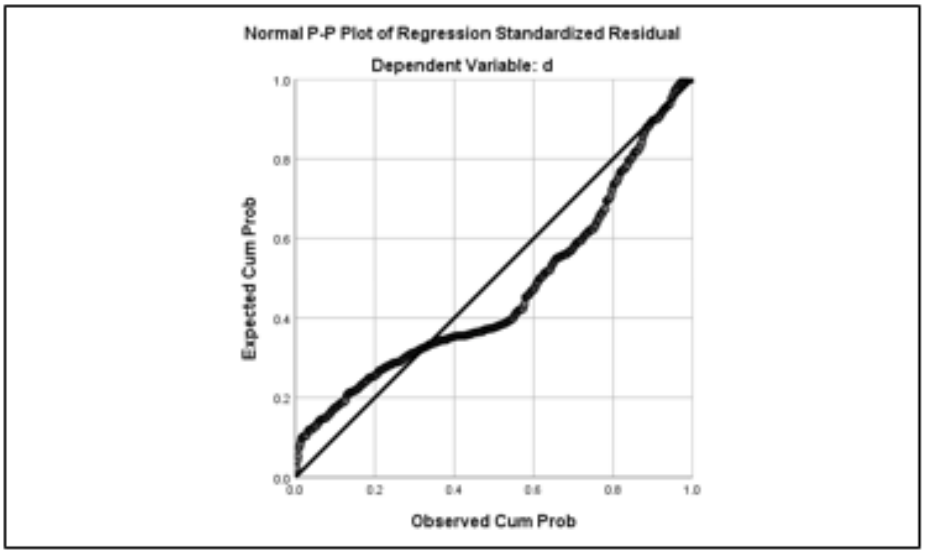

Figure 1: Linear regression model

Looking at the results of the linear regression in Table 5, we find that the slope represents a positive value, which means that for every unit increase of the independent variable, the dependent variable increased 0.006 as much. This result supports our conclusion, which shows that there is a positive correlation between infections and deaths although the correlation is weak.

Table 1: The infections for a single sample consists of 500 days

0.001	0.001	0.077	0.077	0.001	0.154	0.001	0.001	0.001	0.077	0.154	0.077
0.231	0.077	0.308	0.001	0.077	0.692	0.154	0.077	0.001	0.077	0.385	0.001
0.231	0.385	0.077	0.231	0.001	0.231	0.154	0.308	0.231	0.462	0.385	0.001
0.231	0.231	0.154	0.308	0.462	0.308	0.231	0.001	0.077	0.001	0.154	0.154
0.231	0.231	0.154	0.077	0.077	0.231	0.231	0.001	0.001	0.077	0.001	0.077
0.001	0.001	0.154	0.001	0.385	0.615	0.154	0.154	0.077	0.077	0.231	0.077
0.001	0.077	0.001	0.077	0.154	0.154	0.077	0.001	0.308	0.154	0.154	0.385
0.077	0.001	0.001	0.308	0.077	0.231	0.385	0.615	0.615	0.001	0.001	0.077
0.077	0.077	0.154	0.462	0.001	0.154	0.001	0.308	0.001	0.385	0.001	0.231
0.231	0.231	0.077	1.000	0.077	0.001	0.001	0.308	0.154	0.001	0.077	0.154
0.077	0.231	0.231	0.154	0.077	0.154	0.001	0.154	0.154	0.001	0.001	0.001
0.231	0.001	0.001	0.231	0.001	0.231	0.001	0.077	0.001	0.077	0.001	0.001
0.077	0.001	0.001	0.001	0.154	0.385	0.308	0.154	0.077	0.154	0.001	0.001
0.077	0.077	0.001	0.001	0.001	0.077	0.077	0.001	0.077	0.077	0.001	0.001
0.001	0.077	0.231	0.077	0.154	0.001	0.001	0.001	0.001	0.077	0.001	0.077
0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.001	0.001	0.154	0.001	0.001	0.231	0.001	0.001	0.001	0.001	0.001	0.001
0.001	0.001	0.001	0.154	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.001	0.001	0.001	0.001	0.001	0.077	0.077	0.001	0.001	0.001	0.001	0.001
0.001	0.001	0.001	0.001	0.001	0.077	0.077	0.001	0.077	0.077	0.001	0.001
0.154	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.154	0.154	0.001
0.077	0.001	0.077	0.077	0.154	0.077	0.077	0.077	0.154	0.077	0.231	0.154
0.308	0.231	0.001	0.154	0.001	0.308	0.231	0.308	0.231	0.154	0.154	0.462
0.077	0.385	0.001	0.077	0.385	0.385	0.846	0.154	0.154	0.308	0.462	0.385
0.154	0.154	0.308	0.077	0.308	0.385	0.154	0.231	0.154	0.154	0.308	0.001
0.154	0.231	0.154	0.077	0.154	0.077	0.231	0.001	0.001	0.077	0.001	0.077
0.001	0.154	0.077	0.077	0.077	0.001	0.308	0.001	0.001	0.154	0.077	0.077
0.231	0.154	0.001	0.001	0.231	0.077	0.154	0.001	0.154	0.077	0.077	0.001
0.077	0.001	0.077	0.001	0.077	0.001	0.154	0.001	0.001	0.077	0.001	0.001
0.023	0.001	0.001	0.001	0.077	0.077	0.077	0.077	0.154	0.001	0.077	0.001
0.077	0.001	0.154	0.001	0.077	0.231	0.154	0.001	0.154	0.001	0.001	0.001
0.001	0.001	0.001	0.077	0.154	0.154	0.154	0.154	0.001	0.001	0.077	0.231
0.538	0.154	0.154	0.308	0.154	0.231	0.231	0.231	0.231	0.154	0.385	0.308
0.077	0.308	0.077	0.077	0.231	0.154	0.001	0.231	0.615	0.077	0.308	0.077
0.077	0.385	0.231	0.077	0.231	0.231	0.615	0.001	0.077	0.231	0.001	0.231
0.001	0.001	0.231	0.462	0.001	0.308	0.154	0.154	0.001	0.077	0.308	0.154
0.001	0.001	0.154	0.154	0.154	0.077	0.001	0.077	0.154	0.154	0.077	0.001
0.001	0.001	0.077	0.077	0.077	0.077	0.001	0.077	0.001	0.001	0.001	0.077
0.001	0.001	0.001	0.001	0.077	0.077	0.077	0.077	0.077	0.001	0.001	0.077
0.001	0.154	0.154	0.001	0.154	0.077	0.077	0.001	0.001	0.001	0.077	0.001
0.077	0.001	0.001	0.154	0.001	0.077	0.001	0.001	0.001	0.001	0.001	0.001
0.001	0.001	0.077	0.001	0.001	0.001	0.001	0.001				

3.2. COVID-19 Data analysis by copula

Analyzing the data set in Table 1, and Table 2 by using copula functions that follow the Gaussian distribution. The calculations of the measures of association or the correlation coefficients with respect to the Gaussian copula show the results in Table 6.

The results in Table 7 show that the value of correlations with respect to the Gaussian copula is quietly better than the values of classical correlations. That means the data analysis and calculating the correlations is better with copula functions then the classical methods.

Table 2: The deaths for a single sample consists of 500 days

0.004	0.027	0.024	0.022	0.027	0.067	0.036	0.070	0.125	0.146	0.015	0.119
0.070	0.118	0.116	0.048	0.091	0.158	0.107	0.106	0.124	0.095	0.109	0.139
0.070	0.162	0.192	0.156	0.125	0.121	0.241	0.097	0.139	0.364	0.232	0.127
0.060	0.183	0.127	0.285	0.182	0.358	0.207	0.195	0.179	0.122	0.066	0.140
0.194	0.259	0.416	0.091	0.095	0.033	0.145	0.040	0.094	0.089	0.109	0.080
0.186	0.063	0.112	0.185	0.188	0.265	0.428	0.203	0.396	0.200	0.282	0.359
0.225	0.435	0.326	0.428	0.593	0.255	0.292	0.390	0.230	0.446	0.332	0.310
0.361	0.176	0.159	0.441	0.413	0.429	0.493	0.401	0.337	0.304	0.265	0.261
0.353	0.203	0.484	0.539	0.146	0.197	0.125	0.408	0.092	0.374	0.237	0.139
0.328	0.143	0.151	0.297	0.204	0.256	0.151	0.238	0.183	0.101	0.063	0.152
0.186	0.145	0.106	0.058	0.110	0.094	0.043	0.100	0.085	0.015	0.072	0.036
0.137	0.104	0.088	0.124	0.063	0.148	0.004	0.134	0.142	0.186	0.173	0.048
0.155	0.048	0.052	0.122	0.146	0.156	0.151	0.052	0.037	0.067	0.082	0.143
0.139	0.113	0.086	0.037	0.060	0.073	0.082	0.043	0.082	0.067	0.048	0.043
0.009	0.136	0.036	0.021	0.088	0.167	0.079	0.031	0.066	0.031	0.037	0.063
0.055	0.067	0.115	0.060	0.039	0.036	0.046	0.036	0.051	0.046	0.115	0.049
0.080	0.049	0.088	0.027	0.030	0.088	0.055	0.048	0.072	0.077	0.048	0.052
0.055	0.054	0.027	0.031	0.048	0.018	0.025	0.030	0.036	0.022	0.015	0.025
0.024	0.018	0.001	0.013	0.046	0.024	0.022	0.006	0.022	0.060	0.024	0.019
0.039	0.001	0.027	0.010	0.046	0.031	0.033	0.067	0.040	0.043	0.083	0.063
0.128	0.152	0.097	0.046	0.040	0.139	0.136	0.171	0.128	0.395	0.461	0.374
0.279	0.386	0.469	0.365	0.502	0.189	0.335	0.393	0.660	0.599	0.770	0.696
0.811	0.663	0.729	0.990	0.769	0.869	0.753	0.627	0.581	0.410	0.842	0.999
0.689	0.820	0.344	0.532	0.823	0.708	0.848	0.912	0.791	0.674	0.696	0.709
0.690	0.727	0.520	0.632	0.429	0.572	0.389	0.371	0.514	0.674	0.487	0.504
0.317	0.551	0.431	0.411	0.353	0.455	0.277	0.385	0.386	0.452	0.432	0.322
0.311	0.362	0.356	0.396	0.477	0.396	0.383	0.422	0.368	0.358	0.285	0.298
0.455	0.362	0.334	0.276	0.408	0.349	0.402	0.422	0.340	0.347	0.343	0.197
0.291	0.332	0.340	0.311	0.359	0.204	0.243	0.218	0.355	0.183	0.224	0.142
0.159	0.174	0.197	0.188	0.274	0.432	0.314	0.282	0.264	0.291	0.382	0.331
0.246	0.227	0.240	0.264	0.306	0.232	0.347	0.274	0.241	0.234	0.282	0.334
0.298	0.438	0.370	0.367	0.227	0.308	0.359	0.355	0.268	0.341	0.377	0.367
0.405	0.414	0.419	0.562	0.562	0.583	0.729	0.562	0.517	0.517	0.569	0.626
0.545	0.578	0.437	0.678	0.539	0.654	0.620	0.699	0.487	0.601	0.720	0.547
0.592	0.581	0.431	0.399	0.462	0.490	0.526	0.572	0.472	0.489	0.422	0.405
0.385	0.373	0.465	0.562	0.484	0.557	0.501	0.510	0.423	0.387	0.343	0.316
0.331	0.283	0.398	0.273	0.231	0.300	0.319	0.164	0.259	0.273	0.322	0.297
0.222	0.154	0.133	0.237	0.261	0.301	0.264	0.240	0.256	0.177	0.227	0.277
0.255	0.218	0.292	0.219	0.267	0.274	0.307	0.235	0.235	0.222	0.213	0.237
0.355	0.353	0.362	0.227	0.317	0.104	0.197	0.109	0.232	0.109	0.100	0.069
0.088	0.067	0.049	0.039	0.030	0.058	0.060	0.063	0.046	0.094	0.127	0.098
0.045	0.060	0.052	0.060	0.046	0.072	0.121	0.080				

4. Conclusion

The computations of correlation coefficients via copula are a modern technique that has not use to analyze Covid-19 data to best of our knowledge. The correlations within copula shows an improvement on their values comparing to their values with direct approach. The study shows that the relationship between the number of infections and number of deaths is not strongly correlated. This weakness may be due to the fact that the recorded data in ALnajaf Health Directorate is inaccurate and not organized well. We recommend that ALNajaf Health Directorate has to use a

Table 3: The pearson's correlation coeffiecint

Correlations

		infections	deaths
infections	Pearson Correlation	1	.437**
	Sig. (2-tailed)		.000
	N	500	500
deaths	Pearson Correlation	.437**	1
	Sig. (2-tailed)	.000	
	N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4: kendall's tau and spearman's rho correlation coefficients.

Correlations

			infections	deaths
Kendall's tau_b	infections	Correlation Coefficient	1.000	.363**
		Sig. (2-tailed)	.	.000
		N	500	500
	deaths	Correlation Coefficient	.363**	1.000
		Sig. (2-tailed)	.000	.
		N	500	500
Spearman's rho	infections	Correlation Coefficient	1.000	.477**
		Sig. (2-tailed)	.	.000
		N	500	500
	deaths	Correlation Coefficient	.477**	1.000
		Sig. (2-tailed)	.000	.
		N	500	500

**. Correlation is significant at the 0.01 level (2-tailed).

better way of recording to the COVID-19 data to ensure that any study that concern with COVID-19 data set lead to better interpretation. We also mention that data set by using statistical inference with respect to copulas has a better description than classical correlations.

References

[1] A.A. Ahmed and O. Hassan, *Comments on copula functions and their relationship to probability density functions*, Iraqi J. Sci. 2020 (2020) 1115–1122.

[2] S.I. Bangdiwala, *Regression: simple linear*, Int. J. Injury Cont. Safety Promotion 25(1) (2018) 113–115.

Table 5: The coefficients of the linear regression model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.400	.117		3.422	.001
	n	.006	.001	.437	10.843	.000

a. Dependent Variable: d

Table 6: The results of the correlation coefficients via Gaussian copula

Type of copula	ρ	τ	P_c
Gaussian copula	**0.46**	**0.304**	**0.456**

Table 7: Corr. with classical methods and corr. via copula

Correlation	Corr. with classical methods	Corr. with copula
ρ	0.437	0.46
τ	0.36	0.304
P_c	0.47	0.456

[3] L. Benettazzo, *Copula VAR Models With Applications to Genetic Networks*, Universita degli Studi di Padova Dipartimento di Scienze Statistiche Corso di Laurea Magistrale in Scienze Statistiche, 2017.

[4] P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin and S. Zeger, *Springer Series in Statistics*, Springer, New York, 2008.

[5] N.S. Chok, *Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data*, Doctoral dissertation, University of Pittsburgh, (2016).

[6] J.C. De Winter, S.D. Gosling and J. Potter, *Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data*, Psych. Methods 21(3) (2016).

[7] Y. Elouerkhaoui, *Credit correlation*, Palgrave Macmillan, 2017.

[8] J.D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded (Statistics: A Series of Textbooks and Monographs)*, CRC Press; 4th edition, 2003.

[9] S. Kautish, S.L. Peng and A.J. Obaid, *Computational intelligence techniques for combating COVID-19*, Springer International Publishing, 2021.

[10] A.D. Lovie, *Who discovered Spearman's rank correlation?*, Brit. J. Math. Statist. Psych. 48(2) (1995) 255–269

[11] R.B. Nelsen, *An introduction to copulas*, Springer Science & Business Media, 2007.

[12] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publications de l'Institut Statistique de l'Université de Paris. 8 (1959) 229–231.

[13] X. Zhang and H. Jiang, *Application of copula function in financial risk analysis*, Comput. Electric. Engin. 77 (2019) 376–388.