Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds

Sorana-Daniela BOLBOACĂ¹, Lorentz JÄNTSCHI ²

¹“Iuliu Hațieganu” University of Medicine and Pharmacy, 13 Emil Isac, 400023 Cluj-Napoca, Romania; ²Technical University of Cluj-Napoca, 15 Constantin Daicoviciu, 400020 Cluj-Napoca, Romania
sbolboaca@umfcluj.ro, lori@academicdirect.org.

Abstract
A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of molecular descriptors family on structure-activity relationships. Starting from the results obtained by applying of MDF-SAR methodology on pyrimidine derivatives and from the assumption that the measured activity (compounds’ inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used, etc.), the abilities of Pearson, Spearman, Kendall’s, and Gamma correlation coefficients in analysis of estimated toxicity were studied and are presented.

Keywords
Multiple linear regressions, Correlation coefficients, Molecular Descriptors Family on Structure-Activity Relationships (MDF-SAR)

Introduction

QSAR (Quantitative Structure-Activity Relationships) is an approach which is able to indicate for a given compound or a class of compounds which feature of structure characteristics is correlated with its activity [1]. In QSAR analysis were proposed several
approaches for development. Simple and multiple linear regressions is one of the more successful techniques use by many researcher in construct of QSAR models [2-4].

Correlation coefficient is a simple statistical measure of relationship between one dependent and one or more than one independent variables and it is use as a measure of the statistical fit of a regression based model in QSAR [5]. Its squared value (the coefficient of determination) it is most frequently used parameter as a measure of the goodness-of-fit of the model [6-10].

A new approach of molecular descriptors family on structure-activity relationships (MDF-SAR) was developed [11], and proved its usefulness in estimation and prediction of: toxicity [12, 13], mutagenicity [12], antioxidant efficacy [14], antituberculotic activity [15], antimalarial activity [16], antiallergic activity [17], anti-HIV-1 potencies [18], inhibition activity on carbonic anhydrase II [19] and IV [20].

Several correlation coefficients based on different statistical hypothesis are known and most frequently used today: Pearson correlation coefficient, Spearman rank correlation coefficient and Spearman semi-quantitative correlation coefficient, Kendall tau-a, -b and -c correlation coefficients, Gamma correlation coefficient [5].

Starting from the results obtained by applying of MDF-SAR methodology on a sample of sixty-seven compounds and from the assumption that the measured activity (compounds’ inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used), the abilities of Pearson, Spearman, Kendall’s, and Gamma correlation coefficients in analysis of estimated toxicity were studied.

**Multi-varied MDF-SAR model of pyrimidine derivatives**

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of MDF-SAR methodology.

The set of pyrimidine derivatives (2,4-Diamino-5-(substituted-benzyl)-pyrimidine derivatives) with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was previously studied by Ting-Lan Chiu & Sung-Sau So by the use of neural network approach [21].
By applying the MDF-SAR methodology on the sample of sixty-seven pyrimidine derivatives, a multi-varied model with four descriptors revealed to have good performances in prediction and estimation of inhibitory activity.

The multi-varied MDF-SAR model with four descriptors had the following equation:

\[ Y_{\text{est}} = 3.78 + 1.62 \cdot iImrKHt + 2.37 \cdot liMDWHg + 6.40 \cdot IsDrJQt - 8.52 \cdot 10^{-2} \cdot \text{LSPmEQg} \]

Analyzing the MDF-SAR model with four descriptors it could be said that inhibitory activity considers compounds geometry (\(g\)) and topology (\(t\)), being related with the number of directly bonded hydrogen’s (\(H\)) of compounds and with the partial charge (\(Q\)) as atomic properties.

Statistical characteristics of the MDF-SAR model with four descriptors are in table 1 and 2.

**Table 1. Statistical characteristics of the multi-varied MDF-SAR model with four descriptors**

<table>
<thead>
<tr>
<th>Characteristic (notation)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of variable ((v))</td>
<td>4</td>
</tr>
<tr>
<td>Correlation coefficient ((r))</td>
<td>0.9517</td>
</tr>
<tr>
<td>95% Confidence Intervals for (r) (95% CI (r))</td>
<td>[0.9223, 0.9701]</td>
</tr>
<tr>
<td>Squared correlation coefficient ((r^2))</td>
<td>0.9058</td>
</tr>
<tr>
<td>Adjusted squared correlation coefficient ((r^2_{\text{adj}}))</td>
<td>0.8997</td>
</tr>
<tr>
<td>Standard error of estimated ((s_{\text{est}}))</td>
<td>0.1919</td>
</tr>
<tr>
<td>Fisher parameter ((F_{\text{est}}))</td>
<td>149</td>
</tr>
<tr>
<td>Cross-validation leave-one-out (loo) score ((r^2_{\text{cv-loo}}))</td>
<td>0.8932</td>
</tr>
<tr>
<td>Fisher parameter for loo analysis ((F_{\text{pred}}))</td>
<td>130*</td>
</tr>
<tr>
<td>Standard error for leave-one-out analysis ((s_{\text{loo}}))</td>
<td>0.2044</td>
</tr>
<tr>
<td>Model stability ((r^2 - r^2_{\text{cv-loo}}))</td>
<td>0.0126</td>
</tr>
<tr>
<td>(r^2)((iImrKHt, liMDWHg))</td>
<td>0.2020</td>
</tr>
<tr>
<td>(r^2)((iImrKHt, IsDrJQt))</td>
<td>0.0047</td>
</tr>
<tr>
<td>(r^2)((iImrKHt, LSPmEQg))</td>
<td>0.1482</td>
</tr>
<tr>
<td>(r^2)((liMDWHg, IsDrJQt))</td>
<td>0.0003</td>
</tr>
<tr>
<td>(r^2)((liMDWHg, LSPmEQg))</td>
<td>0.0212</td>
</tr>
<tr>
<td>(r^2)((IsDrJQt, LSPmEQg))</td>
<td>0.0664</td>
</tr>
</tbody>
</table>

\* \(p < 0.001\)

**Table 2. Statistics of the regression MDF-SAR model with four descriptors**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>StdError</th>
<th>t Stat</th>
<th>95%CIcoefficient</th>
<th>(r(Y_{m,\text{desc}}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.1999</td>
<td>18.92*</td>
<td>[3.38, 4.18]</td>
<td>n.a.</td>
</tr>
<tr>
<td>iImrKHt</td>
<td>0.0709</td>
<td>22.85*</td>
<td>[1.48, 1.76]</td>
<td>0.4803</td>
</tr>
<tr>
<td>liMDWHg</td>
<td>0.1500</td>
<td>15.81*</td>
<td>[2.07, 2.67]</td>
<td>0.0558</td>
</tr>
<tr>
<td>IsDrJQt</td>
<td>1.4779</td>
<td>4.33*</td>
<td>[3.45, 9.36]</td>
<td>0.0336</td>
</tr>
<tr>
<td>LSPmEQg</td>
<td>0.0182</td>
<td>-4.68*</td>
<td>[-0.12, -0.12]</td>
<td>0.0231</td>
</tr>
</tbody>
</table>

StdError = standard error; t Stat = Student tets parameter; 95% CIcoefficient = 95% confidence interval associated with regression coefficients; \(Y_{m}\) = measured inhibitory activity; desc = molecular descriptor; * \(p < 0.001\)
Graphical representation of the measured versus estimated by MDF-SAR model with four descriptors inhibitory activity is in figure 1.

![Figure 1. Plot of measured vs estimated by MDF-SAR inhibitory activity](image)

Internal validation of the four-varied MDF SAR model with four descriptors was performed through splitting the whole set into training and test sets by applying of a randomization algorithm.

The coefficients for each model obtained in training sets, in conformity with the generic equation $Y_{est} = a_0 + a_1 \cdot iM_{rKHt} + a_2 \cdot lM_{DWHg} + a_3 \cdot lS_{rJQt} - a_4 \cdot 10^{-2} \cdot LSP_{mEQg}$, the number of compounds in training ($N_{tr}$) and test ($N_{ts}$) sets, the correlation coefficient for training ($r_{tr}$) and test ($r_{ts}$) sets with associated 95% confidence intervals (95%CI$_{tr}$ and 95%CI$_{ts}$), the Fisher parameter associated with training (F$_{tr}$) and test (F$_{ts}$) sets, and the Fisher’s Z parameter of correlation coefficients comparison ($Z_{tr-rs}$) are in table 3.
Table 3. Statistics results on training versus test sets

<table>
<thead>
<tr>
<th>a0</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>Ntr</th>
<th>rtr</th>
<th>95%CI_rtr</th>
<th>Ftr</th>
<th>Nts</th>
<th>rts</th>
<th>95%CI_rts</th>
<th>Fts</th>
<th>Zrtr-rts</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.93</td>
<td>1.61</td>
<td>2.43</td>
<td>6.56</td>
<td>9.67-10^-2</td>
<td>35</td>
<td>0.949</td>
<td>0.899, 0.974</td>
<td>67</td>
<td>32</td>
<td>0.958</td>
<td>0.916, 0.980</td>
<td>59</td>
<td>0.418†</td>
</tr>
<tr>
<td>3.98</td>
<td>1.57</td>
<td>2.45</td>
<td>6.55</td>
<td>7.52-10^-2</td>
<td>36</td>
<td>0.951</td>
<td>0.905, 0.975</td>
<td>73</td>
<td>31</td>
<td>0.951</td>
<td>0.899, 0.976</td>
<td>61</td>
<td>0.000†</td>
</tr>
<tr>
<td>3.84</td>
<td>1.55</td>
<td>2.15</td>
<td>9.08</td>
<td>9.12-10^-2</td>
<td>37</td>
<td>0.944</td>
<td>0.893, 0.908</td>
<td>66</td>
<td>30</td>
<td>0.949</td>
<td>0.895, 0.976</td>
<td>55</td>
<td>0.206†</td>
</tr>
<tr>
<td>3.94</td>
<td>1.59</td>
<td>2.42</td>
<td>6.10</td>
<td>8.18-10^-2</td>
<td>38</td>
<td>0.951</td>
<td>0.907, 0.974</td>
<td>78</td>
<td>29</td>
<td>0.947</td>
<td>0.890, 0.975</td>
<td>50</td>
<td>0.144†</td>
</tr>
<tr>
<td>3.91</td>
<td>1.56</td>
<td>2.25</td>
<td>8.22</td>
<td>1.04-10^-1</td>
<td>39</td>
<td>0.963</td>
<td>0.931, 0.981</td>
<td>110</td>
<td>28</td>
<td>0.937</td>
<td>0.867, 0.971</td>
<td>39</td>
<td>1.069†</td>
</tr>
<tr>
<td>4.18</td>
<td>1.51</td>
<td>2.44</td>
<td>6.06</td>
<td>7.22-10^-2</td>
<td>40</td>
<td>0.956</td>
<td>0.917, 0.975</td>
<td>92</td>
<td>27</td>
<td>0.936</td>
<td>0.863, 0.971</td>
<td>35</td>
<td>0.721†</td>
</tr>
<tr>
<td>3.76</td>
<td>1.63</td>
<td>2.32</td>
<td>7.35</td>
<td>1.02-10^-1</td>
<td>41</td>
<td>0.963</td>
<td>0.931, 0.980</td>
<td>116</td>
<td>26</td>
<td>0.935</td>
<td>0.858, 0.971</td>
<td>34</td>
<td>1.104†</td>
</tr>
<tr>
<td>3.97</td>
<td>1.58</td>
<td>2.39</td>
<td>5.11</td>
<td>9.36-10^-1</td>
<td>42</td>
<td>0.956</td>
<td>0.919, 0.976</td>
<td>99</td>
<td>25</td>
<td>0.954</td>
<td>0.896, 0.980</td>
<td>34</td>
<td>0.115†</td>
</tr>
<tr>
<td>3.64</td>
<td>1.64</td>
<td>2.30</td>
<td>7.00</td>
<td>8.15-10^-2</td>
<td>43</td>
<td>0.955</td>
<td>0.917, 0.975</td>
<td>98</td>
<td>24</td>
<td>0.944</td>
<td>0.873, 0.976</td>
<td>37</td>
<td>0.407†</td>
</tr>
<tr>
<td>3.72</td>
<td>1.66</td>
<td>2.43</td>
<td>5.78</td>
<td>8.12-10^-2</td>
<td>44</td>
<td>0.938</td>
<td>0.889, 0.966</td>
<td>72</td>
<td>23</td>
<td>0.964</td>
<td>0.916, 0.985</td>
<td>54</td>
<td>1.030†</td>
</tr>
<tr>
<td>3.59</td>
<td>1.64</td>
<td>2.25</td>
<td>4.94</td>
<td>9.98-10^-2</td>
<td>45</td>
<td>0.947</td>
<td>0.904, 0.970</td>
<td>86</td>
<td>22</td>
<td>0.957</td>
<td>0.898, 0.982</td>
<td>37</td>
<td>0.411†</td>
</tr>
<tr>
<td>3.86</td>
<td>1.55</td>
<td>2.23</td>
<td>8.68</td>
<td>8.86-10^-2</td>
<td>46</td>
<td>0.940</td>
<td>0.894, 0.967</td>
<td>78</td>
<td>21</td>
<td>0.983</td>
<td>0.958, 0.993</td>
<td>43</td>
<td>2.290†</td>
</tr>
<tr>
<td>4.04</td>
<td>1.54</td>
<td>2.36</td>
<td>6.46</td>
<td>7.31-10^-2</td>
<td>47</td>
<td>0.949</td>
<td>0.911, 0.972</td>
<td>96</td>
<td>20</td>
<td>0.963</td>
<td>0.906, 0.985</td>
<td>34</td>
<td>0.538†</td>
</tr>
<tr>
<td>3.63</td>
<td>1.63</td>
<td>2.24</td>
<td>4.27</td>
<td>8.93-10^-2</td>
<td>48</td>
<td>0.940</td>
<td>0.895, 0.966</td>
<td>82</td>
<td>19</td>
<td>0.963</td>
<td>0.904, 0.986</td>
<td>44</td>
<td>0.852†</td>
</tr>
<tr>
<td>3.98</td>
<td>1.57</td>
<td>2.42</td>
<td>6.49</td>
<td>8.59-10^-2</td>
<td>49</td>
<td>0.946</td>
<td>0.905, 0.969</td>
<td>93</td>
<td>18</td>
<td>0.960</td>
<td>0.894, 0.985</td>
<td>36</td>
<td>0.535†</td>
</tr>
<tr>
<td>3.77</td>
<td>1.61</td>
<td>2.32</td>
<td>6.37</td>
<td>8.46-10^-2</td>
<td>50</td>
<td>0.943</td>
<td>0.902, 0.968</td>
<td>91</td>
<td>17</td>
<td>0.974</td>
<td>0.927, 0.991</td>
<td>52</td>
<td>1.294†</td>
</tr>
<tr>
<td>3.67</td>
<td>1.63</td>
<td>2.22</td>
<td>6.56</td>
<td>1.01-10^-1</td>
<td>51</td>
<td>0.954</td>
<td>0.919, 0.973</td>
<td>115</td>
<td>16</td>
<td>0.950</td>
<td>0.858, 0.983</td>
<td>17</td>
<td>0.126†</td>
</tr>
<tr>
<td>3.81</td>
<td>1.61</td>
<td>2.39</td>
<td>6.87</td>
<td>7.70-10^-1</td>
<td>52</td>
<td>0.951</td>
<td>0.916, 0.972</td>
<td>112</td>
<td>15</td>
<td>0.950</td>
<td>0.853, 0.984</td>
<td>22</td>
<td>0.032†</td>
</tr>
<tr>
<td>3.69</td>
<td>1.65</td>
<td>2.36</td>
<td>6.32</td>
<td>8.21-10^-2</td>
<td>53</td>
<td>0.953</td>
<td>0.919, 0.972</td>
<td>118</td>
<td>14</td>
<td>0.956</td>
<td>0.864, 0.986</td>
<td>17</td>
<td>0.128†</td>
</tr>
<tr>
<td>3.97</td>
<td>1.56</td>
<td>2.40</td>
<td>6.16</td>
<td>7.51-10^-2</td>
<td>54</td>
<td>0.951</td>
<td>0.916, 0.971</td>
<td>115</td>
<td>13</td>
<td>0.954</td>
<td>0.851, 0.987</td>
<td>17</td>
<td>0.122†</td>
</tr>
</tbody>
</table>

† p > 0.05; * p < 0.01

Definitions, Formulas, Interpretations, PHP functions, and Results

A number of add notations were used in the study, as follows:

- Pearson product-moment correlation coefficient (named after Karl Pearson (1857 - 1936), a major contributor to the early development of statistics):
  - \( r_{prs} \) = the Pearson correlation coefficient;
  - \( r_{prs}^2 \) = the squared Pearson correlation coefficient;
  - \( t_{prs,df} \) = the Student test parameter, and its significance \( p_{prs,df} \) at a significance level of 5% (where \( df = \) the degree of freedom);

- Spearman's rank correlation coefficient (named after Charles Spearman (1863 - 1945), English psychologist known for his work in statistics - factor analysis, and Spearman's rank correlation coefficient):
  - \( r_{spm} \) = the Spearman rank correlation coefficient
  - \( r_{spm}^2 \) = the squared of Spearman rank correlation coefficient;
  - \( t_{spm,df} \) = the Student test parameter, and its significance \( p_{spm,df} \);
Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds

Sorana-Daniela BOLBOACĂ and Lorentz JÁNTSCHI

- $r_{SO}^2$ = the squared of Spearman semi-Quantitative correlation coefficient;
- $t_{SO}$ = the Student test parameter, and its significance $p_{SO};$

- Kendall’s tau correlation coefficients (named after Maurice George Kendall (1907 - 1983), a prominent British statistician; published in monograph *Rank Correlation* in 1948);
- $\tau_{Ken,a} = $ the *Kendall tau-a* correlation coefficient;
- $\tau_{Ken,a}^2 = $ the squared of *Kendall tau-a* correlation coefficient;
- $Z_{Ken,ta} = $ the $Z$-test parameter of Kendall tau-a correlation coefficient, and its significance $p_{Ken,ta};$
- $\tau_{Ken,b} = $ the *Kendall tau-b* correlation coefficient;
- $\tau_{Ken,b}^2 = $ the squared of *Kendall tau-b* correlation coefficient;
- $Z_{Ken,rb} = $ the $Z$-test parameter of Kendall tau-b, and its significance $p_{Ken,rb};$
- $\tau_{Ken,c}^2 = $ the *Kendall tau-c* correlation coefficient;
- $\tau_{Ken,c}^2 = $ the squared of *Kendall tau-c* correlation coefficient;
- $Z_{Ken,rc} = $ the $Z$-test parameter of *Kendall tau-c*, and its significance $p_{Ken,rc};$

- Gamma correlation coefficient (also known as Goodman and Kruskal's gamma):
- $\Gamma = $ the *Gamma* correlation coefficient;
- $\Gamma^2 = $ the squared of *Gamma* correlation coefficient;
- $Z_{\Gamma} = $ the $Z$-test parameter of *Gamma* correlation coefficient, and its significance $p_{\Gamma}.$

A series of *.php* programs which to facilitate the calculation and to display of above-described correlation coefficients and their statistics (Student-test and $Z$-test parameters and associated significances) were implemented and was use in order to reach the objective of study [22].

**Pearson correlation coefficient**

**Definition:** a measure the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another.

**Assumptions:** both variable (variables $Y_m$ and $Y_{est}$) are interval or ratio variables and are well approximated by a normal distribution, and their joint distribution is bivariate normal [23].
**Formula**

\[
    r_{Prs} = \frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{est-i} - \bar{Y}_{est})}{\sqrt{\left(\sum (Y_{m-i} - \bar{Y}_m)^2\right)\left(\sum (Y_{est-i} - \bar{Y}_{est})^2\right)}}
\]

where \( Y_{m-i} \) is the value of the measured inhibitory activity for compound \( i \) \((i = 1, 2, \ldots, 67)\), \( \bar{Y}_m \) is the average of the measured inhibitory activity, \( Y_{est-i} \) is the value of the estimated inhibitory activity for compound \( i \), and \( \bar{Y}_{est} \) is the average of the estimated inhibitory activity.

**Interpretation**

The Pearson correlation coefficient can take values from -1 to +1. A value of +1 shows that the variables are perfectly linear related by an increasing relationship, a value of -1 shows that the variables are perfectly linear related by a decreasing relationship, and a value of 0 shows that the variables are not linear related by each other. There is considered a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5.

The coefficient of determination (or \( r \) squared) gives information about the proportion of variation in the dependent variable which might be considered as being associated with the variation in the independent variable.

**Related statistics:**

- The squared of Pearson correlation coefficient or Pearson coefficient of determination \( (r_{Prs}^2) \):
  - Describe the proportion of variance in \( Y_m \) that is related with linear variation of \( Y_{est} \);
  - Can take values from 0 to 1.

**Statistical test**

Student t-test was used to determine if the value of Pearson correlation coefficient is statistically significant, at a significance level of 5%.

The null hypothesis vs. the alternative hypothesis was:

- \( H_0: r_{Prs} = 0 \) (there is no correlation between the variables)
- \( H_1: r_{Prs} <> 0 \) (variables are correlated)
For a significance level equal with 5%, a p-value associated to $t_{\text{Prs,df}}$ less than 0.05 means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis. In other words there is a statistically significant linear relationship between the variables.

**PHP implementation**

In order to compute the statistics associated with Pearson correlation coefficient, three functions were implemented:

```php
function coef_rk(&$y1,&$y2)
{
    $my1=m1($y1);
    $dy2=m2($y1, $y1)-$my1*$my1;
    $mx1=m1($y2);
    $my2=m2($y2, $y2);
    $d2x=m2($y2, $y2);
    $myx=m2($y1, $y2);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2, $y2);
    $d2x2=m2($y2, $y2);
    $r2=pow($myx-$mx1*$my1,2)/($d2x2*$dy2);
    return $r2;
}
function t_p($n,$k,$r)
{
    return $r*pow($n-$k-1,0.5)/pow(1-pow($r,2),0.5);
}
function p_t($t,$df)
{
    $p = $df/2;
    $x = 0.5+0.5*$t/pow(pow($t,2)+$df,0.5);
    $beta_gam = exp( -logBeta($p, $p) + $p * log($x) + $p * log(1.0 - $x) );
    return (2.0 * $beta_gam * betaFraction(1.0 - $x, $p, $p) / $p);
}
```

The statistics of Pearson correlation coefficients are computed as follows:

- Pearson correlation coefficient:
  
  $$r_{\text{pe}} = \text{coef_rk}(\text{cmp}[0],\text{cmp}[1]);$$

  where $\text{cmp}[0]$ is the measured inhibitory activity ($Y_m$), and $\text{cmp}[1]$ is the estimated by MDF-SAR model with four descriptor inhibitory activity ($Y_{est}$).

- t Student parameter:

  $$t_{\text{pe}} = \text{t_p}($n,1,pow(r_{\text{pe}},0.5));$$

- Significance of t Student parameter

  $$p_{\text{pe}} = \text{p_t}(t_{\text{pe}},$n-2);$$
**Results:**

$$r_{Prs}^2 = 0.9058$$

$$t_{Prs,1} = 24.99$$

$$p_{Prs,1} = 4.74 \cdot 10^{-33} \%$$


**Spearman’s rank correlation coefficient**

**Definition:** a non-parametric measure of correlation between variable which assess how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Frequently the Greek letter $\rho$ (rho) is use to abbreviate the Spearman correlation coefficient.

Spearman’s rank correlation is satisfactory for testing the null hypothesis of no relationship, but is difficult to interpret as a measure of the strength of the relationship [24].

**Assumptions:**

- Does not required any assumptions about the frequency distribution of the variables;
- Does not required the assumption that the relationship between variable is linear;
- Does not required the variable to be measured on interval or ration scale.

**Formula**

In order to compute the Spearman rank correlation coefficient, the two variables (Y\textsubscript{m}, respectively Y\textsubscript{est}) were converted to ranks (see table 4 for exemplification). For each measured and estimated inhibitory activity a rank was assigned (RankY\textsubscript{m} - for measure inhibitory activity, RankY\textsubscript{est} - for estimated by MDF-SAR model inhibitory activity) according with the position of value into a sort serried of values.

In assignment of rank process, the lowest value had the lowest rank. When there are two equal values for two different compounds (for measured and/or estimated inhibitory activity), the associated rank had equal values and was calculated as means of corresponding ranks. For example, the compounds abbreviated as c\_52 and c\_59 have the same measured inhibitory activity (6.45, see table 4). The rank associated with these values is equal with 13.5 (is the average between the rank for c\_52 - 13 and the rank of c\_59 - 14).
Table 4. Compounds abbreviation, measured and estimated activity and associated ranks

<table>
<thead>
<tr>
<th>Abb.</th>
<th>Y_m</th>
<th>RankY_m</th>
<th>Y_est</th>
<th>RankY_est</th>
<th>Abb.</th>
<th>Y_m</th>
<th>RankY_m</th>
<th>Y_est</th>
<th>RankY_est</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_64</td>
<td>6.07</td>
<td>1</td>
<td>0</td>
<td>6.4626</td>
<td>c_32</td>
<td>6.92</td>
<td>35</td>
<td>0</td>
<td>6.8423</td>
</tr>
<tr>
<td>c_65</td>
<td>6.10</td>
<td>2</td>
<td>0</td>
<td>6.2948</td>
<td>c_66</td>
<td>6.93</td>
<td>36.5</td>
<td>0</td>
<td>6.8225</td>
</tr>
<tr>
<td>c_67</td>
<td>6.18</td>
<td>3</td>
<td>0</td>
<td>6.1479</td>
<td>c_68</td>
<td>6.93</td>
<td>36.5</td>
<td>0</td>
<td>6.9609</td>
</tr>
<tr>
<td>c_54</td>
<td>6.20</td>
<td>4</td>
<td>0</td>
<td>6.1595</td>
<td>c_17</td>
<td>6.96</td>
<td>38</td>
<td>0</td>
<td>6.7150</td>
</tr>
<tr>
<td>c_37</td>
<td>6.23</td>
<td>5</td>
<td>0</td>
<td>6.3859</td>
<td>c_40</td>
<td>6.96</td>
<td>38</td>
<td>0</td>
<td>6.9298</td>
</tr>
<tr>
<td>c_48</td>
<td>6.25</td>
<td>6</td>
<td>0</td>
<td>6.2254</td>
<td>c_45</td>
<td>6.99</td>
<td>40</td>
<td>0</td>
<td>7.0283</td>
</tr>
<tr>
<td>c_31</td>
<td>6.28</td>
<td>7</td>
<td>0</td>
<td>6.3483</td>
<td>c_41</td>
<td>7.02</td>
<td>41</td>
<td>0</td>
<td>7.1919</td>
</tr>
<tr>
<td>c_49</td>
<td>6.30</td>
<td>8</td>
<td>0</td>
<td>6.3528</td>
<td>c_15</td>
<td>7.04</td>
<td>42</td>
<td>0</td>
<td>7.0225</td>
</tr>
<tr>
<td>c_10</td>
<td>6.31</td>
<td>9</td>
<td>0</td>
<td>6.4703</td>
<td>c_28</td>
<td>7.06</td>
<td>43</td>
<td>0</td>
<td>6.8355</td>
</tr>
<tr>
<td>c_56</td>
<td>6.35</td>
<td>10</td>
<td>0</td>
<td>6.3149</td>
<td>c_09</td>
<td>7.20</td>
<td>44</td>
<td>0</td>
<td>7.3115</td>
</tr>
<tr>
<td>c_47</td>
<td>6.39</td>
<td>11</td>
<td>0</td>
<td>6.3859</td>
<td>c_18</td>
<td>7.22</td>
<td>45</td>
<td>0</td>
<td>7.2156</td>
</tr>
<tr>
<td>c_53</td>
<td>6.40</td>
<td>12</td>
<td>0</td>
<td>6.4703</td>
<td>c_43</td>
<td>7.23</td>
<td>46</td>
<td>0</td>
<td>6.8855</td>
</tr>
<tr>
<td>c_52</td>
<td>6.45</td>
<td>13.5</td>
<td>1</td>
<td>6.2913</td>
<td>c_29</td>
<td>7.35</td>
<td>47</td>
<td>0</td>
<td>7.2724</td>
</tr>
<tr>
<td>c_59</td>
<td>6.45</td>
<td>13.5</td>
<td>1</td>
<td>6.3367</td>
<td>c_14</td>
<td>7.41</td>
<td>48</td>
<td>0</td>
<td>7.4072</td>
</tr>
<tr>
<td>c_16</td>
<td>6.46</td>
<td>15</td>
<td>0</td>
<td>6.5851</td>
<td>c_24</td>
<td>7.53</td>
<td>49</td>
<td>0</td>
<td>7.5476</td>
</tr>
<tr>
<td>c_34</td>
<td>6.47</td>
<td>16</td>
<td>0</td>
<td>6.3422</td>
<td>c_22</td>
<td>7.54</td>
<td>50</td>
<td>0</td>
<td>7.1218</td>
</tr>
<tr>
<td>c_58</td>
<td>6.48</td>
<td>17</td>
<td>0</td>
<td>6.5536</td>
<td>c_26</td>
<td>7.66</td>
<td>51.5</td>
<td>6</td>
<td>7.7002</td>
</tr>
<tr>
<td>c_35</td>
<td>6.53</td>
<td>18</td>
<td>0</td>
<td>6.9755</td>
<td>c_08</td>
<td>7.66</td>
<td>51.5</td>
<td>6</td>
<td>7.8841</td>
</tr>
<tr>
<td>c_42</td>
<td>6.55</td>
<td>19</td>
<td>0</td>
<td>6.7654</td>
<td>c_27</td>
<td>7.69</td>
<td>54</td>
<td>7</td>
<td>7.4715</td>
</tr>
<tr>
<td>c_30</td>
<td>6.57</td>
<td>20.5</td>
<td>2</td>
<td>6.5625</td>
<td>c_13</td>
<td>7.69</td>
<td>54</td>
<td>7</td>
<td>5.4895</td>
</tr>
<tr>
<td>c_61</td>
<td>6.57</td>
<td>20.5</td>
<td>2</td>
<td>6.7594</td>
<td>c_12</td>
<td>7.69</td>
<td>54</td>
<td>7</td>
<td>5.7593</td>
</tr>
<tr>
<td>c_33</td>
<td>6.59</td>
<td>22</td>
<td>0</td>
<td>6.8010</td>
<td>c_04</td>
<td>7.71</td>
<td>56.5</td>
<td>8</td>
<td>7.5841</td>
</tr>
<tr>
<td>c_51</td>
<td>6.60</td>
<td>23</td>
<td>0</td>
<td>7.0616</td>
<td>c_11</td>
<td>7.71</td>
<td>56.5</td>
<td>8</td>
<td>7.6497</td>
</tr>
<tr>
<td>c_39</td>
<td>6.65</td>
<td>24</td>
<td>0</td>
<td>6.4993</td>
<td>c_19</td>
<td>7.72</td>
<td>58</td>
<td>0</td>
<td>7.7915</td>
</tr>
<tr>
<td>c_38</td>
<td>6.70</td>
<td>25</td>
<td>0</td>
<td>6.6297</td>
<td>c_23</td>
<td>7.77</td>
<td>59</td>
<td>0</td>
<td>7.7014</td>
</tr>
<tr>
<td>c_57</td>
<td>6.78</td>
<td>26</td>
<td>0</td>
<td>6.7552</td>
<td>c_25</td>
<td>7.80</td>
<td>60</td>
<td>0</td>
<td>7.9130</td>
</tr>
<tr>
<td>c_60</td>
<td>6.82</td>
<td>28</td>
<td>3</td>
<td>6.7091</td>
<td>c_01</td>
<td>7.82</td>
<td>61</td>
<td>0</td>
<td>7.6576</td>
</tr>
<tr>
<td>c_44</td>
<td>6.82</td>
<td>28</td>
<td>3</td>
<td>6.8748</td>
<td>c_21</td>
<td>7.94</td>
<td>62</td>
<td>0</td>
<td>7.8130</td>
</tr>
<tr>
<td>c_55</td>
<td>6.82</td>
<td>28</td>
<td>3</td>
<td>6.9318</td>
<td>c_06</td>
<td>8.07</td>
<td>63</td>
<td>0</td>
<td>8.2391</td>
</tr>
<tr>
<td>c_20</td>
<td>6.84</td>
<td>30</td>
<td>0</td>
<td>7.1067</td>
<td>c_03</td>
<td>8.08</td>
<td>64</td>
<td>0</td>
<td>8.1224</td>
</tr>
<tr>
<td>c_46</td>
<td>6.86</td>
<td>31</td>
<td>0</td>
<td>6.5813</td>
<td>c_07</td>
<td>8.12</td>
<td>65</td>
<td>0</td>
<td>8.1353</td>
</tr>
<tr>
<td>c_50</td>
<td>6.89</td>
<td>33</td>
<td>4</td>
<td>6.4794</td>
<td>c_05</td>
<td>8.18</td>
<td>66</td>
<td>0</td>
<td>8.0372</td>
</tr>
<tr>
<td>c_62</td>
<td>6.89</td>
<td>33</td>
<td>4</td>
<td>6.6942</td>
<td>c_02</td>
<td>8.35</td>
<td>67</td>
<td>0</td>
<td>8.2702</td>
</tr>
<tr>
<td>c_63</td>
<td>6.89</td>
<td>33</td>
<td>4</td>
<td>6.8475</td>
<td>c_01</td>
<td>8.38</td>
<td>67</td>
<td>0</td>
<td>8.2702</td>
</tr>
</tbody>
</table>

The method of rank assignment for more than two equal values of measured and/or estimated inhibitory activity is the same as for two equal values. If there are an odd number of compounds which have the same measured value (see compounds c_60, c_44, and c_55 from table 2) then the rank will be an integer ((27+28+29)/3 = 28, see the rank for c_60, c_44, and c_55).
In studied example, there are equal values for measured activity: five situations of two equal values (c_52-c_59, c_30-c_61, c_66-c_36, c_26-c_08, and c_04-c_11), and three situations of three equal values (c_60-c_44-c_55, c_50-c_62-c_63, and c_27-c_13-c_12).

By conversion of the measured and estimated inhibitory activity to ranks, the distribution of ranks does not depend on the distribution of measured, respectively estimated inhibitory activity.

The formula for calculation of the Spearman rank correlation coefficient is:

\[
    r_{\text{Spm}} = \frac{\sum (R_{Ym-i} - \bar{R}_{Ym})(R_{Y\text{est}-i} - \bar{R}_{Y\text{est}})}{\sqrt{\left(\sum (R_{Ym-i} - \bar{R}_{Ym})^2\right)\left(\sum (R_{Y\text{est}-i} - \bar{R}_{Y\text{est}})^2\right)}}
\]

where \(R_{Ym-i}\) is the rank of the measured inhibitory activity for compound \(i\), \(\bar{R}_{Ym-i}\) is the average of the measured inhibitory activity, \(R_{Y\text{est}-i}\) is the rank of the estimated by MDF-SAR inhibitory activity for compound \(i\), and \(\bar{R}_{Y\text{est}-i}\) is the average of the estimated inhibitory activity.

The simple formula for \(r_{\text{Spm}}\) is based on the difference between each pairs of ranks:

\[
    r_{\text{Spm}} = 1 - \frac{6\sum D^2}{n(n^2 - 1)}
\]

where \(D\) is the differences between each pair of ranks (e.g. \(D = R_{Ym-1} - R_{Y\text{est}-1}\)) and \(n\) is the volume of the sample.

The formula of the *Spearman semi-quantitative method* is:

\[
    r_{sQ} = \sqrt{\frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{\text{est}-i} - \bar{Y}_{\text{est}})}{\sqrt{\left(\sum (Y_{m-i} - \bar{Y}_m)^2\right)\left(\sum (Y_{\text{est}-i} - \bar{Y}_{\text{est}})^2\right)}} \cdot \sqrt{\frac{\sum (R_{Ym-i} - \bar{R}_{Ym})(R_{Y\text{est}-i} - \bar{R}_{Y\text{est}})}{\sqrt{\left(\sum (R_{Ym-i} - \bar{R}_{Ym})^2\right)\left(\sum (R_{Y\text{est}-i} - \bar{R}_{Y\text{est}})^2\right)}}}
\]

**Interpretation**

- Identical with Pearson correlation coefficient.

**Related statistics**:

- \(r_{\text{Spm}}^2\) = the squared of *Spearman* rank correlation coefficient;
- \(r_{sQ}^2\) = the squared of semi-quantitative correlation coefficient.
Statistical significance:

- Compute by the use of a permutation test (a statistical test in which the reference distribution is obtained by permuting the observed data points across all possible outcomes, given a set of conditions consistent with the null hypothesis);
- Comparing the observed $r_{Spm}$ with published tables for different levels of significance (e.g., 0.05, 0.01…). It is a simple solution when the researchers want to know the significance within a certain range or less than a certain value;
- Tested by applying the Student t-test (for sample sizes > 20): the method used in this study.

The null hypothesis vs. the alternative hypothesis for Spearman rank correlation coefficient was:

$H_0$: $r_{Spm} = 0$ (there is no correlation between the ranked pairs)
$H_1$: $r_{Spm} <> 0$ (ranked pairs are correlated)

The null hypothesis vs. the alternative hypothesis for semi-quantitative correlation coefficient was:

$H_0$: $r_{sQ} = 0$ (there is no correlation between the ranked pairs)
$H_1$: $r_{sQ} <> 0$ (ranked pairs are correlated)

**PHP implementation**

The formulas for Spearman and respectively semi-quantitative correlation coefficients used two defined above functions ($t_p$ and respectively $p_t$). The Spearman rank correlation coefficient used the $coef_rk$ function defined as:

```php
function coef_rk(&$y1,&$y2){
    $my1=m1($y1);
    $dy2=m2($y1,$y1)-$my1*$my1;
    $mx1=m1($y2);
    $mxy=m2($y2,$y1);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2,$y2);
    $dx2=$mx2-$m2x;
    $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
    return $r2;
}
```

where

```php
function m1(&$v){
    $rez=0;
    $n=count($v);
    for($i=1;$i<$n;$i++)
        $rez+=$v[$i];
    return $rez/$n;
}
```
Spearman correlation coefficient

The statistics of Spearman rank correlation coefficients are computed as follows:

- Spearman correlation coefficient:
  
  \[ \text{spearman coefficient:} \]
  
  \[ \text{spearman coefficient:} \]
  
  where \( \text{poz[0]} \) is the position on sort series of measured inhibitory activity, and \( \text{poz[1]} \) is the position on sort serried of estimated inhibitory activity by MDF-SAR model with four descriptor.

- t Student parameter:
  
  \[ \text{t student parameter:} \]
  
  - Significance of t Student parameter
    
    \[ \text{significance of t student parameter:} \]
    

Semi-quantitative correlation coefficient

The statistics of semi-quantitative correlation coefficients are computed as follows:

- Semi-quantitative correlation coefficient:
  
  \[ \text{semi-quantitative coefficient:} \]
  
  - t Student parameter:
    
    \[ \text{t student parameter:} \]
    
    - Significance of t Student parameter
      
      \[ \text{significance of t student parameter:} \]

Results:

\[ r_{spr}^2 = 0.8606 \]

\[ t_{spr,1} = 20.03 \]

\[ p_{spr,1} = 1.62 \cdot 10^{-29} \]
\[ r_{sQ}^2 = 0.8829 \]
\[ t_{sQ} = 22.14 \]
\[ p_{sQ} = 5.57 \cdot 10^{-32} \]

**Kendall’s rank correlation coefficients**

**Definition**

Kendall-tau is a non-parametric correlation coefficient that can be used to assess and test correlations between non-interval scaled ordinal variables. Frequently the Greek letter \( \tau \) (tau), is used to abbreviate the Kendall tau correlation coefficient.

The Kendall tau correlation coefficient is considered to be equivalent to the Spearman rank correlation coefficient. While Spearman rank correlation coefficient is like the Pearson correlation coefficient but computed from ranks, the Kendall tau correlation rather represents a probability.

There are three Kendall’s tau correlation coefficient known as \( \tau \)-a, \( \tau \)-b, and \( \tau \)-c.

**Formula**

Let \((Y_{m-i}, Y_{est-i})\) and \((Y_{m-j}, Y_{est-j})\) be the pair of measured and estimated inhibitory activity. If \(Y_{m-j} - Y_{m-i}\) and \(Y_{est-j} - Y_{est-i}\), where \(i < j\) have the same sign the pair is *concordant*, if have opposite signs the pair is *discordant*.

In a sample of \(n\) observations it can be found \(n(n-1)/2\) pairs corresponding to choices \(1 \leq i < j \leq n\).

The formulas of Kendall’s tau correlation coefficients are as follows:

- **Kendall tau-a correlation coefficient (\(\tau_{Ken,a}\))**:
  \[ \tau_{Ken,a} = \frac{(C-D)}{[n(n-1)/2]} \]

- **Kendall tau-b correlation coefficient (\(\tau_{Ken,b}\))**:
  \[ \tau_{Ken,b} = \frac{(C-D)/\sqrt{[n(n-1)/2-t)(n(n-1)/2-u)]}} \]
  where \(t\) is the number of tied \(Y_m\) values and \(u\) is the number of tied \(Y_{est}\) values.

- **Kendall tau-c correlation coefficient (\(\tau_{Ken,c}\))**:
  \[ \tau_{Ken,c} = \frac{2(C-D)/n^2} \]
Interpretation:

• If the agreement between the two rankings is perfect and the two rankings are the same, the coefficient has value 1.
• If the disagreement between the two rankings is perfect and one ranking is the reverse of the other, the coefficient has value -1.
• For all other arrangements the value lies between -1 and 1, and increasing values imply increasing agreement between the rankings.
• If the rankings are independent, the coefficient has value 0.

Related statistics:

• $\tau_{Kena}^2$ = the squared of Kendall tau-a correlation coefficient;
• $\tau_{Kemb}^2$ = the squared of Kendall tau-b correlation coefficient;
• $\tau_{Ken,c}^2$ = the squared of Kendall tau-c correlation coefficient.

Statistical significance:

Statistical significance of the Kendall’s tau correlation coefficient is tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Kendall’s tau correlation coefficients was:

• Kendall tau-a correlation coefficient:
  \[ H_0: \tau_{Kena} = 0 \] (there is no correlation between the two variables)
  \[ H_1: \tau_{Kena} <> 0 \] (the two variables are correlated)
• Kendall tau-b correlation coefficient:
  \[ H_0: \tau_{Kemb} = 0 \] (there is no correlation between the two variables)
  \[ H_1: \tau_{Kemb} < > 0 \] (the two variables are correlated)
• Kendall tau-c correlation coefficient:
  \[ H_0: \tau_{Ken,c} = 0 \] (there is no correlation between the two variables)
  \[ H_1: \tau_{Ken,c} < > 0 \] (the two variables are correlated)

PHP implementation

*Kendall* function was implemented in order to calculate the Kendall’s tau correlation coefficients:

```php
function Kendall(&$cmp){
```
$n = count($cmp[0]);
$spz = 0;
if(!is_numeric($cmp[0][0])) $spz = 1;
$C = 0;
$D = 0;
$E = 0;
for($i=$spz; $i<$n-1; $i++)
    for($j=$i+1; $j<$n; $j++){
        $sgx = 0;
        $sgy = 0;
        if($cmp[0][i]>$cmp[0][j]) $sgx = 1;
        if($cmp[0][i]<$cmp[0][j]) $sgx = -1;
        if($cmp[1][i]>$cmp[1][j]) $sgy = 1;
        if($cmp[1][i]<$cmp[1][j]) $sgy = -1;
        if($sgx*$sgy>0) $C++;
        if($sgx*$sgy<0) $D++;
        if($sgx*$sgy==0) $E++;
        if($sgx==0)$tied_x[$i][]=$j;
        if($sgy==0)$tied_y[$i][]=$j;
    }
$u1 = 0;
$vt = 0;
$vu = 0;
$sv2t = 0;
$sv2u = 0;
if(isset($tied_x))
    if(is_array($tied_x))
        foreach($tied_x as $vx){
            $nt = count($vx)+1;
            $t1 += $nt*($nt-1);
            $vt += $nt*($nt-1)*(2*$nt+5);
            $sv2t += $nt*($nt-1)*($nt-2);
        }
if(isset($tied_y))
    if(is_array($tied_y))
        foreach($tied_y as $vy){
            $nu = count($vy)+1;
            $u1 += $nu*($nu-1);
            $vu += $nu*($nu-1)*(2*$nu+5);
            $sv2u += $nu*($nu-1)*($nu-2);
        }
$sv1 = $st1*$su1;
$stl /= 2;
$sul /= 2;
$sv2 = $sv2t*$sv2u;
\$S = \$C - \$D;
\$n = \$n - \$pz;
\$cn2 = \$n*(\$n-1)/2;
\$tau\_a2 = pow(\$S,2)/pow(\$cn2,2);
\$v\_tau\_a = \$cn2*(2*\$n+5)/9;
\$z\_tau\_a = \$S/pow(\$v\_tau\_a,0.5);
\$T = (\$cn2-\$tl)*(\$cn2-\$tl);
\$tau\_b2 = pow(\$S,2)/\$T;
\$vT0 = \$v\_tau\_a - (\$vt + \$vu)/18;
\$vt1 = \$v1/(4*\$cn2);
\$vT2 = \$v2/(18*\$cn2*(\$n-2));
\$v\_tau\_b = pow(\$vt0 + \$vt1 + \$vt2, 0.5);
\$z\_tau\_b = \$S/\$v\_tau\_b;
\$gamma = pow((\$C - \$D)/(\$C + \$D),2);
\$v\_gamma = (2*\$n+5)/9.0/\$cn2;
\$z\_gamma = \$gamma/pow(\$v\_gamma,0.5);
\$tau\_c2 = 4*pow(\$S,2)/pow(\$n,4);
\$z\_tau\_c = \$z\_tau\_b*(\$n-1)/\$n;
return array( \$tau\_a2, \$z\_tau\_a, \$tau\_b2, \$z\_tau\_b,
\$tau\_c2, \$z\_tau\_c, \$gamma, \$z\_gamma );
}\n
where C is the number of concordant pairs (C = (<, <) or (>, >)), D is the number of discordant pairs (D = (<, >) or (>, <)), and E is the number of equal pairs (E = (=, .) or (., =)).

**Results:**

- **Kendall's $\tau_a$** correlation coefficient and associated statistics:
  \[ \tau_{\text{Ken,a}}^2 = 0.6129 \]
  \[ Z_{\text{Ken,ra}} = 9.37 \]
  \[ p_{\text{Ken,ra}} = 7.44 \times 10^{-21} \] (4)

- **Kendall's $\tau_b$** correlation coefficient and associated statistics:
  \[ \tau_{\text{Ken,b}}^2 = 0.6177 \]
  \[ Z_{\text{Ken,rb}} = 9.37 \]
  \[ p_{\text{Ken,rb}} = 7.26 \times 10^{-21} \] (5)

- **Kendall's $\tau_c$** correlation coefficient and associated statistics:
  \[ \tau_{\text{Ken,c}}^2 = 0.5948 \]
  \[ Z_{\text{Ken,rc}} = 9.23 \]
  \[ p_{\text{Ken,rc}} = 2.70 \times 10^{-20} \] (6)
**Gamma correlation coefficient**

**Definition**

The Gamma correlation coefficient ($\Gamma$, gamma) is a measure of association between variables that comparing with Kendall’s tau correlation coefficients is more resistant to tied data [25], being preferable to Spearman rank or Kendall tau when data contain many tied observations [26].

**Formula**

The formula for Gamma correlation coefficient is:

$$\Gamma = \frac{(C-D)}{(C+D)}$$

where the significance of C and D were described above.

**Interpretation:**

- In the same manner as the Kendall tau correlation coefficient.

**Related statistics:**

- $\Gamma^2$ = the squared of Gamma correlation coefficient.

**Statistical significance:**

Statistical significance of Gamma correlation coefficient was tested by the $Z$-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Gamma correlation coefficients was:

$H_0$: $\Gamma = 0$ (there is no correlation between the two variables)

$H_1$: $\Gamma \neq 0$ (the two variables are correlated).

**PHP implementation**

The function which computes the Gamma correlation coefficient was presented at Kendall’s tau correlation coefficient, in PHP implementation chapter.

**Results:**

$$\Gamma^2 = 0.6208$$

$$Z_\Gamma = 7.43$$

$$p_\Gamma = 1.11 \cdot 10^{-13}$$
Conclusions

All seven computational methods used to evaluate the correlation between measured and estimated by MDF-SAR model inhibitory activity are statistically significant (p-value always less than 0.0001, correlation coefficients always greater than 0.5).

More research on other classes of biologic active compounds may reveal whether it is appropriate to analyze the MDF-SAR models using the Pearson correlation coefficient or other correlation coefficients (Spearman rank, Kendall’s tau, or Gamma correlation coefficient).

Acknowledgement

Research was partly supported by UEFISCSU Romania through the project ET46/2006.

References


[23] Pearson's Correlation Coefficient [online], Available at: [http://www.texasoft.com/winkpear.html](http://www.texasoft.com/winkpear.html)
