The International Journal of Biological Markers, Vol. 17 no. 2, pp. 148-151

Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval

R. Artusi[1], P. Verderio[1], E. Marubini[1,2]

[1]Operative Unit of Medical Statistics and Biometry, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan
[2]Institute of Medical Statistics and Biometry, Università degli Studi di Milano, Milan - Italy

The International Journal of Biological Markers, like many other biomedical journals dealing with cancer research, often publishes papers whose principal aim it is to test the association between quantative measurements of biological variables. When these are expressed on continuous scales, the statistics most frequently adopted to test their association are the Bravais-Pearson (parametric) and the Spearman (non-parametric) correlation coefficients. In this note the correlation coefficient estimate (statistic) will be denoted by the Latin letter r, while the "true" correlation coefficient (parameter) of the underlying population will be denoted by the Greek letter ρ.

The Bravais-Pearson correlation coefficient (ρ_{BP}) is a suitable measure of association when n couples of continuous data ((y_i, x_i) with i=1,2,...n), collected on the same experimental unit, follow a bivariate normal distribution. In this case the only relationship that can be postulated is the linear one. Two different regression lines (see Fig. 1) can be defined: the first (l_1) corresponding to the linear regression of y on x and the second (l_2) corresponding to the linear regression of x on y. The two straight lines intersect at a point whose coordinates are the means of the observed y_i and x_i, respectively; this point is the vertex of an angle θ, defined by l_1 and l_2, which is an expression of the strength of the linear association between y and x. The Bravais-Pearson correlation coefficient (ρ_{BP}) is the geometrical mean of the slopes of the two regression lines and corresponds to the cosine of θ. In absence of association the two straight lines are perpendicular (θ = 90°), so that ρ_{BP} = cos 90° = 0. When there is a complete association the two straight lines overlap: if the resulting single straight line has a positive slope (i.e. y increases with increasing values of x), θ = 0° and ρ_{BP} = cos 0° = 1; if it has a negative slope (i.e. y decreases with increasing values of x), θ = 180° and ρ_{BP} = cos 180° = -1.

The Spearman correlation coefficient (ρ_S) is usually adopted when the assumption of the bivariate normal distribution is not tenable. It is known that ρ_S is computed as ρ_{BP}, changing the integer 1,2,...n to $y_1,y_2,...y_n$ according to their relative magnitude; the same procedure is performed for $x_1,x_2,...x_n$. This transformation makes it possible to move from the scales in which the original data are collected towards the same scale, i.e. that of ranks. The ranks do not follow the normal bivariate distribution and therefore the correlation coefficient cannot

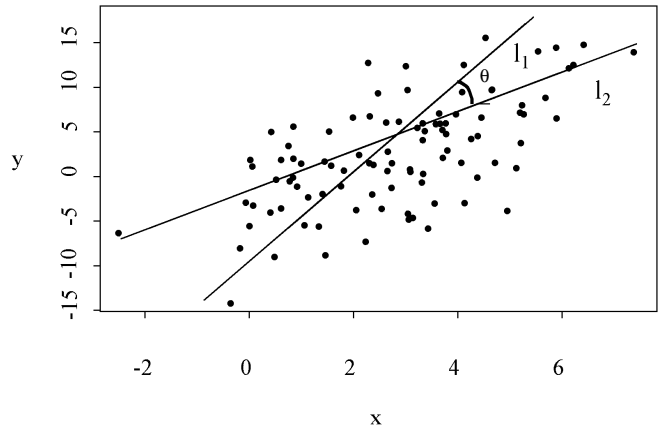

Fig. 1 - *Geometrical interpretation of the Bravais-Pearson correlation coefficient.*

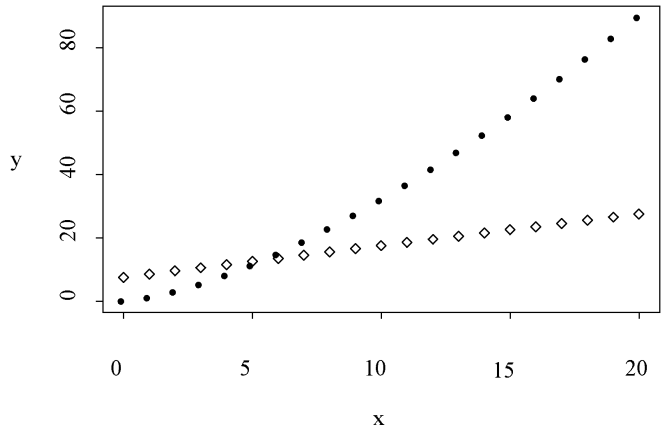

Fig. 2 - *Monotonic relationship between data on original scales and corresponding linear relationship between data expressed on rank scale.*

0393-6155/148-04$02.00/0

be geometrically interpreted as before. Even though ρ_S cannot be thought of as showing the extent of the *linear* relationship between the variables underlying the ranks, ρ_S can be considered an index of the general monotonicity of the underlying relationship. Recall that a relationship between two variables is monotonic if its graphical representation does not show any "peaks" and "valleys". For example, in Figure 2 the relation $y_i = x_i^{3/2}$ (x_i=i, i=0,1,…20) is drawn (dot plot); this is a monotonic increasing relationship between the two variables. After replacing y_i and x_i with the corresponding ranks, the monotonicity of their relationship implies a linear relationship between the corresponding ranks, as shown by the straight line (diamond plot) in the same Figure. Figure 3 may aid to better grasp the meaning of the two coefficients:

a) a straight relationship between original variables is translated into r = 1, like for both coefficients;
b) absence of monotonicity is reflected by a null value of both coefficients;
c) at a glance, a monotonic but *not linear* relation of y_i against x_i on the original scales emerges. The computation of the Bravais-Pearson correlation coefficient (r_{BP} = 0.77) is misleading as it captures only the linear component of the relationship between the two original variables. Instead of the "true" relationship between the two variables (exponential in this case), a naive reader could be led to rely upon the linear relationship like the one drawn in the panel. The Spearman correlation coefficient (r_S = 1) on the other hand informs the reader about the absolute monotonicity which, however, can be grasped only by a graphical representation of the values on the original scales.

The different meaning of the two correlation coefficients should also be taken into account when conclu-

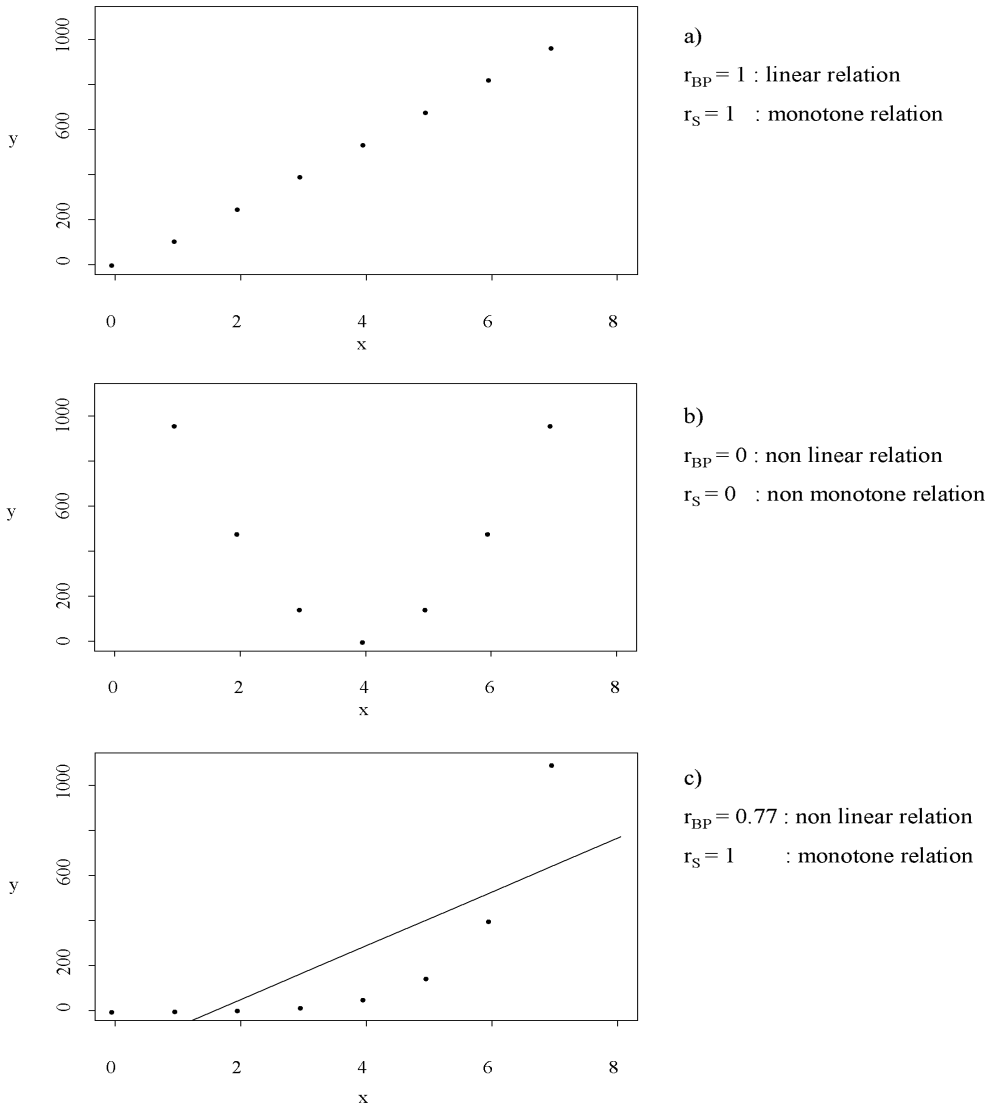

Fig. 3 - *Correlation coefficients as an expression of tendency towards linearity and/or monotonicity.*

TABLE I - VEGF CONCENTRATION IN LYSED WHOLE BLOOD AND PLATELET COUNT IN TEN HEALTHY SUBJECTS

Subject ID	VEGF (pg/mL)	Platelet ($x10^{-3}/\mu L$)
N1	612	376
N2	160	188
N3	531	255
N4	309	206
N5	321	231
N6	262	227
N7	411	233
N8	196	184
N9	450	296
N10	756	261

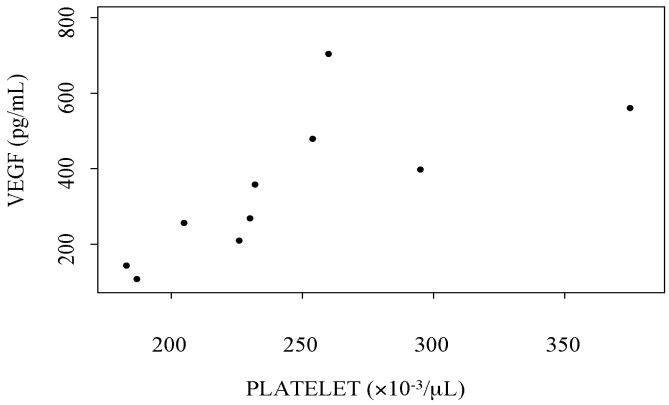

Fig. 4 - *Relationship between VEGF in lysed whole blood and platelet count in ten healthy subjects.*

sions are drawn after rejection of the null hypothesis H_0: $\rho = 0$, where ρ is pertinent to both correlation coefficients. As regards ρ_{BP}, H_0 is tested by resorting to a t-test with n-2 degrees of freedom. Namely:

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

For sufficiently large sample sizes, say n>10, the same statistic can be adopted to test the null hypothesis even for ρ_S (1, 2). For sample sizes of ten or less a specific probability table is available (3). The latter reports the thresholds of the estimated Spearman correlation coefficient corresponding to the sampling distribution of the Spearman correlation coefficient under H_0 for $\alpha = 0.05$ and $\alpha = 0.01$ (two-tailed). As an example we considered the vascular endothelial growth factor (VEGF) concentration in lysed whole blood and the platelet count reported by Dittadi et al (4) in ten healthy subjects (data in Table I).

The calculated values for r_{BP} and r_S (0.723 and 0.915, respectively) suggest that the relationship between the two variables is monotonic but not strictly linear (see Fig. 4). The authors considered ρ_S as measure of association. By computing r_S to test H_0, the calculated value ($r_S = 0.915$) results to be statistically different from zero, as r_S is greater than both the tabulated thresholds (0.648 and 0.794, for n = 10) for $\alpha = 0.05$ and $\alpha = 0.01$ (two-tailed test) and this implies rejecting H_0.

When an estimate of ρ is computed it is always advisable to give, together with the point estimate, the pertinent confidence interval (CI). The latter provides a measure of precision and allows to draw conclusions about the quantitative (clinical, biological, etc.) relevance of the association in the underlying population.

The CI of the Bravais-Pearson correlation coefficient is estimated by means of the transformation of ρ_{BP} suggested by Fisher (z-transformation) (5). This transformation is approximately normally distributed with variance $\sigma_z^2 = 1/(n-3)$, independent of ρ_{BP}. The z-transformation is not appropriate for the Spearman correlation coefficient because the sampling distribution of this coefficient can be defined only under H_0. For sufficiently large sample sizes, say n>5, the confidence interval of ρ_S can be estimated by resorting to the bootstrap resampling method (6, 7). The latter can be adopted to try to bypass the lack of knowledge of the sampling distribution of the Spearman correlation coefficient. The bootstrap resampling method allows to gain better knowledge of this distribution by calculating the Spearman correlation coefficient (r_S^*) in a large number (at least 1000) of bootstrap samples. These are obtained by random resampling *with replacement* from the original set of data; each of the bootstrap samples has the same size as the original one. Among the available bootstrapping algorithms for resampling planes, the Bias Corrected and Accelerated (BCa) method seems to be preferable for several reasons: no estimate of the variance of ρ_S is needed, no invalid parameter values can be obtained, and the corresponding coverage error is closest to the nominal one. It is worth noting that, due to the random resampling, the same bootstrapping algorithm applied to the same original set of data provides similar but not exactly equal results. In the appendix we report the commands to run the SAS macros, available in the file jackboot.sas at the web site

http://ftp.sas.com/techsup/download/stat/ (8).

For the sake of illustration the aforementioned SAS macros have been utilized to compute the bootstrap CI (α=0.05) for the Spearman correlation coefficient on the data by Dittadi et al (4) and reported in Table I. By applying the BCa bootstrap resampling method to 1000 bootstrap samples, we obtained an approximate CI for r_S equal to 0.65 |—-| 1.00.

APPENDIX

Below we report the SAS commands to be run for computing the confidence interval (α=0.05) of the Spearman correlation coefficient to evaluate the as-

sociation between the vascular endothelial growth factor (VEGF) concentration in lysed whole blood and the platelet count (data in Table I) (4). The bootstrap estimate of the confidence interval is computed by running the SAS macros %BOOT and %BOOTCI to perform the Bias Corrected and Accelerated (BCa) method on 1000 bootstrap samples. The macros are available in the file jackboot.sas, which can be freely downloaded at the web site http://ftp.sas.com/techsup/download/stat/ (8).

```
/* ******************************************************************/
/* Input the data of Table I                                        */
/* ******************************************************************/

data aa; input VEGF PLATELET; cards;
612     376
160     188
531     255
309     206
321     231
262     227
411     233
196     184
450     296
756     261
;
run;


/* ******************************************************************/
/* Include and submit the macros in the file jackboot.sas           */
/* ******************************************************************/

%include 'C:\ijbm\jackboot.sas';  run;


/* ******************************************************************/
/* write and submit the macro %ANALYZE for Spearman correlation     */
/* ******************************************************************/

%macro analyze(data=,out=);
     proc corr spearman noprint data=&data
       outs=&out(where=(_type_='CORR' & _name_='VEGF')
              rename=(PLATELET=corr)
              keep=PLATELET _type_ _name_ &by);
       var VEGF PLATELET;
       %bystmt;
     run;
  %mend;
```

```
/* ******************************************************************/
/* Execute the macro %BOOT                                          */
/* ******************************************************************/

%boot(data=aa,
samples=1000,
residual=,
equation=,
size=,
balanced=,
random=0,
stat=CORR,
id=,
biascorr=1,
alpha=.05,
print=1,
chart=1
)


/* ******************************************************************/
/* Execute the macro %BOOTCI                                        */
/* ******************************************************************/

%bootci(
method=BCa,
stat=CORR,
student=,
id=,
alpha=.05,
print=1
)


/* ******************************************************************/
/* Print the approximate lower (ALCL) and upper (AUCL) limits of the  */
/* BCa confidence interval                                          */
/* ******************************************************************/

proc print data=bootci; var alcl aucl; run;
```

Address for correspondence:
Prof. Ettore Marubini
Unità Operativa di Statistica Medica e Biometria
Istituto Nazionale per lo Studio e la Cura dei Tumori
Via Venezian, 1
20133 Milano, Italy
e-mail: biom@istitutotumori.mi.it

REFERENCES

1. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 1981; 35: 124-32.
2. Kenney JF, Keeping ES. Mathematics of statistics – part two. New York: D. Van Nostrand Company, Inc., 1951.
3. Kramer MS. Clinical epidemiology and biostatistics. Berlin: Springer-Verlag, 1988.
4. Dittadi R, Meo S, Fabris S, Gasparini G, Contri D, Medici M, Gion M. Validation of blood collection procedures for determination of circulating vascular endothelial growth factor (VEGF) in different blood compartments. Int J Biol Markers 2001; 16: 87-96.
5. Bossi A, Cortinovis I, Duca P, Marubini E. Introduzione alla statistica medica. Rome: La Nuova Italia Scientifica, 1992.
6. Efron B, Tibshirana RJ. An introduction to the bootstrap. London: Chapman and Hall, 1993.
7. Carpenter J, Bithell J. Bootstrap confidence interval: when, which, what? A practical guide for medical statisticians. Stat Med 2000; 19: 1141-64.
8. SAS Institute Inc. SAS Campus Drive Care, NC 27513, USA.

Received: January 15, 2002
Accepted: March 1, 2002