South Eur J Orthod Dentofac Res

SEJODR

StaTips Part VI: Bivariate correlation

Perinetti, Giuseppe *

** Private practice, Nocciano (PE), Italy*

ABSTRACT

A very common situation in medical research, including orthodontics, is when a researcher has to verify the association between 2 variables, best referred to as bivariate correlation. Bivariate correlation is an analysis that measures the strength of relationship between two variables through the calculation of different correlation coefficients. The most common correlation coefficients are: Pearson (r), Kendall (rho), Spearman (rho) and the point-biserial (rpb). The choice of the correct coefficient is based on the type of data to be analysed and, for some of them, the existence of assumptions for using parametrical tests. Indications on how to choose the correct coefficient and about their interpretation are provided.

Perinetti G. StaTips Part VI: Bivariate correlation. South Eur J Orthod Dentofac Res. 2019;6(1):2-5.

FRAMING OF THE PROBLEM

A very common situation in medical research, including orthodontics, is when a researcher has to quantify the association between 2 variables, a procedure best referred as bivariate correlation.[1] Such bivariate correlation is the simplest case of analysis of association between 2 variables only, while in case of association among 3 or more variables, it would be more appropriate to use multiple regression models (not dealt herein). A typical example of bivariate correlation would be to verify whether the entity of refereed pain upon application of the orthodontic force is related to the age of the patients (provided that force and other conditions are the same for all the patients).

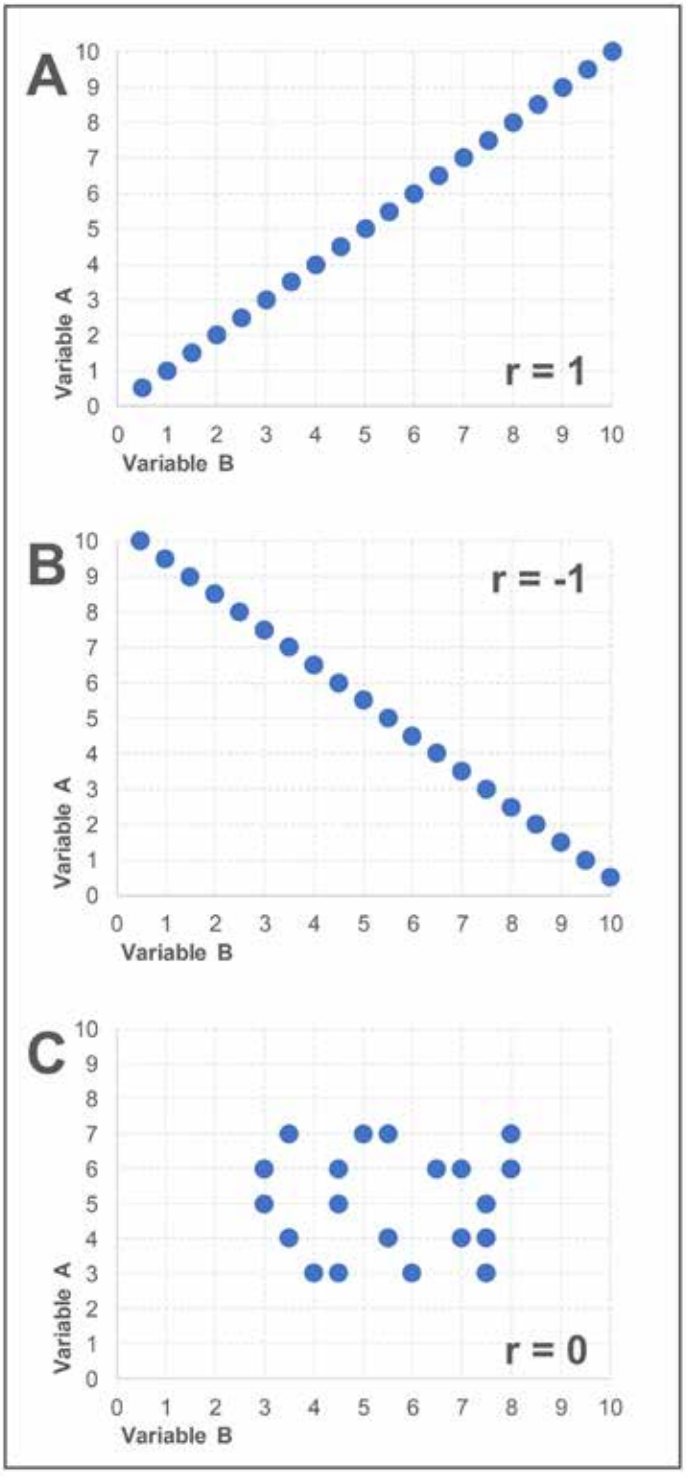
Bivariate correlation explores the association between variables, where the term association refers to any relationship (linear and not linear). Even though the term correlation is often used to refer only to a linear relationship between 2 variables, herein this term will be used in either case. More specifically, bivariate correlation is an analysis that measures the strength of relationship between 2 variables through the calculation of different correlation coefficients. Relationship between 2 variables (linear or not) has two distinct aspects, which are strength and direction, denoted by the absolute value and sign of the correlation coefficient, respectively. Regarding the strength, the value of the correlation coefficients range between -1 and +1. A value of ±1 indicates a perfect relationship between the two variables, closer the correlation coefficient to zero, weaker the relationship will be. Regarding the direction of the relationship, a positive sign indicates a positive relationship (when one variable increases the other does the same), while a negative sign indicates a negative relationship (when one variable increases the other decreases).[1] For each coefficient, it is important to calculate, the corresponding P value for the null hypothesis that the coefficient itself would be equal to zero (i.e. P >0.05, no significant correlation exists). Most of the available statistical packages are able to calculate all of these coefficients and corresponding P values.

It must be pointed out that correlation analysis, thorough its coefficients, is an interdependency measure between 2 variables that does not determine cause and effect. Therefore, when correlation coefficients show a significant relationship between 2 variables, it means only that when there is a systematic change in one variable, there is also a systematic change in the other.

The most common correlation coefficients are: Pearson (r),[2] Kendall (tau),[3] Spearman (rho)[4] and the point-biserial (rpb).[5] The choice of the correct coefficient is based on the type of data to be analysed (continuous, ordinal and dichotomous) as reported below and summarised in Table. Examples of linear relationship between 2 variables with corresponding r correlation coefficients are showed in Figure 1, while the special case of non-linear relationship is detailed below.

Corresponding Author:
Perinetti Giuseppe
Via San Lorenzo 69/1,
65010 Nocciano (PE), Italy.
e-mail: G.Perinetti@yahoo.com

Figure 1. *The correlation coefficient in cases of different directions of relationship.*



*In this example, the Pearson (**r**) correlation coefficient has been used. In **A**, a positive perfect linear relationship between the 2 variables with an r coefficient equal to +1; in **B**, a perfect negative linear relationship between the 2 variables with an r coefficient equal to -1. In **C**, absence of linear relationship between the 2 variables with an r equal to zero.*

PEARSON CORRELATION COEFFICIENT

The r correlation coefficient also referred as Pearson product-moment correlation coefficient is likely the most widely used correlation test to measure the degree of the linear relationship between 2 continuous variables. This r coefficient requires the existence of assumptions for using parametrical methods [6]; therefore, both variables should be normally distributed, and data has to be equally distributed about the regression line (equal variances or homoscedasticity). Moreover, while r coefficient assumes that a linear relationship exists between the two variable, computation of this coefficient may not detect any non-linear relationship. Therefore, it is suggested to initially evaluate the data through graphical displays to explore the nature of associations between the variables.

Of note, the value of the r coefficient is not influenced by the slope of the linear relationship between the 2 variables (Figure 2A and 2B), and it is undetermined in the case one of the 2 variables is constant (Figure 2C). On the contrary, the value of the r coefficient is influenced by the noisiness (i.e. random distribution) of the points around the regression line of the relationship between the variables (Figure 3A and 3B). The smaller the noisiness and the greater the r coefficient, and *vice versa*.

Table. *Most common correlation coefficients according to the nature of the 2 variables under analysis (see text for details).*

Type of data of the 2 variables	Normal distribution and homoscedasticity	Relationship	Correlation coefficient
Both continuous	Required	Linear	Pearson (r)
Both continuous	Not required	Linear or not linear	Kendall (tau) or Spearman (rho)
Both or at least 1 ordinal	Not required (for the eventually present continuosus data set)	Monotomic (linear or not linear)	
One continuous and the other naturally dichotomous	Required (for the continuous variable data set)	-	Point biserial (r_{pb})
One continuous and the other dichotomous but underling a continuity between categories	Required (for the continuous variable data set)	-	Biserial (r_b)

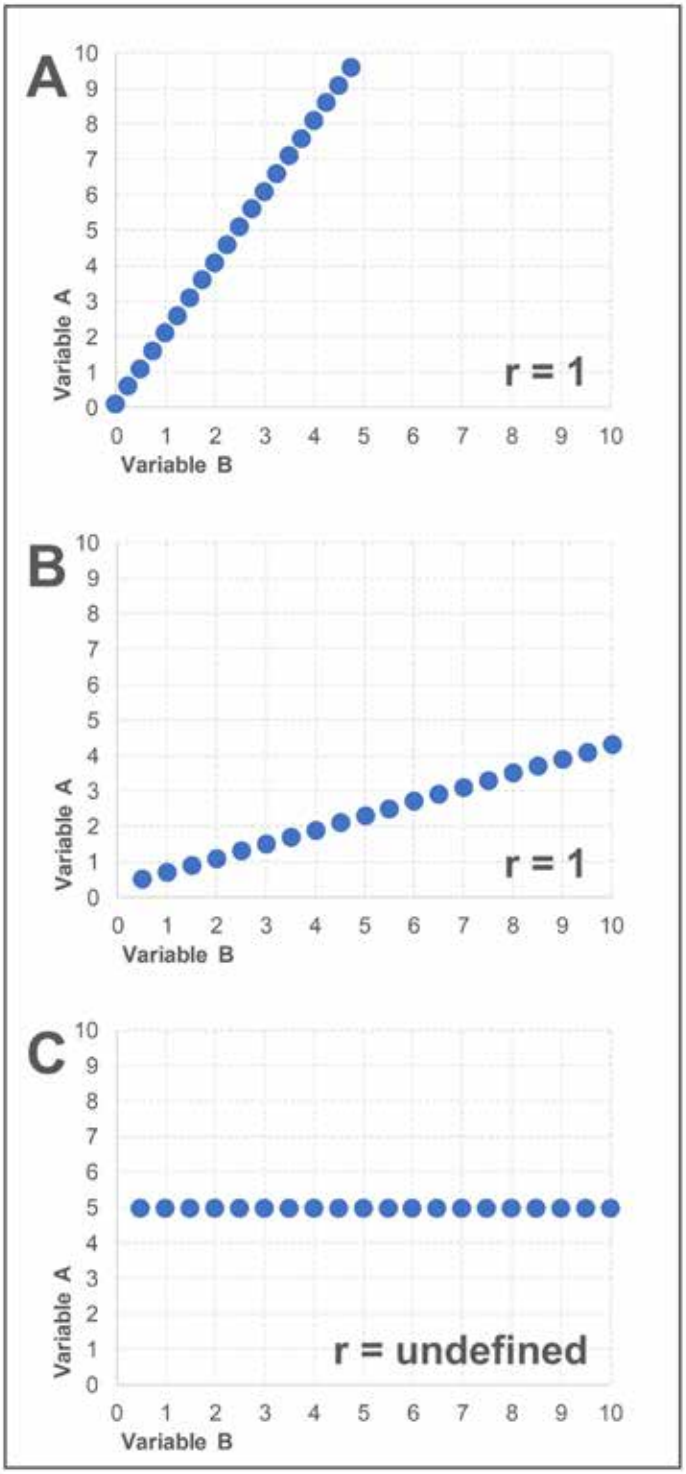
SPEARMAN AND KENDALL CORRELATION COEFFICIENTS

The Kendall (tau) and Spearman (rho) correlation coefficients are non-parametric measures of rank correlation between the rankings of two variables. Therefore, these tests do not require the exists of assumptions for using the r correlation coefficient (or more in general parametric tests), and may be used for ordinal variables, or for continuous data sets failing the assumption of normal distribution and homoscedasticity (Table).

The tau and rho correlation coefficients also do not require a linear relationship between the 2 variables being correlated (as it is for the r coefficient). Indeed, tau and rho coefficients assess monotonic relationships (whether linear or not). A monotonic relationship is a relationship where: 1) as the value

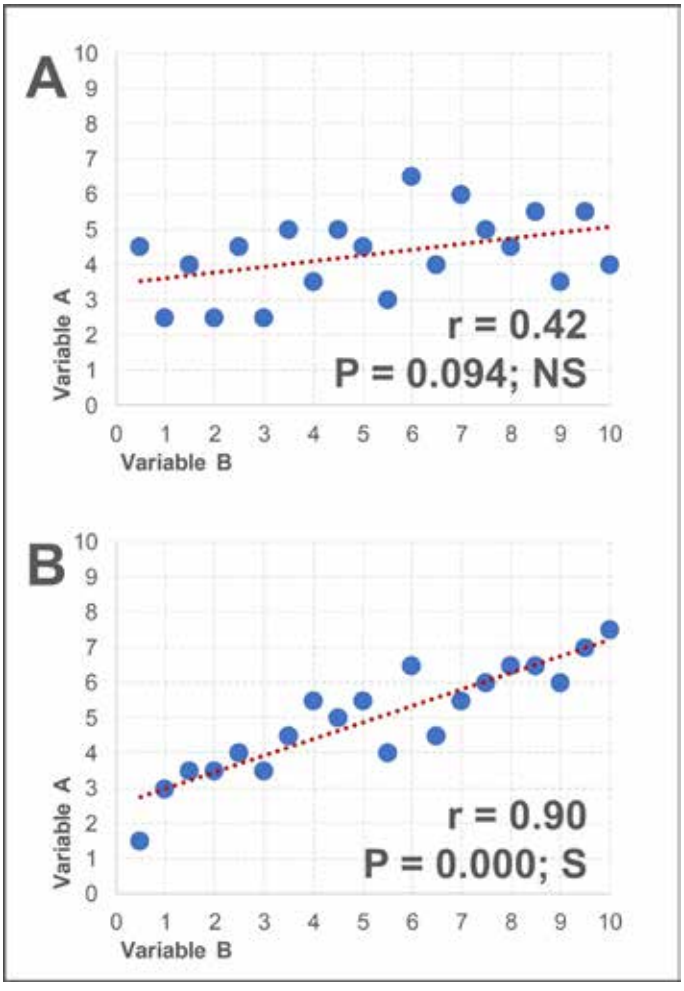
South Eur J Orthod Dentofac Res

Figure 2. *The Pearson correlation coefficient in cases of different slopes of linear relationship.*



*The Pearson (**r**) correlation coefficient is not influenced by the slope of the of the relationship between the 2 variables. Irrespective of the different slopes of linear relationships in **A** and **B**, the r correlation coefficients are the same. In **C**, the special case when one variable is constant. In this case the r correlation coefficient is undefined, and correlation does not exist.*

Figure 3. *The Pearson correlation coefficient in cases of different noisiness of linear relationship.*



*The Pearson (**r**) correlation coefficient and corresponding P value are influenced by the noisiness of the points around the regression line (red) of the of the relationship between the 2 variables. In **A**, a medium relationship with non-significant P value due to a noteworthy noisiness of the points; in **B**, a strong correlation coefficient with a significant P value due to a lower noisiness of the points around the regression line (red). **S**, significant; **NS**, not significant.*

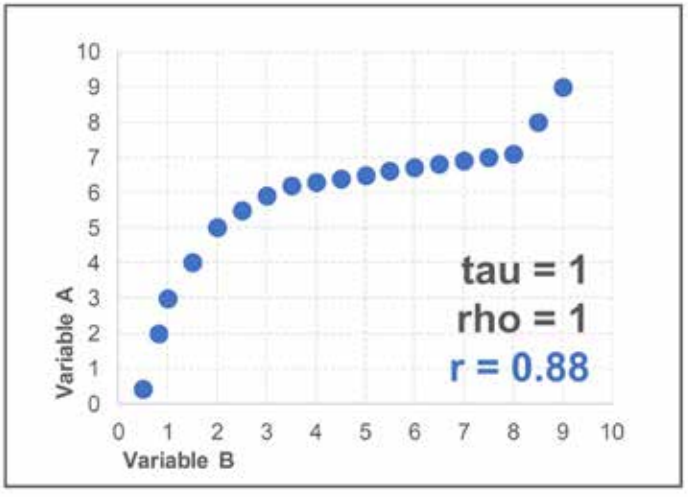
of one variable increases, so does the value of the other variable; or 2) as the value of one variable increases, the other variable value decreases. If there are no repeated data values, a perfect tau or rho coefficient of +1 or –1 occurs when each of the 2 variables is a perfect monotone function of the other (Figure 4).

POINT-BISERIAL CORRELATION COEFFICIENT

The point-biserial correlation (rpb) coefficient is a special case of the r coefficient where one variable is continuous and the other is naturally dichotomous (Figure 5). The categories of the dichotomous variable (nominal binary variables) do not have a natural ordering (i.e. male/female, treated/not treated). The rpb coefficient is calculated as the r coefficient, wherein the dichotomous variable is coded as 0 or 1 (regardless of which of

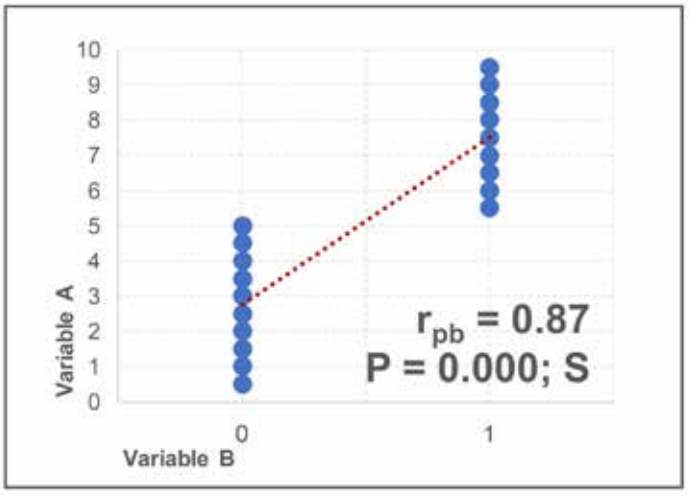
the 2 categories is coded into 0 or 1). Nevertheless, the point-biserial correlation requires the existence of assumption for using parametrical tests (as for the r coefficient) for the continuous data set belonging to each of the 2 categories of the dichotomous variable. These continuous data set by must be normally distributed with homoscedasticity. It is not recommended to artificially recode a continuous or ordinal variable into a dichotomous variable because binary data carries less variance information, making correlation analysis less reliable. A slightly different situation is when the dichotomous variable carries an intrinsic underlying continuity between the 2 categories, as for instance when referring to a successful or unsuccessfully treatment for a given malocclusion. In this case a biserial correlation coefficient (rb) should be used instead of the rpb coefficient.

Figure 4. *The Kendall and Spearman correlation coefficients in case of non-linear monotonic relationship between variables.*



*Irrespective of the non-linear relationship between the variables, the Kendall (**tau**) and Spearman (**rho**) correlation coefficients are equal to 1. On the contrary, the Pearson (**r**) correlation coefficient (in blue) is only 0.88. Even if the variables A and B would be continuous, the r correlation coefficient would not be indicated because of their non-linear relationship.*

Figure 5. *The point biserial correlation coefficient.*



*A strong point biserial (**rpb**) correlation coefficient and corresponding significant P value. In this case, the variable B is naturally dichotomous (i.e. full qualitative, such as a male and female recoded indifferently into zero and 1). **S**, significant.*

GENERAL INTERPRETATION OF THE CORRELATION COEFFICIENTS

As a rule of thumb, meaningful correlations considered clinically relevant are those with a correlation coefficient of at least ±0.4. For a more precise interpretation, Cohen's standard [7] may be followed. Accordingly, coefficients between 0.10 and 0.29 would represent a small correlation, coefficients between 0.30 and 0.49 would be indicative of a medium correlation, and coefficients above 0.49 would represent a large correlation. As stated above, when calculating a correlation coefficient, it is important to include the corresponding P value or confidence interval to ensure that the retrieved value is significantly different from zero.

CONFLICT OF INTEREST

The Author declares no conflict of interest.

REFERENCES

1. Glantz S. Primer of Biostatistics. 7th ed. Columbus: McGraw-Hill Education; 2011. 7th ed. ed. Columbus, OH: McGraw-Hill Education; 2011.
2. Pearson K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. 1895;58:240-42.
3. Kendall M. A New Measure of Rank Correlation. Biometrika. 1938;30:81-89.
4. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15:72-101.
5. Sheskin DJ. Handbook of parametric and nonparametric statistical procedure. 5th ed. ed. Boca Raton, FL: Chapman and Hall/CRC; 2011.
6. Perinetti G. StaTips Part I: Choosing statistical test when dealing with differences. South Eur J Orthod Dentofac Res. 2016;3:4-5.
7. Cohen J. A power primer. Psychol Bull. 1992;112:155-9.