Applied statistics in vascular surgery Part 2: Correlation analysis and its common misconceptions

Constantine N. Antonopoulos, John D. Kakisis

Department of Vascular Surgery, Medical School, National and Kapodistrian University of Athens, Attikon University Hospital, Athens, Greece

Abstract:

Correlation is a statistical technique assessing the association between two quantitative variables. Before performing a correlation analysis, a scatter plot of the variables used should be drawn to check for linearity. For a normal distribution, the *"Pearson correlation coefficient"* is used, with the assumption that the two variables have to be measured on either an interval or ratio scale, outliers should be taken into consideration and the scatter plot should present fair homoscedasticity. When these assumptions are violated, *"Spearman's rank-order correlation coefficient"* is used, which is the nonparametric alternative version of the Pearson correlation test. Both tests' results range from -1.0 to +1.0, where r > 0 indicates a positive association, r < 0 indicates a negative relationship and r = 0 indicates no relationship. Of note, the closer r is to +1 or -1, the more closely the two variables are related. In case of non-linear (curvilinear) correlation, in which the ratio of change is not constant or when the variable's distribution is not normal, the researcher can perform a logarithmic or other type of transformation for one or both variables. Although very simple in use, correlation analysis carries many misconceptions and misuses. When dealing with them, the researcher should be aware that a conclusion about individuals should never be reached based on group-level data and that correlation does not imply causality.

INTRODUCTION

Increase your chocolate consumption and you might win a Nobel Prize?

A very interesting study appeared in 2012 in the distinguished journal "*New England Journal of Medicine*"[1]. The authors collected a list of countries ranked in terms of Nobel laureates per capita from "*Wikipedia*" and data on per capita yearly chocolate consumption from the official website of the "*Association of Swiss Chocolate Manufacturers*". A statistical correlation test was applied thereafter, and, surprisingly, there was a close, significant linear correlation (r = 0.791, P<0.0001) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries. Moreover, the slope of the regression line allowed them to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1. The authors concluded that "chocolate consumption enhances cognitive function, which is a "*sine qua non*" for winning the Nobel Prize". But should we all change our diet and start consuming more chocolate to increase the chances of winning a Nobel Prize?

Author for correspondence:

Constantine N. Antonopoulos

Department of Vascular Surgery, Athens University Medical School, Attikon University Hospital, Athens, Greece
E-mail: kostas.antonopoulos@gmail.com
ISSN 1106-7237/ 2019 Hellenic Society of Vascular and Endovascular Surgery Published by Rotonda Publications

What does correlation mean?

Correlation is a statistical term, indicating how strongly pairs of variables are related and implicates association between two quantitative variables. A common example is height and weight; taller people are usually heavier than shorter people, which highlights a correlation between height and weight in the humankind. A classical example of correlation in vascular surgery is the size of the aneurysm and its % percentage of risk for rupture; bigger aneurysms tend to rupture more often. When applying a correlation test in statistics, the researcher needs to answer two basic questions; 1) whether this relationship is positive or negative and 2) which is the strength of the relationship? A positive correlation indicates that both variables increase or decrease in parallel (Figure 1), whereas in a negative correlation the change between the two variables occurs in opposing directions so that increase in one is followed by decrease in the other (Figure 2)[2]. As a result, we conclude that height and weight have a positive correlation. On the contrary, it has been reported that the incidence of diabetes mellitus is rising and at the same time the incidence of aneurysms is declining[3], underlying a negative correlation.

Measures of correlation; the correlation coefficient

In order to measure the direction and strength of the association between two variables, a statistical estimator should be used, which is generally called "*correlation coefficient*" (r). It ranges from -1.0 to +1.0; r > 0 indicates a positive association, r < 0 indicates a negative relationship and r = 0 indicates no relationship, while the closer r is to +1 or -1, the more closely the two variables are related. As a general rule, r=-1.0 to -0.5 or r=0.5 to 1.0 indicates a strong correlation, r=-0.5 to -0.3 or

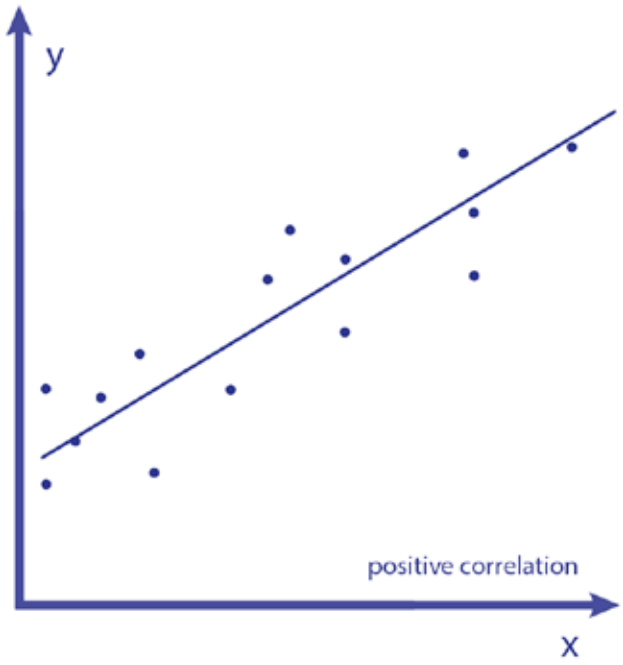

Figure 1.

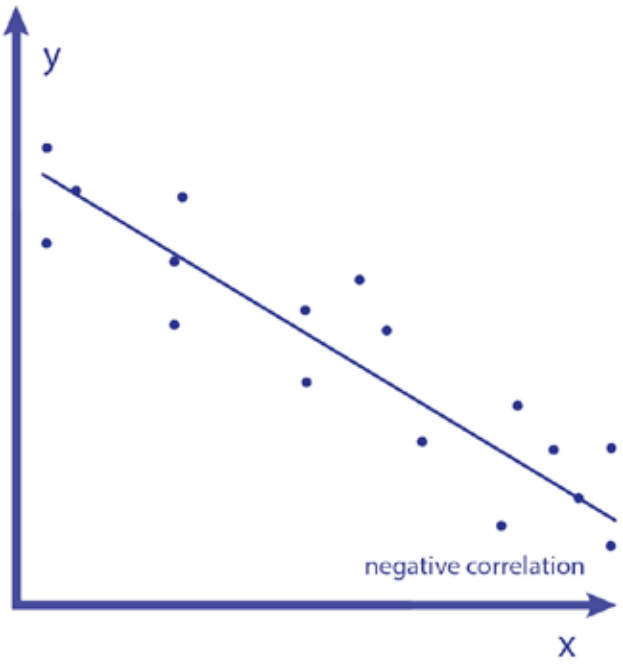

Figure 2.

r=0.3 to 0.5 indicates a moderate correlation, r=-0.3 to -0.1 or r=0.1 to 0.3 indicates a weak correlation, while r=-0.1 to 0.1 indicates a none or a very weak correlation. In any case, p value<0.05 indicates statistical significance, which is the probability that the researcher has found the observed result, or a more extreme one, when the correlation coefficient was in fact zero (null hypothesis). However, someone should be careful, as statistically significant result does not necessarily mean that there is a strong correlation; it simply tests the null hypothesis[2].

Which correlation coefficient should I use?

In cases of normal distribution[4], the "*Pearson product-moment correlation coefficient*", also called "*Pearson correlation coefficient*" estimates the degree to which a relationship is linear. Its main action is to draw a line of best fit through the data of two variables and indicate how far away all these data points are to this line of best fit. However, it does not represent the slope of the line of best fit. Basic assumptions of *Pearson correlation coefficient* include that the two variables have to be measured on either an interval or ratio scale, while they can be measured in entirely different units. Furthermore, the two variables should form a linear relationship, which can be checked by plotting them on a scatterplot and visually inspect its shape. Usually, the one variable is plotted on the x-axis (horizontally) and the other variable is plotted on the y-axis (vertically).

Moreover, outliers should be taken in to consideration, as they might pose a very large effect on the line of best fit and largely affect the estimate of *Pearson correlation coefficient*. Additionally, the plot should present fair homoscedasticity, which means that the variances along the line of best fit should remain almost the same along the line. Another important characteristic is that *Pearson correlation coefficient* does not take into consideration whether a variable has been classified as a dependent or independent variable. Although *Pearson correlation coefficient* is widely used, it is common in some publications to report the *coefficient of determination, r^2*, which is the square of the *Pearson correlation coefficient r* **(i.e., r^2)**. This index represents the proportion of the variance that is shared by both variables and provides a measure of the amount of variation that can be explained by the model.

The nonparametric version of the Pearson product-moment correlation is called "*Spearman's rank-order correlation coefficient*", (ρ, also signified by r_s) and measures the strength and direction of association between two ranked variables. This is an alternative to the Pearson correlation index, which is used when assumptions of the Pearson correlation are violated. Of note, in cases of non-linear (curvilinear) correlation, in which the ratio of change is not constant or when the variable's distribution is non-normal, the researcher can perform a logarithmic or other type of transformation for one or both variables.

Other types of correlation coefficient

In cases of ordinal association between two measured quantities, the "*Kendall rank correlation coefficient*", commonly referred to as "*Kendall's tau coefficient*", is used and investigates the similarity of the orderings of the data when ranked by each of the quantities[5]. When one variable is continuous and the other variable is dichotomous then the "*point-biserial correlation*" should be used. (eg. correlation between a continuous variable, which is the monthly income measured in Euros and a binary variable, which is gender, with males and females as categories)[6]. Additionally, the "*biserial correlation*",

which is different from the "*point-biserial" correlation*", is recommended when the dichotomous (binary) variable has an underlying continuous distribution; for example, if 0=low intelligence quotient (IQ), and 1=high IQ, the researcher should use the biserial and not the point-biserial - correlation.

Major pitfalls in correlation analysis or "*should I consume more chocolate*"?

Although many papers might present data on the correlation between two variables, some important issues require special attention when running a correlation test. First of all, a conclusion about individuals should never be reached based on group-level data. As a result, a correlation coefficient at country level, must not be used to reach a conclusion about the individual level. Therefore, given that no data are known on how much chocolate the Nobel laureates consumed, any conclusions are rather speculative. Another major misinterpretation of correlation is the idea that it implies causality. Correlation only assesses the intensity of association between two variables and never explains the nature of this agreement[7]. The two variables may show a correlation not because they are influenced by each other but because they are both influenced by the same confounder. As a result, chocolate consumption and winning the Nobel Prize do not have a causal relation. In order to point out meaningless correlation, researcher use the term "*nonsense" or "spurious" correlation*" in which "no sensible natural causal interpretation can be provided"[8]. A classification of correlations has been provided by Haig in order to highlight errors in interpreting statistically significant correlations. Consequently, someone has to be very careful when dealing with significant associations.

CONCLUSIONS

The correlation coefficient is a popular measure of the association between two variables and can easily summarize a scatterplot in a single number. Two main estimators are commonly used in research, namely the "*Pearson product-moment correlation"* and its nonparametric alternative version, called "*Spearman's rank-order correlation coefficient"*. Although very simple in implementation and interpretation, researchers should clearly understand the assumptions behind conducting a correlation analysis and explain them in their methods in order to avoid common errors and ecological fallacies.

No conflict of interest.

REFERENCES

1 Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. N Engl J Med. 2012;367(16):1562-4

2 Swinscow TDV. 11. Correlation and regression. Statistics at Square One. 9th ed: BMJ Publishing Group, Revised by M J Campbell, University of Southampton; 1997

3 Antonopoulos CN, Liapis CD. Aneurysms and Diabetes Mellitus: A Strange Symbiosis? Cardiology. 2018;141(2):123-4

4 Antonopoulos C, Kakisis J. Applied statistics in vascular surgery Part 1: Choosing between parametric and non-parametric tests. HJVES. 2019;1(1)

5 David ST, Kendall MG, Stuart A. Some questions of distribution in the theory of rank correlation. Biometrika. 1951;38(1-2):131-40

6 Linacre J. The Expected Value of a Point-Biserial (or Similar) Correlation. Rasch Measurement Transactions. 2008;22(1):1154

7 Veličković V. What Everyone Should Know about Statistical Correlation. American Scientist. 2015;103:26-9.

8 Haig B. D. What Is a Spurious Correlation? Understanding Statistics. 2 (2)2003. p. 125-32