STATISTICS FOR RESEARCHERS

Principles of Correlation Analysis

NJ Gogtay, UM Thatte

Introduction

The field of medicine often requires drawing inferences regarding the association or relationship between two or more variables. In an earlier article on "Measures of Association" we introduced the concept of finding associations [relationships] between two variables that were binary and categorical in nature.[1] Therein, we explored several possible relationships between these binary variables and understood metrics such as absolute risk, relative risk and odds ratio.

In the present article, we discuss how to establish a relationship or an association between two quantitative variables, i.e., variables that can be "measured".[2] As an example, we could perhaps ask the question "Is there a relationship between the number of hours of work put in by a sales representative and the actual sales of a product?" Or "Is there a relationship between maternal age [measured in years] and parity [total number of pregnancies that a woman has carried past 20 weeks of pregnancy]? Correlation analysis helps answer questions such as these.

Definition of Correlation, its Assumptions and the Correlation Coefficient

Correlation, also called as correlation analysis, is a term used to denote the association or relationship between two (or more) quantitative variables. This analysis is fundamentally based on the assumption of a straight –line [linear] relationship between the quantitative variables. Similar to the measures of association for binary variables, it measures the "strength" or the "extent" of an association between the variables and also its direction.

The end result of a correlation analysis is a *Correlation coefficient* whose *v*alues range from -1 to +1. A correlation coefficient of +1 indicates that the two variables are perfectly related in a positive [linear] manner, a correlation coefficient of -1 indicates that two variables are perfectly related in a negative [linear] manner, while a correlation coefficient of zero indicates that there is no linear relationship between the two variables being studied. These are depicted in Figures 1 and 2.

Eyeballing and Analyzing the Data for Correlation - Construction of the Scatter Plot/Scatter Diagram

A correlation analysis begins with the construction of a *scatter plot* or *scatter diagram* [a graphical representation of the data] with one variable on the X-axis and the other on the Y-axis. Let us understand this with an example.

We had carried out a study[3] earlier that evaluated whether two modalities of the informed consent process – the written informed consent process, and the audio visual [AV] recording of this (in the same clinical trial) were different from each other in terms of the extent of understanding of the study by the participant using a pre-validated questionnaire. This questionnaire gave a "total score" [a quantitative measure] at the end of administration. One of the study objectives was to see if there was a relationship between the time (in minutes) taken to administer the consent in the two groups [again a quantitative measure] and the total score. Table 1 gives data on individual participants in both groups for time taken to consent [measured in minutes] and the total

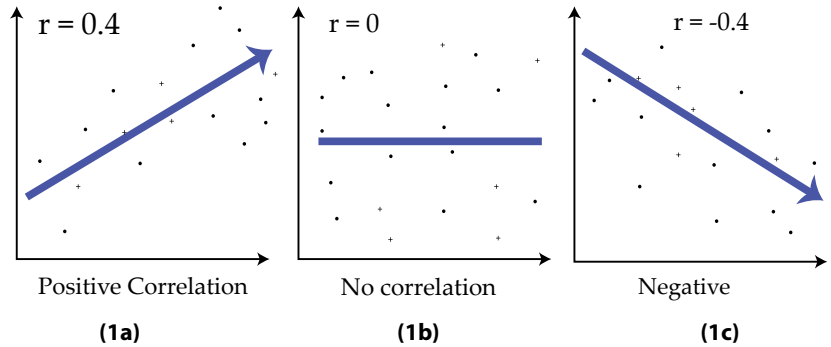


Fig. 1: Scatter Plot showing Correlation between two variables. Note: Fig. 1a shows a weak positive correlation, Fig. 1b shows no correlation and Fig. 1c shows a weak negative correlation

Department of Clinical Pharmacology, Seth GS Medical College & KEM Hospital, Mumbai, Maharashtra
Received: 04.12.2016; Accepted: 10.12.2016

score obtained by the participant [presented as a number].

The *scatter plot* or *scatter diagram* of the total score on the Y axis with the time taken to administer consent on the X axis, enables us to get a feel of the relationship (if any) between the two. Each point on the scatter plot represents the values of X and Y as a *single coordinate*. The closer the points are to a straight line, the stronger is the linear relationship between two variables.

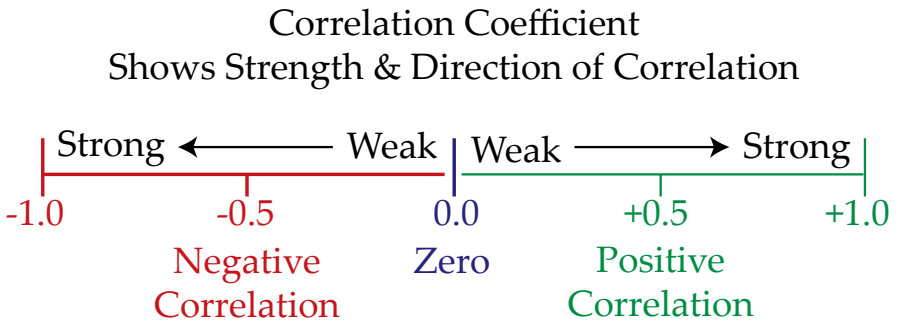


Fig. 2: The spectrum of the correlation coefficient (-1 to +1)

Table 1

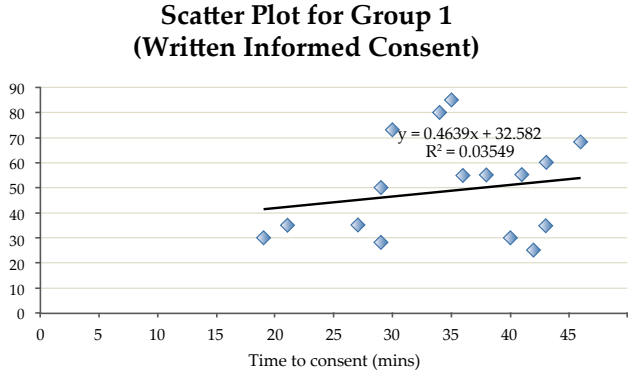
Participant Number	Group 1 Written informed consent [WIC] Total score [n=17]	Time to administer WIC [minutes]	Group 2 AV consent Total Score [n=21]	Time to administer AV consent [minutes]
1	30	73	44	75
2	29	28	37	42
3	42	25	44	20
4	40	30	42	20
5	40	30	46	55
6	43	35	38	90
7	29	50	43	30
8	36	55	38	73
9	38	55	46	104
10	43	60	45	81
11	46	68	44	149
12	41	55	42	60
13	34	80	41	58
14	35	85	39	54
15	27	35	44	120
16	21	35	37	60
17	19	30	38	80
18			43	60
19			23	58
20			45	80
21			27	70

Two scatter plots, one for each group can be easily constructed using Microsoft Excel and those for our example are shown below.

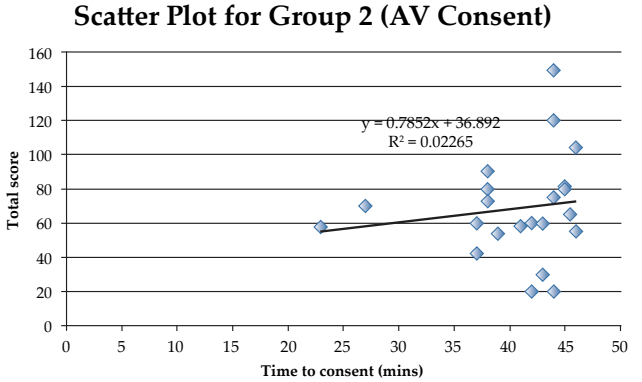
Both scatter plots from our study show a weak, positive, linear relationship between the total scores and the time taken to administer the consent.

The advantage of the scatter plot is that it is simple to construct, is non-mathematical in nature and is unaffected by any extreme values that may be present in the data set. It also tells us immediately if there are outliers or if the relationship is actually non-linear or not entirely linear. A line is usually drawn through the points on a scatter plot to identify linearity in the relationship. This line is called the *regression line* or the *least squares line*, because it is determined such that the sum of the squared distances of all the data points from the line is the lowest possible. This will be discussed in greater detail in the next article on regression analysis.

The disadvantage of a scatter plot is that it does not give us one single value that will help us to understand whether or not there is a correlation between the variables



Scatter plot 1: Written informed consent [Total score *vs.* time to administer consent]



Scatter plot 2: AV consent group [Total score *vs.* time to administer consent]

being studied and hence we need to go a step ahead now to calculate a correlation coefficient.

Calculating the Correlation Coefficients - Karl Pearson's Correlation Co-efficient r and Spearman's Correlation Co-efficient rho (ρ)

A correlation coefficient is that single value or number which establishes a relationship between the two variables being studied. Two methods are used to calculate this value, *viz*. the Karl Pearson's product moment correlation coefficient r or more simply Karl Pearson's correlation coefficient r and the Spearman's rank correlation coefficient rho (ϱ) or Spearman's rho (ϱ) in short.

The Pearson's correlation coefficient establishes a relationship between the two variables based on three assumptions. These are-

a. Relationship is linear
b. Variables are independent of each other
c. Variables are normally distributed.[4]

On the other hand, the Spearman's rho (ϱ) is based on the ranks given to the observations and not on their actual values and is used when the assumptions of the Pearson's coefficient are not met. It can be thus considered as the non-parametric equivalent of the Pearson's coefficient. This is a robust coefficient and can also be used when one of the variables is ordinal[4] in nature. For example, if you want to find the relationship between the weight (measured in kg, continuous, quantitative data) and socioeconomic stratum (ordinal data – higher, middle, lower, etc.) the Spearman rho (ϱ) could be used.

Normality, we know from an earlier article on distributions is commonly tested using the Kolmogorov Smirnov test.[5] In this example, when the variables in the two groups were tested for normality and were found not to follow a normal distribution, we calculated the Spearman's rho (ϱ). The ϱ value obtained in our study for the written informed consent group was 0.2 while that for the AV consent group was 0.15.

Figure 2 describes the interpretation of this correlation coefficient and places the relationship in perspective. In our case, the values of 0.2 and 0.15 indicate a weak positive correlation between the two variables interpreted to mean that the time taken to administer consent is weakly, though positively related to the understanding of consent as assessed by the total scores.

When the relationship or association between more than two quantitative variables is to be studied, other correlation coefficients such as the Sample Multiple Correlation Coefficient can be used

What Correlation Coefficients do NOT do

Correlation coefficients do not give information about whether one variable moves in response to another. There is no attempt to establish one variable as "dependent" and the other as "independent". We shall discuss the concept of independent and dependent variables in the next article on regression analysis. Relationships identified using correlation coefficients should be interpreted for what they are: associations, and **not** causal relationships (see below).

Testing for Significance after Calculating the Correlation Coefficients

Any relationship or association between two variables should be assessed not just for the strength and direction [as given by the correlation coefficients r or ϱ], but also by whether the relationship is "significant" [given by the p value]. Hence testing for significance answers the question "how reliable is the correlation analysis?"

When we calculate correlation coefficients from the given data, what we calculate really are the *sample* correlation coefficients. We now need to apply "tests of significance"[6] to see how close these sample correlation coefficients are to the true population value; i.e., the *population* correlation coefficients. Both the p values obtained in our study were > 0.05 indicating a lack of a significant relationship between the time taken to administer consent and the total score. It is important to remember here that if the sample size is sufficiently large, even small correlation coefficients will achieve statistical significance without being clinically meaningful.

Coefficient of Determination – r^2 [r square]

This is the square of the coefficient of correlation r^2, which is calculated by squaring the value of the "r" obtained. In our study, this would be 0.2 x 0.2 = 0.04 or 4% for the written, informed consent group and 0.15 x 0.15 = 0.02 or 2% for the AV Consent group. This would mean that only 4% and 2% variability respectively in the total score can be accounted for by the time taken to administer the consent.

Correlation and Causation

One common error that often occurs is confusing correlation with causation. All that correlation shows is that the two variables are associated and nothing more. Any judgment regarding cause and effect must be made on the basis of the investigator's knowledge and biological plausibility. This is easily seen in an interesting study by Messerli FH[7] who showed that

greater a country's annual per capita chocolate consumption, more were the number of Nobel Laureates per 10 million population and thus established a "relationship" or "association" between chocolate consumption and getting a Nobel prize!

Factors that Affect a Correlation Analysis

Several factors must be considered when a correlation analysis is planned. These include:

i. Correlation analysis should not be used when data is repeated measures of the same variable from the same individual at the same or varied time points. For example, if you have measured pain scores in patients with Rheumatoid arthritis at monthly intervals over 2 years in a study, it is inappropriate to find out a correlation coefficient for this data.

ii. It is useful to draw a scatter plot as an important pre-requisite to any correlation analysis as it helps eyeball the data for outliers, non-linear relationships and heteroscedasticity

iii. An outlier is essentially an infrequently occurring value in the data set. It is important to remember that even a single outlier can dramatically alter the correlation coefficient.

iv. If there is a non-linear relationship between the quantitative variables, correlation analysis should not be performed. For example, during the growth phase in adolescence, there would a linear relationship between height and weight, as both increase. However, this relationship ceases once a person enters adulthood.

v. If the dataset has two distinct subgroups of individuals whose values for one or both variables differ considerably from each other, a false correlation may be found, when none may exist. An example given by Aggarwal and Ranganathan[8] illustrates this point well. If you were to plot heights (on X-axis) and hemoglobin levels (on Y-axis), of a group of men (n=20) and women (n=20), most women may end up in the left lower corner (shorter and lower hemoglobin) and most men in the right upper corner (taller and higher hemoglobin). Analysis would suggest a relationship with a positive "r" value between height and hemoglobin levels!

vi. The sample size should be appropriately calculated *à priori*.[9] Small sample sizes may show a false positive relationship.

vii. If one data set forms part of the second data set, for example, height at age 12 (X - axis) and height at age 30 (Y-axis) we would expect to find a positive correlation between them because the second quantity "contains" the first quantity.

viii. Heteroscedasticity is a situation in which one variable has unequal variability across the range of values of the second variable. For instance, if one were to plot time on the X-axis and the Sensex on the Y-axis, one would find a great variability in the Sensex as compared to the relative stability in time.

Conclusion

In summary, correlation coefficients are used to assess the strength and direction of the linear relationships between pairs of continuous variables. When both variables are normally distributed we use Pearson's correlation coefficient "r". Otherwise, we use Spearman's correlation coefficient rho (ϱ), which is non–parametric in nature, and is more robust to outliers than is the Pearson's correlation coefficient "r".

Correlation analysis is seldom used alone and is usually accompanied by the regression analysis. The difference between correlation and regression lies in the fact that while a correlation analysis stops with the calculation of the correlation coefficient and perhaps a test of significance, a regression analysis goes ahead to expresses the relationship in the form of an equation and moves into the realm of prediction. The next article in the series will deal with regression analysis.

References

1. Gogtay NJ, Deshpande S, Thatte UM. Measures of Association. *J Assoc Phy Ind* 2016; [in press]
2. Deshpande S, Gogtay NJ, Thatte UM. Data types. *J Assoc Phy Ind* 2016; 64:64-65.
3. Figer BH, Chaturvedi M, Thaker SJ, Gogtay NJ, Thatte UM. A comparative study of the informed consent process with or without audio-visual recording. *Nat Med J Ind* 2017; in press.
4. Deshpande S, Gogtay NJ, Thatte UM. Data types. *J Assoc Phy Ind* 2016; 64:64-5.
5. Gogtay NJ, Deshpande S, Thatte UM. Normal distributions, p values and confidence intervals. *J Assoc Phy Ind* 2016; 64:74-6.
6. Deshpande S, Gogtay NJ, Thatte UM. Which test where? *J Assoc Phy Ind* 2016; 64:64-66.
7. Messerli FH. Chocolate consumption, cognitive function and Nobel Laureates. *N Engl J Med* 2012; 367:1562-4.
8. Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspect Clin Res* 2016; 7:187-90.
9. Gogtay NJ, Thatte UM. Samples and their sizes- the bane of researchers. *J Assoc Phy Ind* 2016; 64:68-71.