JDMS 1:35–39, January/February 1990

Interpretation of the Correlation Coefficient: A Basic Review

RICHARD TAYLOR, EdD, RDCS

A basic consideration in the evaluation of professional medical literature is being able to understand the statistical analysis presented. One of the more frequently reported statistical methods involves correlation analysis where a correlation coefficient is reported representing the degree of linear association between two variables. This article discusses the basic aspects of correlation analysis with examples given from professional journals and focuses on the interpretations and limitations of the correlation coefficient. No attention was given to the actual calculation of this statistical value.

Key words: correlation coefficient, *r* coefficient, regression equation, coefficient of determination.

The review of any medical or scientific journal articles cannot be undertaken without being constantly subjected to the statistical analysis and interpretation of research data. Often the trauma of these mysterious numbers and symbols can be avoided by simply reading the abstract containing the summary and conclusions. Even then, statistical terms and symbols that require at least a minimal understanding of the statistical concepts are frequently reported in the abstract summary. For example, in a journal article about the echocardiographic analysis of prosthetic valve replacements, it was noted in the abstract that valvular gradients correlated with prosthetic size ($r = 0.57$) and were higher ($P < 0.001$) across small (19 to 23 mm) versus large (25 to 31 mm) valves.[1] To fully understand the clinical significance of this research finding would require some knowledge of what the statistical symbols represented and an interpretation of the given statistical values.

Ideally, a good course in basic statistics would be helpful for all sonographers. This is particularly relevant to those who are consumers of the published literature and research in the various specialties. However, this is not always possible; publications like the recent *JDMS* article by Khamis[2] give some valuable help and insight into the statistical puzzle. This article only touched on one piece of the puzzle as the author masterfully explained the meaning of the *P* value and its relationship to the test hypothesis. It is my hope that another part of the puzzle can be added to help understand more fully the statistical presentations often encountered in our professional journals.

From the Department of Cardiology, Logan General Hospital, Logan, West Virginia.

Correspondence: Richard Taylor, EdD, RDCS, Cardiac Laboratory, Logan General Hospital, Logan, WV 25601.

Correlation analysis is one of the most widely used and reported statistical methods in summarizing medical and scientific research data. In this article the basic aspects of correlation analysis will be reviewed with emphasis placed upon the interpretations and limitations of the correlation coefficient. No focus will be given to the actual calculation of this statistical value.

It is often useful to determine if a relationship exists between two different variables. If so, how significant or how strong is this association between the two variables? For example, is there a relationship between the years of service as a sonographer and scores achieved on the registry examination? The correlation coefficient or *r* coefficient is a statistic used to measure the degree or strength of this type of relationship. As previously mentioned, this important statistic is reported extensively in the health science journals. The following examples serve to illustrate this point: "A correlation coefficient of 0.94 was noted between the Doppler-derived transaortic gradient and the catheterization-derived transaortic gradient in the evaluation of 30 adult patients with aortic stenosis."[3] The reported correlation ($r = 0.92$) between the echocardiographically and hemodynamically derived mitral valve areas was statistically significant.[4] In these examples, statistical correlation analysis was used to determine the strength of the association between clinical data which was derived noninvasively (Doppler echocardiography) compared with invasively derived data (catheterization) to evaluate valvular stenosis. The clinical feasibility of a strong correlation between these two sets of data or variables should be obvious. Intuition and empirical observation may indicate that certain variables are linearly related, but in its most basic sense the coefficient of correlation measures the degree to which the two variables are related.[5]

The correlation coefficient is often referred to as Pearson's product-moment *r* or *r* coefficient.[6] The correlation *r* value requires both a magnitude and a direction of either positive or negative. It may take on a range of values from −1 to 0 to +1, where the values are absolute and nondimensional with no units involved. A correlation coefficient of zero indicates that no association exists between the measured variables. The closer the *r* coefficient approaches ± 1, regardless of the direction, the stronger is the existing association indicating a more linear relationship between the two variables. The strength of the correlation is not dependent on the direction or the sign. Thus, $r = 0.90$ and $r = -0.90$ are equal in the degree of association of the measured variables. A positive correlation coefficient indicates that an increase in the first variable would correspond to an increase in the second variable, thus implying a direct relationship between the variables. A negative correlation indicates an inverse relationship whereas one variable increases, the second variable decreases.

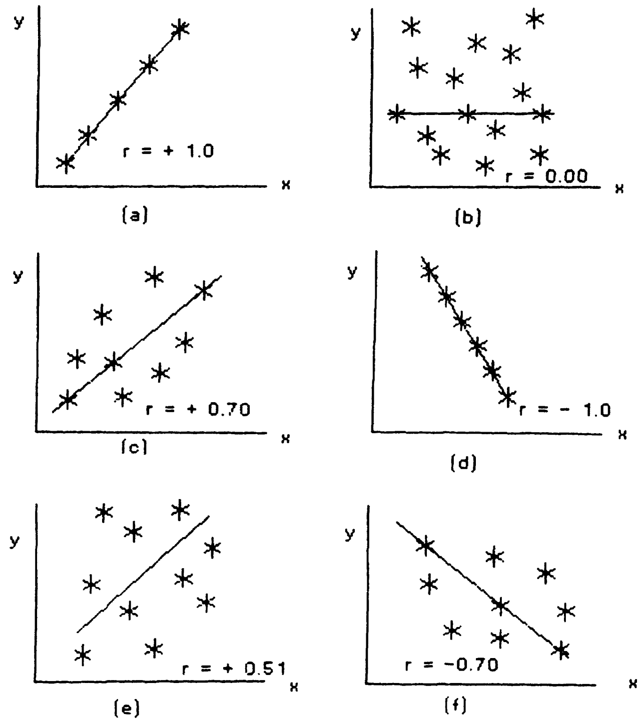

FIG. 1A through F. Examples of various values of *r*. Each graph illustrates the correlation indicated by the specific *r*-value equation shown.

A graph can be useful to illustrate the concept of correlation and to visualize the relationship which exists between the variables. For illustration purposes, two variables are labeled as variable *x* and variable *y* and plotted on a graph. Figure 1 illustrates some different sets of data and how they are summarized by a correlation coefficient value or *r* coefficient. Note that in real-world situations, *x* and *y* would represent the two variables being statistically analyzed such as high school GPA versus SAT scores, or the level of blood serum cholesterol versus heart disease risk to determine the degree of the relationship.

Figure 1a illustrates a perfect positive correlation of $r = 1.0$. It can be noted that all data observations fall on a line, thereby the term *perfect linear correlation* becomes appropriate. The correlation is positive in direction because as variable x increases, variable y varies in the same direction. The relationship of the variables in Figure 1d is also a perfect correlation, but negative in direction ($r = -1.0$). All the data observations still lie on a single line, but as variable x increases, variable y decreases. An example of a negative correlation might involve the number of pull-ups performed relative to the percentage of body fat whereas with body fat percentage increases, a decrease in the number of pull-ups is observed. "In real life, there are always random variations in our observations; hence, a perfect linear relationship is extremely rare."[6] Although the r coefficient value is no longer perfect, the correlation remains higher as the data observations fall closer to a straight line (Fig. 1c) and the coefficient value decreases as the data points deviate more from the straight line (Fig. 1e). If there is no linear relationship between the variables, r will be virtually zero and the data points on the corresponding graph will be randomly scattered and approximate a circle (Fig. 1b). It is important to understand that it is possible to obtain a nonzero value for r even when no correlation actually exists.[5] Also, good to high correlations exist even though they are less than a perfect 1.0. Now what about the statistical significance of those correlation coefficients other than zero and $r = 1.0$?

Like any statistical value, the correlation coefficient is of little importance unless it can be properly interpreted. Like all scale values, the correlation coefficient is difficult to interpret. Labeling systems exist to roughly categorize r values where correlation coefficients (in absolute value) which are ≤ 0.35 are generally considered to represent low or weak correlations, 0.36 to 0.67 modest or moderate correlations, and 0.68 to 1.0 strong or high correlations with r coefficients ≥ 0.90 very high correlations.[5,7] However, merely describing a correlation coefficient of $r = 0.55$ as a moderate correlation is not meaningful.

A basic question which needs to be answered relates to the statistical significance of the correlation coefficient and the random chance of observing a given value of r when, in fact, no real correlation exists. Statistical tables exist which define what r coefficients must be observed before the correlation is said to be statistically significant.[5] The critical values for correlation statistical significance vary as to the sample size used and the level of significance. If it could be assumed that the observed correlation of $r = 0.55$ mentioned previously came from 35 interval-scaled paired observations which were obtained randomly, and that both x and y variables are normally distributed, then it could be concluded that variables x and y represent a correlation coefficient ($r = 0.55$) which is significantly ($P < .01$) different from zero. In other words, the relationship existing between these variables is statistically significant. However, to add to the difficulty of correlation interpretations, a statistically significant correlation coefficient is not necessarily an important one. A statistically significant r coefficient merely indicates that the observed sample data provides ample evidence to reject the null hypothesis that the population correlation coefficient parameter (*rho*) is zero thereby concluding that the population correlation coefficient is not equal to zero. For example, given a large sample size ($n > 100$), a correlation coefficient as small as $r = 0.20$ can be significantly different from zero at $\alpha = 0.05$. This degree of linear correlation would have little practical importance as we shall see later.

It can be seen that the correlation coefficient is an abstract measure and not given to a direct precise interpretation. It can be said that the higher the absolute value of the correlation coefficient, the stronger the relationship. Although the correlation coefficient is the best known and subject to statistical testing, perhaps the *coefficient of determination* is more meaningful.[8] The coefficient of determination can be used to more fully interpret r and is obtained by simply squaring the correlation coefficient r. The coefficient of determination (r^2) is defined as the percent of the variation in the values of the dependent variable (y) that can be "explained" by variations in the value of the independent variable (x).[5,7,8]

This technique results in a percent value which makes it easier to interpret more precisely. Thus, if a correlation coefficient of $r = 0.20$ was observed between variable x and variable y, then the coefficient of determination is $r^2 = 0.04$. This means that only 4% of the total variation in variable y can be explained or accounted for by variation in variable x. Therefore, even though the $r = 0.20$ was statistically significant at $\alpha = 0.05$ with a large sample in the example noted above, it can be seen that only 4% of the total variation of variable y

can be explained by variation in x. The coefficient of determination technique is a more conservative measure of the relationship between the two variables and is preferred by many statisticians, but is seldom reported in research data statistical analyses.

As is true of all statistical methods and procedures, correlation analysis is subject to limitations and misinterpretations that can be serious. In addition to the limitations previously noted, it must be understood that although it measures how closely the variable points approximate a straight line, it does not validly measure the strength of a nonlinear relationship.[6] Spurious or accidental associations between variables may also exist. Browner and Newman[9] wrote that, "It is a mistake to believe a research hypothesis just because the P value is statistically significant." This is particularly true when low prior probability makes a particular association unrealistic or "unsuspected." They noted that the finding of a significant P value dealing with a correlation between coffee drinking and pancreatic cancer ($P < .05$) did not establish the truth of the research hypothesis; subsequent studies failed to confirm the association. Research problems such as data contamination, lab error, sample bias, or poor research design could also cause problems with reliable conclusions.

One of the most frequent and serious misuses of correlation analysis is to interpret a high correlation between variables as a cause-and-effect relationship.[6,8] Correlation analysis measures a relationship or association; it does not define the explanation or its basis. For instance, there is a significant association between a child's foot size and handwriting ability, but it might be presumptuous to claim a large foot causes better handwriting.[6] Statistics do not lie, but they sometimes lead us to reaching false conclusions. Caution must be exercised to avoid this pitfall. Statistical data might indicate that 99.9% of all people who died of cancer drank some water within the previous month. The data speaks for itself, but it would be easy to be deceived into believing that a cause-and-effect relationship, however ridiculous, exists here.

In correlation analysis, the purpose is to measure the closeness of the linear relationship between the defined variables. The correlation coefficient indicates how closely the data fit a linear pattern. Generally, correlation analysis also includes further investigation into defining the pattern of the existing relationship. This procedure is known as *regression analysis*. A mathematical equation is developed for the line of best fit representing the data. From this regression equation, prediction becomes possible where either variable can be predicted based on a value of the other variable. Predicting unknown values (dependent variable) from given values (independent variable) is common and widely used in medical science as well as business forecasting, economics, and education. For example, a research study abstract reported that the correlation between Doppler pressure half-time measurements for mitral valve area relative to catheterization measurements was good ($r = 0.85$, $y = 84x + 0.17$).[10] Upon interpreting this correlation analysis, we see a statistically significant linear relationship ($r = 0.85$, $P < .001$) which exists between the Doppler evaluation of mitral valve area and catheter-derived mitral valve areas. The r coefficient would fall into a general category label of "high" and would yield a coefficient of determination (r^2) or 0.72 or 72%, meaning that 72% of the variability noted in catheterization-derived mitral valve area measurements could be accounted for by Doppler pressure half-time method. Therefore, it appears we have a good association here of clinical usefulness. Finally, the pattern of the linear relationships between Doppler compared with catheter measurements for mitral area assessment is defined by the regression equation $y = 0.84x + 0.17$ where x = Doppler pressure half-time mitral valve area and y = catheterization-derived mitral valve area. Therefore, if we determine the value for Doppler pressure half-time mitral area to be 1.2 cm^2, then the catheterization-derived area (y) could be predicted to be 1.18 cm^2 [0.84 (1.2) + 0.17]. Correlation is important here also because the closer the data "fits" the line, that is the higher the correlation coefficient, the better the predictions become as to reducing the potential errors. It should be noted that although correlation analysis often routinely includes regression analysis in the "package," it is possible to focus on either correlation coefficients or regression equations independently.

Although beyond the scope of this article, different formulas exist to calculate the coefficient of correlation depending upon the nature of the variables and samples. Computer programs are available to routinely perform this task. Regardless of the technique or formula used, the interpretation of the correlation coefficient is basically the same and is generally left to the research consumer.

CONCLUSION

The ability to interpret research reports and professional literature becomes hampered without a basic understanding of statistics. This article should shed some light onto a widely used statistical procedure known as *correlation analysis*. The purpose of this statistical method is to give us a statistic known as the *correlation coefficient* which is a summary value of a large set of data representing the degree of linear association between two measured variables. This statistic serves to reduce the large amounts of data down to a manageable form for sonographers to review. For this goal to be realized, sonographers must understand what the statistical correlation coefficient represents and what it means.

ACKNOWLEDGMENTS

The author wishes to thank Alisa Bentley, Susan Chafin, Betty Ellis, Anna Marshall, RN, and Brady Toler for their gracious assistance in the preparation of this manuscript.

REFERENCES

1. Goldrath N, Zimes R, Vered Z: Analysis of Doppler-obtained velocity curves in functional evaluation of mechanical prosthetic valves in the mitral and aortic positions. *J Am Soc Echo*, 1988;1:211–225.
2. Khamis JF: Statistics referesher II: Choice of sample size. *JDMS* 1989;4:176–183.
3. Stamm BR, Martin RP: Quantification of pressure gradients across stenotic valves by Doppler ultrasound. *J Am Coll Cardiol* 1983;2:707–718.
4. Martin RP, Rakowski H, Kleiman JH, et al: Reliability and reproducibility of two-dimensional echocardiographic measurement of the stenotic mitral valve orifice area. *Am J Cardiol* 1979;43:560–568.
5. Weber JC, Lamb DR: *Statistics and Research in Physical Education*. St. Louis: CV Mosby Co, 1970, pp 59–64, 222.
6. Kuma JW: *Basic Statistics for the Health Sciences*. Palo Alto: Mayfield Publishing Co., 1984, pp 158–169.
7. Mason RO, Lind DA, Marchal WG: *Statistics: An Introduction*. New York: Harcourt Brace Jovanovich, Inc, 1983, pp 368–383.
8. Congelosi VE, Taylor PE, Rice PF: *Basic Statistics: A Real World Approach*, ed 3. St. Paul: West Publishing Co, 1983, pp 315–336.
9. Browner WS, Newman TB: Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459–2463.
10. Smith MD, Handshoe R, Handshoe S, et al: Comparative accuracy of two-dimensional echocardiography and Doppler pressure half-time methods in assessing severity in mitral stenosis in patients with and without prior commissurotomy. *Circulation* 1986;73:100–107.