Reasons for Teaching and Using the Signed Coefficient of Determination Instead of the Correlation Coefficient

by John N. Zorich, Jr. www.JohnZorich.com

THIS IS THE MSWORD DOCUMENT FROM WHICH THE *JOURNAL OF THE AMERICAN MATHEMATICAL ASSOCIATION OF TWO YEAR COLLEGES* CREATED THE PUBLISHED 2018 VERSION.
THE 1ST PAGE OF THE AS-PUBLISHED VERSION CAN BE FOUND AT THE END OF THIS DOCUMENT;
THE FULL AS-PUBLISHED VERSION, PAGES 48-51+58, CAN BE FOUND STARTING AT:
https://amatyc.site-ym.com/page/EducatorSummer2018

Does the simple linear correlation coefficient, typically symbolized by r or r_{XY}, have meaning that is mathematically rigorous? In other words, what does it mean when one r value is twice as large as another? Does it mean "twice as much correlation?" Does r measure or quantify the "strength" of correlation? Can any of the definitions or formulas for the correlation coefficient help answer these questions?

Correlation Coefficient

Mathematical correlation and the simple linear correlation coefficient became important parts of statistical method after their discovery in 1888 by Francis Galton (Walker, 1929). He defined them by saying that "two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other" (Galton, 1888, p. 135); he provided no mathematically rigorous meaning (in the sense provided in the introduction) for what he called his "index of co-relation" (p. 143), which he symbolized by r, other than to say that it "measures the closeness of co-relation" (p. 145). As early as 1892, r came to be called the "coefficient of correlation" (Edgeworth, 1892, p. 191) and was then viewed (incorrectly) as a "proportion" (Edgeworth, 1893, p. 674).

The following is a chronological list of examples of definitions that have been given for r since Galton's time; none of these definitions help answer the questions posed previously:

- "r measures the correspondence between deviations from their means of the two series of observations" (Bowley, 1901, p. 320),
- "[r is the] amount of dependence one variable has upon another" (Baten, 1938, p. 170),
- "[r is the] degree of association between two variables" (Duncan, 1986, p. 820),
- "[r is the] extent to which the scattergraph of the relationship between two variables fits a straight line" (Miles & Shevlin, 2001, p. 20), and
- "[r] measures the strength … of the linear relationship between two variables" (Bluman, 2015, p. 543).

Formulas for the Correlation Coefficient

There are many strikingly different looking but mathematically identical formulas for the correlation coefficient (Symonds, 1926; Rogers & Nicewander, 1988). Although some of them are useful for introducing novice students to the general concept of correlation (Zorich, 2017), none of them help answer the questions initially posed. The following is a sampling of such formulas.

In 1896, the first "basic formula for estimating the correlation coefficient had finally been presented" (Stigler, 1986, p. 343); its presenter was Karl Pearson. That is why the correlation coefficient has most often been defined as "Pearson's *r*"*:*

$$\text{Formula 1: Pearson's } r = \frac{\text{covariance of } X, Y}{\sqrt{(\text{variance of } X)(\text{variance of } Y)}}.$$

Formulas 2, 3, and 4 have also been used to define the correlation coefficient. For example, Formula 2 was used by Yule in 1910 (p. 538), Formula 3 by Ezekiel in 1930 (p. 118), and Formula 4 by Dixon and Massey in 1969 (p. 203). The meanings of symbols used in these formulas are:

$b_1 =$ the slope of the linear regression of *Y* on *X*,

$b_2 =$ the slope of the linear regression of *X* on *Y*,

$S_X =$ standard deviation of the *X* values,

$S_Y =$ standard deviation of the *Y* values,

$S_{Ye} =$ the standard deviation of the values derived from the linear regression equation at each *X* value in an *X*, *Y* data set (i.e., $Y_e = a + bX$), and

r (in Formulas 2 and 3) takes its sign from the slopes of the linear regressions.

$$\text{Formula 2:} \quad r = \sqrt{b_1 b_2}$$

$$\text{Formula 3:} \quad r = \frac{S_{Ye}}{S_Y}$$

$$\text{Formula 4:} \quad r = b_1 \left(\frac{S_X}{S_Y} \right)$$

Based upon Formula 3, it is tempting to describe the correlation coefficient as the *proportion* of the total variation (measured in units of standard deviation) that can be explained by the linear dependence of *Y* on *X*. In fact, that is exactly how it was described in what may have been the first book ever published solely on correlation: "[Formula 3] is then a measure of ... the amount of *correlation*.... The [correlation] coefficient is simply a measure of how large the variation in the estimated values is, in proportion to the variation in the original values" (Ezekiel, 1930, pp. 118–119). Ezekiel eventually

realized that it is mathematically invalid to consider a ratio of standard deviations to be a proportion, as evidenced by the fact that the third edition of his book defined "the proportion of variation in *Y* accounted for by *X*" (Ezekiel & Fox, 1959, p. 127) as a ratio of *variances* (rather than standard deviations), and stated that the "square root of this proportion [rather than the proportion itself] … is termed the *coefficient of correlation*" (p. 127).

No formula and/or definition for *r* itself has ever been used to provide a mathematically rigorous explanation of what it means when one *r* value is twice that of another. Even lengthy philosophical and technical discussions failed in this regard (e.g., 27 pages in Pearson (1911, pp. 152–178), and 115 pages in Yule (1912, pp. 157–253, 317–334)). As is generally well known, the only way to provide such an explanation is to transform each *r* value into its respective *coefficient of determination*.

Coefficient of Determination

The coefficient of determination has been symbolized in various sources by r^2, r_{XY}^2, R^2, or R_{XY}^2. It equals the square of the correlation coefficient and is therefore commonly referred to as "*R*-squared." The earliest known reference to it is from Wright (1921): "Another coefficient which it will be convenient to use, the *coefficient of determination* [emphasis added] of *X* by *A*, ... measures the *fraction* [emphasis added] of complete determination for which factor *A* is directly responsible" (p. 562).

Almost every textbook that discusses r^2 describes it as a *better* statistic than *r* for quantifying the explainable proportion of variation in the *Y* values in an *X*, *Y* data set. The following chronological list contains examples of definitions that have been given for r^2 since Wright's time; they are mathematically rigorous because of their valid use of the words *percentage* and *proportion*:

- "[r^2] may be said to measure the per cent [sic] to which the variance in *Y* is determined by *X*, since it measures that proportion of all the elements of variance in *Y* which are also present in *X*." (Ezekiel, 1930, p. 120);
- "[r^2 is the] proportion of the variance of *Y* that can be attributed to its linear regression on *X*" (Snedecor & Cochran, 1967, p. 176);
- "[r^2 is the] percentage of the total variation in *Y* which can be attributed to the linear relationship with *X*" (Bohrnstedt & Knoke, 1988, p. 270); and
- "[r^2 is the] proportion of the variability in *Y* ... predicted by the relationship with *X*" (Gravetter & Wallnau, 2000, p. 565).

Formulas for the Coefficient of Determination

The square of any of the formulas for the correlation coefficient could be used to calculate the coefficient of determination. For example, using Formula 3, we can derive

$$\text{Formula 5: } r^2 = \frac{S_{Ye}^2}{S_Y^2} = \frac{V_{Ye}}{V_Y} = \frac{(V_Y - V_{Yu})}{V_Y},$$

where V_{Yu} is defined by Formula 6, and V_{Ye} is derived via Formula 7.

$$\text{Formula 6: } V_{Yu} = \sum \frac{(Y - Y_e)^2}{(n-1)}$$

$$\text{Formula 7: } V_Y = V_{Ye} + V_{Yu}$$

In Formula 7, the total variation of *Y* is broken down arithmetically into two components, one of which (V_{Ye}) represents the variation due to the linear relationship between *X* and *Y*, and the other of which (V_{Yu}) represents the remainder of the variation (variation that is *un*explained by, or is not due to, a linear relationship between *X* and *Y*). Thus, these formulas help provide a mathematically rigorous meaning for r^2, namely that it is a decimal fraction between 0 and 1, a fraction whose numerator is the amount of *Y* variation caused by the linear regression dependence of *Y* on *X* (as measured by V_{Ye}), and whose denominator is the total variation of *Y* (as measured by V_Y). Therefore, a linear coefficient of determination whose value is exactly twice that of another means that exactly twice as much of the variation in *Y* is due to its linear correlation with *X*. If we apply the mathematical rigor of the previous example to the correlation coefficient, it will not be valid. This is emphasized in the warnings which follow.

Warnings

The following chronological list contains examples of warnings given about *r* since 1921:

- "For many purposes it is enough to look on it [the correlation coefficient] as giving an *arbitrary scale* [emphasis added] between +1 for perfect positive correlation, 0 for no correlation, and −1 for perfect negative correlation" (Wright, 1921, p. 558);
- "A person who has no knowledge of statistics might easily be led to the *erroneous* [emphasis added] idea that a correlation of *r* = 0.80 is 'twice is good' as a correlation of *r* = 0.40, or that a correlation of *r* = 0.75 is 'three times as good' or 'three times as strong' as a correlation of *r* = 0.25" (Freund, 1960, p. 333);
- "The correlation coefficient *r* is not a proportion and one cannot talk about one correlation coefficient being twice another, nor about one correlation coefficient being 0.2 more than another; *its scale must be regarded as ordinal* [emphasis added]" (Selkirk, 1981, p. 17); and
- "While a [correlation coefficient] value of 0.00 indicates no linear relationship and a value of ±1.00 indicates a perfect linear relationship, *values between these extremes have no direct interpretation* [emphasis added]" (Healey, 1984, p. 267).

History

The practice of reporting correlation as *r* rather than r^2 originated with Galton's biological report in 1888. Unfortunately, those who first applied correlation to other fields continued that practice.

Most prominent among such people was Karl Pearson, who from "the mid-1890s to the First World War ... dominated statistical theory in Britain" (MacKenzie, 1981, p. 10). In Pearson's 1896 article that introduced the formula for Pearson's *r*, he repeatedly compared *r* values (not r^2 values) from different data sets, thereby giving the reader the erroneous impression that *magnitudes* of *r* values can be compared directly (i.e., without first converting them into r^2 values). The rest of his early statistical papers in which correlation figured prominently were similarly *r* rather than r^2 focused (E. S. Pearson, 1956).

Another early major contributor in this field was R. A. Fisher. "His books—notably the many editions of *Statistical Methods for Research Workers* … have become classics" (MacKenzie, 1981, p. 183).

In the 1925 first edition of that book, in the 38-page chapter entitled "The Correlation Coefficient," Fisher explains the meaning of the coefficient of determination; but he does so in only one sentence. In it, ρ refers to the population correlation coefficient and ρ^2 to the population coefficient of determination: "Of the total variance of *y*[,] the fraction $\left(1-\rho^2\right)$ is independent of *x*, while the remaining fraction, ρ^2, is determined by, or calculable from, the value of *x*" (Fisher, 1925, p. 145). In the 1958 thirteenth and final edition of the same-titled book, he included that same one sentence (Fisher, 1958, p. 182).

That sentence was not intended by Fisher to urge the reader to use r^2 (i.e., the sample statistic corresponding to ρ^2) rather than *r* as a measure of correlation strength, as evidenced by the fact that in neither the first nor thirteenth edition of his book was there any subsequent discussion or even mention of the meaning or interpretation of r^2 or ρ^2. Instead, he focused on *r*, the correlation coefficient; he thereby led many of his readers to erroneously conclude that magnitudes of correlation coefficients can be compared to each other on a linear scale ("on a conventional scale," as he described it (Fisher, 1925, pp. 153–154; 1958, p. 190).

Practical Uses for the Correlation Coefficient

There are some practical applications for which the correlation coefficient is more appropriate than the coefficient of determination. For example:

- In the Finance industry, "beta" is a measure of an investment's volatility versus a benchmark (such as the S&P 500). This next formula is a "common expression for beta" (Wikipedia, 2017):

Formula 10: $\text{beta} = r\left(\frac{S_A}{S_B}\right)$, where

$S_A =$ standard deviation of the daily % return from an investment (e.g., a stock) over a given time period and

$S_B =$ standard deviation of the benchmark's daily % return over that same time period.

- In Materials Science, one goal of mechanical product design is to ensure that the distribution of product strengths does not overlap the distribution of anticipated stresses. Statistical analysis of that overlap requires calculation of the standard deviation of the strength–stress distribution; the calculation is given by Dovich (1990, p. 58):

Formula 11: $S_{(X-Y)} = \sqrt{(S_X)^2 + (S_Y)^2 - 2r(S_X)(S_Y)}$,

where *S* stands for the standard deviation of the indicated variable or difference, *X* and *Y* refer to the strength and stress variables respectively, and *r* is the correlation coefficient between *X*, *Y* values paired by individual on-test units of product.

Conclusion

Wherever the correlation coefficient might be used as a measure of strength and an indicator of direction of simple linear correlation, it would be better to use a signed version of the coefficient of determination (the sign being the same as for the correlation coefficient). It is proposed that such a value be symbolized by ${}_S R^2$ and be called the "signed-*R*-squared" or "signed coefficient of determination." One formula that provides both its magnitude and sign, where *r* is the correlation coefficient, is:

Formula 12: ${}_S R^2 = \frac{r^3}{|r|}$.

As discussed previously, it would be misleading to think of *r* as a proportion or to give it a direct interpretation. As a scale, it is arbitrary and primarily ordinal in nature. Therefore, it would be better if textbooks and instructors taught measurement of the strength of simple linear correlation only in terms of ${}_S R^2$, and taught *r* only in the contexts of correlation history and practical applications.

References

Baten, W. D. (1938). *Elementary mathematical statistics*. New York, NY: John Wiley and Sons.

Bluman, A. G. (2015). *Elementary statistics: A step by step approach* (7th annotated instructor's ed.). New York, NY: McGraw-Hill.

Bohrnstedt, G. W., & Knoke, D. (1988). *Statistics for social data analysis*. Itasca, IL: F. E. Peacock.

Bowley, A. L. Sir. (1901). *Elements of statistics*. London, UK: P. S. King and Son.

Dixon, W. J., & Massey, F. J., Jr. (1969). *Introduction to statistical analysis* (3rd. ed.). New York, NY: McGraw-Hill.

Dovich, R. A. (1990). *Reliability statistics*. Milwaukee, WI: ASQC Quality Press.

Duncan, A. J. (1986). *Quality control and industrial statistics* (5th ed.). Homewood, IL: Irwin.

Edgeworth, F. Y. (1892). Correlated averages. *Philosophical Magazine, 5*(34), 190–204. doi: 10.1080/14786449208620307

Edgeworth, F. Y. (1893). Statistical correlation between social phenomena. *Journal of the Royal Statistical Society, 56*, 670–675.

Ezekiel, M. (1930). *Methods of correlation analysis*. New York, NY: John Wiley and Sons.

Ezekiel, M., & Fox, K. A. (1959). *Methods of correlation and regression analysis: Linear and curvilinear* (3rd ed.). New York, NY: John Wiley and Sons.

Fisher, R. A. Sir. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Fisher, R. A. Sir. (1958). *Statistical methods for research workers* (13th ed., rev.). New York: Hafner.

Freund, J. E. (1960). *Modern elementary statistics* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London, 45*, 135–145.

Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth.

Healey, J. F. (1984). *Statistics: A tool for social research*. Belmont, CA: Wadsworth.

MacKenzie, D. A. (1981). *Statistics in Britain, 1865–1930: The social construction of scientific knowledge*. Edinburgh, Scotland: Edinburgh University Press.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Thousand Oaks, CA: Sage.

Pearson, E. S., (ed.). (1956). *Karl Pearson's early statistical papers*. Cambridge, UK: University Press.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London, Series A, 187*, 253–318 (note: page references in the body of this present article are to a reprint in *Karl Pearson's Early Statistical Papers* by E. S. Pearson, cited previously). doi: 10.1098/rsta.1896.0007

Pearson, K. (1911). *The grammar of science. Part I–Physical.* (3rd ed., rev.). London, UK: Adam and Charles Black.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66.doi: 10.1080/00031305.1988.10475524

Selkirk, K. E. (1981). *Rediguide 32: Correlation and regression*. Nottingham, UK: Nottingham University.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, IA: Iowa State University Press.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.

Symonds, P. M. (1926). Variations of the product-moment (Pearson) coefficient of correlation. *Journal of Educational Psychology, 17,* 458–469. doi: 10.1037/h0070082

Walker, H. M. (1929). *Studies in the history of statistical method: With special reference to certain educational problems*. Baltimore, MD: Williams and Wilkins. doi: 10.1037/13379-000

Wikipedia. (2017). *Beta (finance)* [Webpage]. Retrieved April 12, 2017 from https://en.wikipedia.org/wiki/Beta_(finance)

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Yule, G. U. (1910). The applications of the method of correlation to social and economic statistics. *Bulletin de L'Institut International de Statistique, 18*(1), 537–551.

Yule, G. U. (1912). *An introduction to the theory of statistics* (2nd ed.). London, UK: Charles Griffin.

Zorich, J. (2017). Four formulas for teaching the meaning of the correlation coefficient. *MathAMATYC Educator, 8*(3), 4–7.

Attachments: see next page

John N. Zorich received an MS degree in botany from the University of California, Davis, in 1979. He has worked in medical-device design and manufacturing as an independent statistical consultant and instructor since 1999. Annually since 2005, he has taught a course in introductory and applied statistics for the Biotechnology Center of Ohlone College (Newark, CA). Previously, he taught courses in applied statistics at Silicon Valley Polytechnic Institute and for the Silicon Valley ASQ Biomedical Division, and he has tutored introductory statistics at Lone Star College (Kingwood, TX). He currently resides near Houston, TX.

MathAMATYC Educator

Volume 9 • Number 3 • Summer 2018

About the Cover

This photo, via flickr.com, was takenwith a Sony SLT-A33onFebruary 24, 2012 by Raymond Adriano. Photo was cropped for publication. https://www.flickr.com/photos/aokpalad/6927185143/in/photolist-byBCNB-CtNP3i-2pVJd3-CZS6Eu-pPgPJj-CVFDHY-dM7t31-pytYvf-7Q4YW2-oRxcUN-HNBuN-U4MutG-cmmtWG-k7heU8-Yk2oN4-o9vQ2z-8gQWsa-SzXDk6-nbtxzM-6MX3rk

Staff

Editor: Johanna Debrecht
Northern Virginia CC
1200 Campus Drive Sterling, VA 20164
(703) 404-7320
jdebrecht@nvcc.edu

Assistant Editor: George Alexander
Madison Area TC
1701 Wright St. Madison, WI 53704
(608) 246-6187
galexander@madisoncollege.edu

Production Manager: Anthony Piccolino
4200 Congress Ave.
Lake Worth, FL 33461
(561) 523-4197
piccolia@palmbeachstate.edu

Graphic Designer: Michael Glowacki
existdesignstudio.com
hello@existdesignstudio.com

Features

MathAMATYC Educator's Departments

Use It Now

In the Spotlight

The Problem Section

Reasons for Teaching and Using the Signed Coefficient of Determination Instead of the Correlation Coefficient

John N. Zorich, Jr.
Ohlone College

Does the simple linear correlation coefficient, typically symbolized by r or r_{XY}, have meaning that is mathematically rigorous? In other words, what does it mean when one r value is twice as large as another? Does it mean "twice as much correlation?" Does r measure or quantify the "strength" of correlation? Can any of the definitions or formulas for the correlation coefficient help answer these questions?

Correlation Coefficient

Mathematical correlation and the simple linear correlation coefficient became important parts of statistical method after their discovery in 1888 by Francis Galton (Walker, 1929). He defined them by saying that "two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other" (Galton, 1888, p. 135); he provided no mathematically rigorous meaning (in the sense provided in the introduction) for what he called his "index of co-relation" (p. 143), which he symbolized by r, other than to say that it "measures the closeness of co-relation" (p. 145). As early as 1892, r came to be called the "coefficient of correlation" (Edgeworth, 1892, p. 191) and was then viewed (incorrectly) as a "proportion" (Edgeworth, 1893, p. 674).

The following is a chronological list of examples of definitions that have been given for r since Galton's time; none of these definitions help answer the questions posed previously:

- "r measures the correspondence between deviations from their means of the two series of observations" (Bowley, 1901, p. 320),
- "[r is the] amount of dependence one variable has upon another" (Baten, 1938, p. 170),
- "[r is the] degree of association between two variables" (Duncan, 1986, p. 820),
- "[r is the] extent to which the scattergraph of the relationship between two variables fits a straight line" (Miles & Shevlin, 2001, p. 20), and
- "[r] measures the strength … of the linear relationship between two variables" (Bluman, 2015, p. 543).

Formulas for the Correlation Coefficient

There are many strikingly different looking but mathematically identical formulas for the correlation coefficient (Symonds, 1926; Rogers & Nicewander, 1988). Although some of them are useful for introducing novice students to the general concept of correlation (Zorich, 2017), none of them help answer the questions initially posed. The following is a sampling of such formulas.

In 1896, the first "basic formula for estimating the correlation coefficient had finally been presented" (Stigler, 1986, p. 343); its presenter was Karl Pearson. That is why the correlation coefficient has most often been defined as "Pearson's r":

$$\text{Formula 1:}\quad \text{Pearson's } r = \frac{\text{covariance of } X, Y}{\sqrt{(\text{variance of } X)(\text{variance of } Y)}}.$$

Formulas 2, 3, and 4 have also been used to define the correlation coefficient. For example, Formula 2 was used by Yule in 1910 (p. 538), Formula 3 by Ezekiel in 1930 (p. 118), and Formula 4 by Dixon and Massey in 1969 (p. 203). The meanings of symbols used in these formulas are:

b_1 = the slope of the linear regression of Y on X,

b_2 = the slope of the linear regression of X on Y,

S_X = standard deviation of the X values,

S_Y = standard deviation of the Y values,

S_{y_e} = the standard deviation of the values derived from the linear regression equation at each X value in an X, Y data set (i.e., $Y_e = a + bX$), and

r (in Formulas 2 and 3) takes its sign from the slopes of the linear regressions.