Four Formulas for Teaching the Meaning of the Correlation Coefficient

THIS IS THE MSWORD DOCUMENT FROM WHICH THE *JOURNAL OF THE AMERICAN MATHEMATICAL ASSOCIATION OF TWO YEAR COLLEGES* CREATED THE PUBLISHED MAY 2017 VERSION.
THE 1ST PAGE OF THE AS-PUBLISHED VERSION CAN BE FOUND AT THE END OF THIS DOCUMENT;
THE FULL AS-PUBLISHED VERSION, PAGES 4-7, CAN BE FOUND STARTING AT:
https://amatyc.site-ym.com/page/EducatorMay2017

SUMMARY

The best equation for introducing students to the meaning of the correlation coefficient is not Karl Pearson's "product-moment" formula. Better alternatives are presented and discussed here. The relationship between the correlation and regression coefficients is reviewed. Some historical discussion is included.

1. Introduction

Almost a century ago, Symonds (1926) published a paper that contained 52 different formulas for calculating the "product-moment (Pearson) coefficient of correlation" (a.k.a., correlation coefficient, index of co-relation, or simply "r"); decades before computers, he focused on formulas useful for "ease of computation". Rodgers and Nicewander (1988) included 13 such formulas, where "each formula suggests a different way of thinking about" the correlation coefficient; the authors focused on "interesting" formulas (p. 59); some of their 13 formulas are not found among Symond's 52. This present paper focuses on 4 formulas useful for teaching the correlation coefficient's meaning to novice students; one of the 4 is not found in either of those previous papers.

Which characteristics of a formula are useful for explaining a difficult concept? At the top of such a list would certainly be that it is both simple and instructive. For example, the concept of "momentum" as taught in introductory physics classes is explained using:

Momentum = Mass x Velocity

Contrast that with the following formula for the correlation coefficient that is found most prominently in many introductory statistics textbooks published in the past 100 years:

Standard Formula: $$r = \frac{\sum(Xi-\bar{X})(Yi-\bar{Y})}{\sqrt{\sum(Xi-\bar{X})^2 \sum(Yi-\bar{Y})^2}}$$

Having taught the correlation coefficient in classrooms and seminars for many years, this author concludes that the Standard Formula does not belong in introductory statistics textbooks, since it is far from simple and provides virtually no instructive value *to the average novice student*, who

can be more confused than educated by it. Instead, introductory textbooks should include one or more of the following four formulas.

2. Four Simple and Instructive Formulas

An unnecessary obstacle to simplifying the formula for the correlation coefficient is the commonly held belief that the formula must output the coefficient's "sign", which is said to indicate whether the correlation is "positive" or "negative". It can be argued that the sign is irrelevant and that correlation is neither positive nor negative, as shown by the following formula and discussion (the formula is from Rogers and Nicewander, p. 62):

Formula 1: $$\boldsymbol{r = b\frac{Sx}{Sy}}$$

where

- $\boldsymbol{b}$ = slope of the least-squares linear regression line, using Y on the vertical axis and X on the horizontal; the linear regression equation is typically given as $Y = a + bX$
- $\boldsymbol{Sy}$ = standard deviation of the plotted vertical-axis sample-data values
- $\boldsymbol{Sx}$ = standard deviation of the plotted horizontal-axis sample-data values.

As seen in Formula 1, the sign of the correlation coefficient is always the same as that of the slope, which is typically referred to as the "regression coefficient". The sign of the regression coefficient (i.e., the "*b*" in $Y = a + bX$) tells us whether the relationship between *X* and *Y* is negative or positive (i.e., whether the plotted linear regression line is an increasing or decreasing function); the *sign* of the correlation coefficient provides no additional information. If the world had never heard of the correlation coefficient's sign, nothing would have been lost, and the gain would have been many fewer confused students.

The next formula is one of Symond's. However, it was printed incorrectly (the square root sign is missing). The correct formula (from Rodgers & Nicewander, p. 62) is:

Formula 2: $$\boldsymbol{r = \sqrt{b_X b_Y}}$$

where

- $\boldsymbol{b_X}$ = slope of the least-squares linear regression line when Y is plotted on the vertical axis, and X is plotted on the horizontal
- $\boldsymbol{b_Y}$ = slope of the least-squares linear regression line when X is plotted on the vertical axis, and Y is plotted on the horizontal
- $\boldsymbol{r}$ = takes the same sign as $\boldsymbol{b_X}$ and $\boldsymbol{b_Y}$, which always have the same sign.

Formula 2 says that the correlation coefficient is the geometric average of the slopes of the two possible ways to plot the data, i.e., *Y* vs. *X* and *X* vs. *Y*.

Francis Galton discovered the correlation coefficient in 1888. He had been frustrated by the problematic fact that a plot of *Y* vs. *X* yields not only a different slope than a plot of *X* vs. *Y* but

also a *different relationship*. To explain that problem, let the equation for Y vs. X be represented as $Y = a + b_X X$, and the equation for X vs. Y as $X = c + b_Y Y$. If the latter equation is rearranged to yield $Y = (X - c)/b_Y$, then amazingly $(X - c)/b_Y$ does not equal $a + b_X X$. Referring to a paper Galton published a few years before, he wrote that the "stature of the father is correlated to that of the adult son, and the stature of the adult son to that of the father...; ...what I there called 'regression,' is different in the different cases" (Galton 1888, p. 143). Twenty years later, he reminisced that "I could not [*yet*] see my way to express the results of [*such regression analyses*] in a single formula" (Galton 1908, p. 302). Galton found his "single formula" when he discovered the correlation coefficient.

To make that discovery, Galton changed how he plotted his *X,Y* data. Instead of scaling the horizontal and vertical plot axes in the same units as the raw data, he scaled them in units of "probable error" (which is a 19th century term that is arithmetically related to the 20th century's "standard deviation"). In other words, he plotted values that had been "transmuted into terms of a new scale" (Galton 1888, p. 136). In effect, he divided each of the individual X and Y values by their respective probable error. These transmuted values are denoted by X_T and Y_T, where the subscript "T" indicates transmuted raw data. Plots of Y_T vs. X_T and of X_T vs. Y_T exhibited the same "problematic fact" that was mentioned above (that is, $(X_T - c)/b_Y$ did not equal $a + b_X X_T$). Amazingly the two plots had the *identical* linear regression *slope.* It is interesting to note that in his "haste to prepare a paper" on this discovery before anyone else (Galton 1890, p. 421), he did not clearly explain what *meaning* to assign to this identical slope that he there called "r", other than what he gave in the very last words of the paper: "*r* measures the closeness of correlation" (Galton 1888, p.145). He did not improve upon that explanation in his subsequent abbreviated version of the publication (Galton 1889).

Formula 3 is the most direct calculation of the slope of the line on Galton's plot of "transmuted" *X,Y* values (i.e., the slope of Y_T vs. X_T). It is a classic Cartesian-coordinate calculation of "rise over run", where the rise is $Yei - \bar{Y}$ and the run is $Xi - \bar{X}$, and where the rise and run values are standardized by dividing by their respective standard deviations.

Formula 3: $$\boldsymbol{r} = \frac{(\boldsymbol{Yei} - \bar{\boldsymbol{Y}})/\boldsymbol{Sy}}{(\boldsymbol{Xi} - \bar{\boldsymbol{X}})/\boldsymbol{Sx}}$$

where

- $\bar{\boldsymbol{Y}}$ = arithmetic mean of the plotted vertical-axis sample-data values (i.e., the *Yi's*)
- $\bar{\boldsymbol{X}}$ = arithmetic mean of the plotted horizontal-axis sample-data values (i.e., the *Xi's*)
- $\boldsymbol{Sy}$ = standard deviation of the plotted vertical-axis sample-data values
- $\boldsymbol{Sx}$ = standard deviation of the plotted horizontal-axis sample-data values
- $\boldsymbol{Xi}$ = any arbitrarily chosen sample-data value plotted on the horizontal axis
- $\boldsymbol{Yei}$ = $a + bXi$ = the output of the least-squares linear regression equation derived from the "*Y* on *X*" plot of the sample-data; this *Yei* value must be calculated with the *Xi* value used in Formula 3's denominator.

The simplest and possibly the most instructive formula for the *magnitude* of the correlation coefficient is this last one, which is a ratio of two standard deviations (the formula is from Rogers and Nicewander, p. 62):

Formula 4 (*magnitude only*): $$\boldsymbol{r = \frac{S(Yei)}{Sy}}$$

where *S(Yei)* is the standard deviation of all the *Yei* values derived from all the plotted *Xi* sample-data values, and *Sy* is the standard deviation of all the plotted *Yi* sample-data values.

To understand the importance of Formula 4, it is necessary to understand that for a linear-regression slope (i.e., "*b*") *of a given value*, the smallest possible standard deviation for the vertical-axis sample-data *Yi* values is when the plotted *X,Y* points occur exactly on the linear regression line. The *S(Yei)* in Formula 4 is that smallest standard deviation. When any of the plotted *X,Y* points deviate from that line (in such a way that "*b*" remains that same "given value"), the result is a larger standard deviation, which is the *Sy* in Formula 4.

Formula 4 says that the correlation coefficient equals the fraction of the total observed variation in *Y* (as measured by *Sy*) that can be explained by a perfectly linear relationship between *X* and *Y* (as measured by *S(Yei)*). The remaining variation is caused by other factors (e.g., measurement error or confounding factors). For a more mathematically rigorous explanation, it is necessary to square both sides of the formula, thereby converting the standard deviations into variances and the correlation coefficient into what is known as the Coefficient of Determination; the correct statement then is that the coefficient of determination equals the fraction of the observed *variance* in *Y* that can be explained by a perfectly linear relationship between *X* and *Y*. However, since the correlation coefficient value is the same, no matter whether the linear regression plot is *Y* vs. *X* or *X* vs. *Y*, the best explanatory statement for use with novice students may be this: The correlation coefficient represents the fraction of the observed co-variation between the *X* and *Y* variables that can be explained by a perfectly linear relationship between them.

Figure 1. Identical Very High Correlation Coefficients But Very Different Slopes

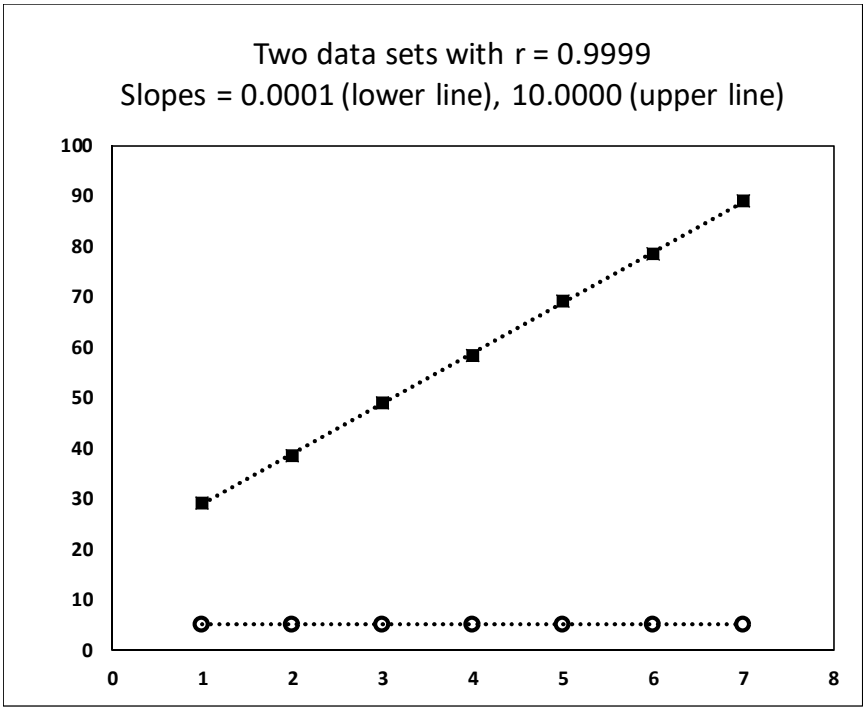


Figure 2. Identical High Correlation Coefficients But Very Different Slopes

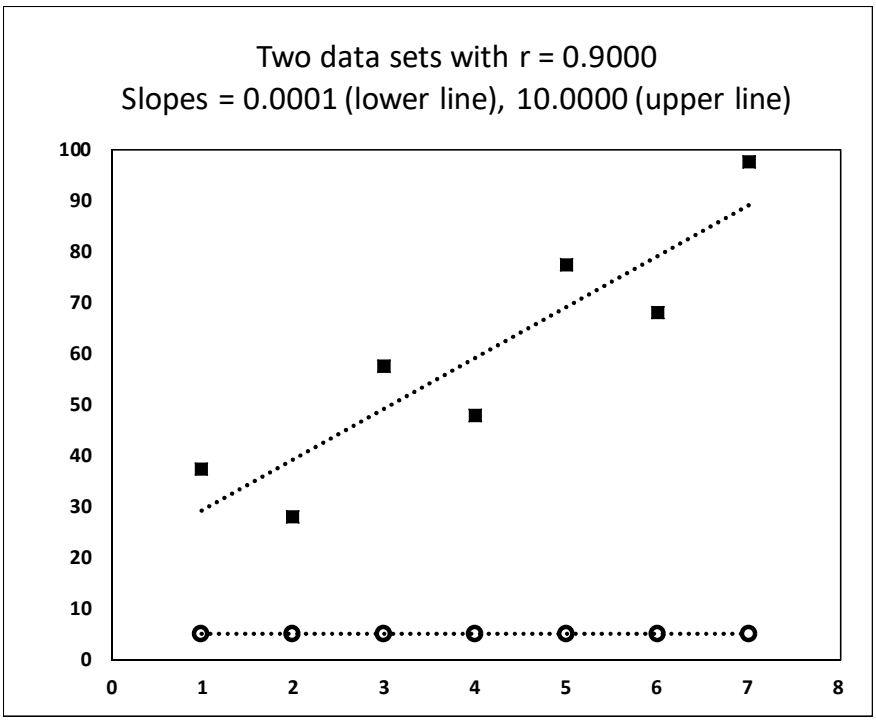


Figure 3. Identical Slopes But Different Correlation Coefficients

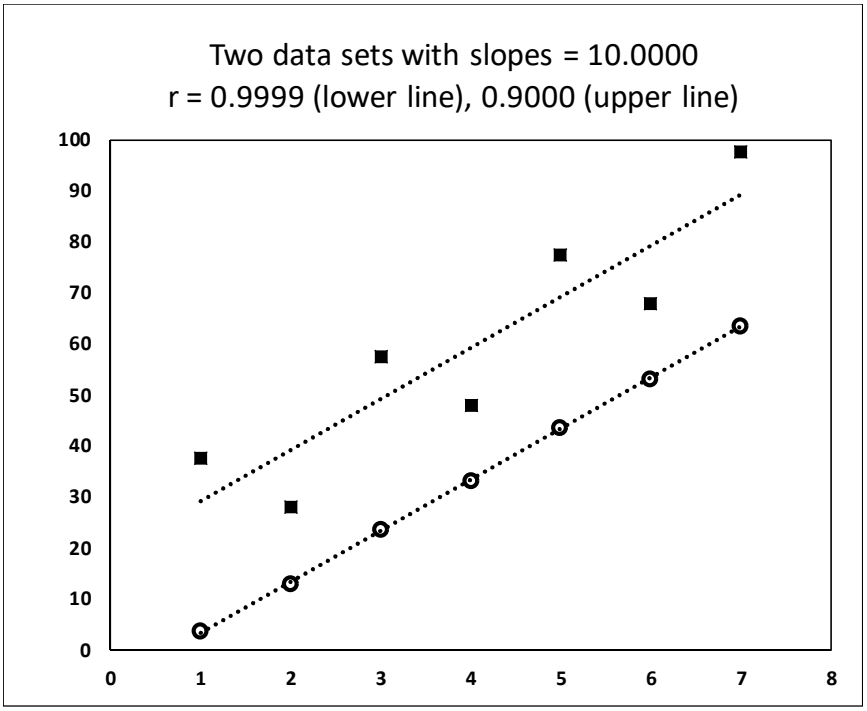
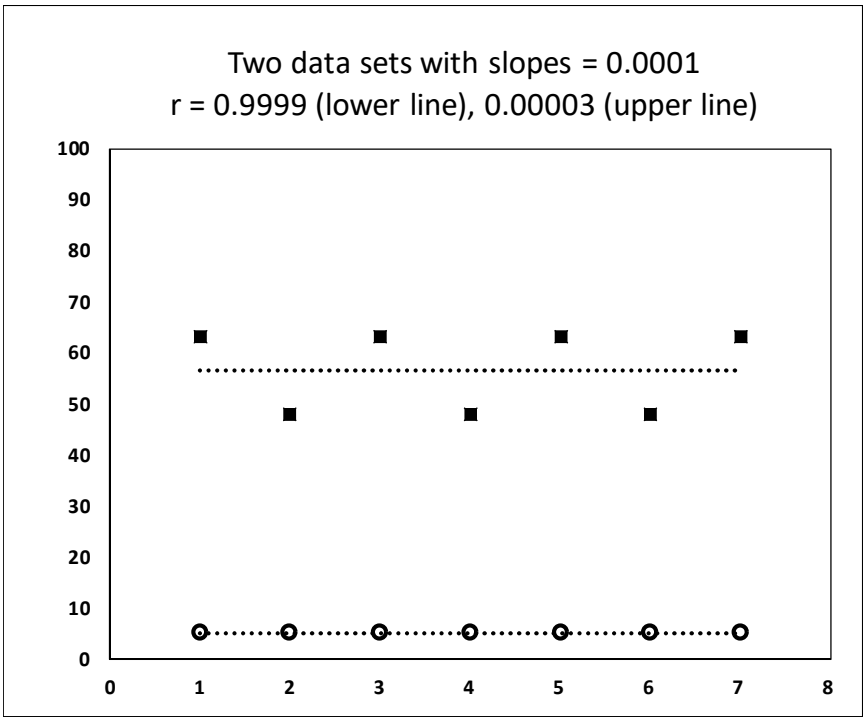


Figure 4. Identical Slopes But Very Different Correlation Coefficients



3. Concluding Remarks

In Galton's words, two variables "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". He also said that the correlation coefficient is a "simple number" that expresses..."the *nearness* with which they *vary* together" (Galton 1888, pp. 135-136, emphases added). Looking only at Figures 3 and 4, it is tempting to agree with Galton's definition that the correlation coefficient is a "measure of the *closeness* of the *co-relation*" (Galton 1888, p. 140; emphases added), because in each of those figures the *X,Y* data-set that displays the larger correlation coefficient (i.e., the lower line on each chart) also displays much *more* "closeness" and "nearness" to its line than does the other data-set. However, that definition cannot be used to explain why the correlation coefficients in Figure 2 are *identical* even though the upper line on the chart was produced with a data-set that displays much *less* "closeness" and "nearness" than does the other data-set.

Therefore, it might have been better for Galton to have defined the correlation coefficient as:

> *A measure of the closeness of the co-relation when comparing linear regression plots of X,Y data-sets that have the same regression coefficient (i.e., same slope).*

That definition helps explain Figures 3 and 4. In order to also help explain Figures 1 and 2, it might have been even better to define the correlation coefficient as:

> *The fraction of the observed co-variation between the X,Y variables that can be explained by a perfectly linear relationship between them*.

To help explain any of those figures, the Standard Formula can be used. However, these four other formulas dissect out the correlation coefficient's meaning so that it can be grasped more easily by novice students.

REFERENCES:

1. Galton, F. (1888), "Co-relations and their Measurement, chiefly from Anthropometric Data", *Proceedings of the Royal Society*, 45, 135-145.

2. Galton, F. (1889), "Correlations and their Measurement, chiefly from Anthropometric Data", *Nature*, 39, 238.

3. Galton, F. (1890), "Kinship and Correlation", *North American Review*, 150, 419-431

4. Galton, F. (1908), *Memories of My Life*, London: Methuen & Co.

5. Rogers, J. L., and Nicewander, W. (1988), "Thirteen Ways to Look at the Correlation Coefficient", *The American Statistician*, 42:1, 59-66.

6. Symonds, P. M. (1926), “Variations of the Product-Moment (Pearson) Coefficient of Correlation”, *Journal of Educational Psychology*, XVII, 458-469.

Four Formulas for Teaching the Meaning of the Correlation Coefficient

John Zorich
Ohlone Community College

Almost a century ago, Symonds (1926) published a paper that contained 52 different formulas for calculating the "product-moment (Pearson) coefficient of correlation" (a.k.a., correlation coefficient, index of co-relation, or simply *r*); decades before computers, he focused on formulas useful for "ease of computation." Rodgers and Nicewander (1988) included 13 such formulas, where "each formula suggests a different way of thinking about" the correlation coefficient; the authors focused on "interesting" formulas (p. 59), and some of their 13 formulas are not found among Symonds' 52. This present article focuses on four formulas useful for teaching the correlation coefficient's meaning to novice students; one of the four is not found in either of those previous articles.

Which characteristics of a formula are useful for explaining a difficult concept? At the top of such a list would certainly be that it is both simple and instructive. For example, the concept of momentum, as taught in introductory physics classes, is explained using

$$\text{Momentum} = \text{Mass} \times \text{Velocity}.$$

Contrast that with the following standard formula for the correlation coefficient that is found most prominently in many introductory statistics textbooks published in the past 100 years:

$$r = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2 \sum\left(Y_i - \overline{Y}\right)^2}}.$$

Having taught the correlation coefficient in classrooms and seminars for many years, this author concludes that the standard formula does not belong in introductory statistics textbooks, since it is far from simple and provides virtually no instructive value to the average novice student, who can be more confused than educated by it. Instead, introductory textbooks should include one or more of the following four formulas.

Four Simple and Instructive Formulas

An unnecessary obstacle to simplifying the formula for the correlation coefficient is the commonly held belief that the formula must output the coefficient's sign, which is said to indicate whether the correlation is positive or negative. It can be argued that the sign is irrelevant and that correlation is neither positive nor negative, as shown by the following formula and discussion (Rogers & Nicewander, 1988, p. 62):

$$r = b\frac{S_x}{S_y}, \qquad (1)$$

where

- $b =$ slope of the least-squares linear regression line, using Y on the vertical axis and X on the horizontal; the linear regression equation is typically given as $Y = a + bX$.
- $S_y =$ standard deviation of the plotted vertical-axis sample-data values.
- $S_x =$ standard deviation of the plotted horizontal-axis sample-data values.

As seen in formula 1, the sign of the correlation coefficient is always the same as that of the slope, which is typically referred to as the regression coefficient. The sign of the regression coefficient (i.e., the b in $Y = a + bX$) tells us whether the relationship between X and Y is negative or positive (i.e., whether the plotted linear regression line is an increasing or decreasing function); the *sign* of the correlation coefficient provides no additional information. If the world had never heard of the correlation coefficient's sign, nothing would have been lost, and the gain would have been many fewer confused students.

The next formula is one of Symonds' (1926). However, it was printed incorrectly, with the square root sign missing. The correct formula (Rodgers & Nicewander, 1988, p. 62) is:

$$r = \sqrt{b_X b_Y}, \qquad (2)$$