test is $\sup\{\Pr(|\bar{X}_A - \bar{X}_B| > a) \mid .75 \le p_1 = p_2 \le 1\}$. The power function $K(p_1,p_2,n_1,n_2)$ is $\Pr(|\bar{X}_A - \bar{X}_B| > a)$. To construct the tables, we evaluate the power function (for given n_1 and n_2 and $a = .1$) at the points (p_1,p_2) in

$$P = \{(m_1/100, m_2/100) \mid 75 \le m_1, m_2 \le 100 \text{ and } |m_1 - m_2| \ge 10\},$$

leaving out the so-called indifference region (see Bickel and Doksum 1977), that is, those points (p_1,p_2) such that $|p_1 - p_2| < .1$. The average power reported in the tables is then the average value of K over the set P. The size reported is similarly computed by taking the supremum over a finite set.

One effect of class size on the test is that in general, large imbalances in class size are associated with larger test size and small imbalances with smaller test size. Indeed, when the normal approximation applies, one can show that among all n_1, n_2 with $n_1 + n_2 = k$, k fixed, the test has minimum size when $n_1 = n_2 = k/2$ for k even and $n_1 = [k/2]$ or $n_1 = [k/2] + 1$ when k is odd. To see this, note that under H_0,

$$\begin{aligned}\Pr(\bar{X}_A - \bar{X}_B > .1) &+ \Pr(\bar{X}_A - \bar{X}_B < -.1)\\ &= \Pr((\bar{X}_A - \bar{X}_B)/\sqrt{pq(1/n_1 + 1/n_2)} \ge l)\\ &\quad + \Pr((\bar{X}_A - \bar{X}_B)/\sqrt{pq(1/n_1 + 1/n_2)} \le -l)\\ &= \Pr(Z \ge l) + \Pr(Z \le -l),\end{aligned}$$

where $l = .1/[pq(1/n_1 + 1/n_2)]^{1/2}$ and Z is a standard normal random variable. Using the usual techniques from calculus gives the desired result.

In Table 5 we find a similar pattern using the pdf given in (8) and the computer. Note the small percentage variation in mean power for the values in the table in comparison to the large percentage variation in size. The difference between class sizes (50, 50) and (5, 95) is a small increase in power but a fourfold increase in size.

[*Received October 1986. Revised June 1987.*]

REFERENCES

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

Cohen, Ayala (1983), "On the Effect of Class Size on the Evaluation of Lecturers' Performance," *The American Statistician*, 37, 331–333.

Thirteen Ways to Look at the Correlation Coefficient

JOSEPH LEE RODGERS and W. ALAN NICEWANDER*

In 1885, Sir Francis Galton first defined the term "regression" and completed the theory of bivariate correlation. A decade later, Karl Pearson developed the index that we still use to measure correlation, Pearson's r. Our article is written in recognition of the 100th anniversary of Galton's first discussion of regression and correlation. We begin with a brief history. Then we present 13 different formulas, each of which represents a different computational and conceptual definition of r. Each formula suggests a different way of thinking about this index, from algebraic, geometric, and trigonometric settings. We show that Pearson's r (or simple functions of r) may variously be thought of as a special type of mean, a special type of variance, the ratio of two means, the ratio of two variances, the slope of a line, the cosine of an angle, and the tangent to an ellipse, and may be looked at from several other interesting perspectives.

INTRODUCTION

We are currently in the midst of a "centennial decade" for correlation and regression. The empirical and theoretical developments that defined regression and correlation as statistical topics were presented by Sir Francis Galton in 1885. Then, in 1895, Karl Pearson published Pearson's r. Our article focuses on Pearson's correlation coefficient, presenting both the background and a number of conceptualizations of r that will be useful to teachers of statistics.

We begin with a brief history of the development of correlation and regression. Following, we present a longer review of ways to interpret the correlation coefficient. This presentation demonstrates that the correlation has developed into a broad and conceptually diverse index; at the same time, for a 100-year-old index it is remarkably unaffected by the passage of time.

The basic idea of correlation was anticipated substantially before 1885 (MacKenzie 1981). Pearson (1920) credited Gauss with developing the normal surface of n correlated variates in 1823. Gauss did not, however, have any particular interest in the correlation as a conceptually distinct notion; instead he interpreted it as one of the several parameters in his distributional equations. In a previous his-

*Joseph Lee Rodgers is Associate Professor and W. Alan Nicewander is Professor and Chair, Department of Psychology, University of Oklahoma, Norman, Oklahoma 73019. The authors thank the reviewers, whose comments improved the article.

torical paper published in 1895, Pearson credited Auguste Bravais, a French astronomer, with developing the bivariate normal distribution in 1846 (see Pearson 1920). Bravais actually referred to one parameter of the bivariate normal distribution as "une correlation," but like Gauss, he did not recognize the importance of the correlation as a measure of association between variables. [By 1920, Pearson had rescinded the credit he gave to Bravais. But Walker (1929) and Seal (1967) reviewed the history that Pearson both reported and helped develop, and they supported Bravais's claim to historical precedence.] Galton's cousin, Charles Darwin, used the concept of correlation in 1868 by noting that "all the parts of the organisation are to a certain extent connected or correlated together." Then, in 1877, Galton first referred to "reversion" in a lecture on the relationship between physical characteristics of parent and offspring seeds. The "law of reversion" was the first formal specification of what Galton later renamed "regression."

During this same period, important developments in philosophy also contributed to the concepts of correlation and regression. In 1843, the British philosopher John Stuart Mill first presented his "Five Canons of Experimental Inquiry." Among those was included the method of concomitant variation: "Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation." Mill suggested three prerequisites for valid causal inference (Cook and Campbell 1979). First, the cause must temporally precede the effect. Second, the cause and effect must be related. Third, other plausible explanations must be ruled out. Thus the separability of correlation and causation and the specification of the former as a necessary but not sufficient condition for the latter were being recognized almost simultaneously in the established discipline of philosophy and the fledgling discipline of biometry.

By 1885 the stage was set for several important contributions. During that year, Galton was the president of the Anthropological Section of the British Association. In his presidential address, he first referred to regression as an extention of the "law of reversion." Later in that year (Galton 1885) he published his presidential address along with the first bivariate scatterplot showing a correlation (Fig. 1). In this graph he plotted the frequencies of combinations of children's height and parents' height. When he smoothed the results and then drew lines through points with equal frequency, he found that "lines drawn through entries of the same value formed a series of concentric and similar ellipses." This was the first empirical representation of the isodensity contour lines from the bivariate normal distribution. With the assistance of J. D. Hamilton Dickson, a Cambridge mathematician, Galton was able to derive the theoretical formula for the bivariate normal distribution. This formalized mathematically the topic on which Gauss and Bravais had been working a half century before. Pearson (1920) stated that "in 1885 Galton had completed the theory of bi-variate normal correlation" (p. 37).

In the years following 1885, several additional events added mathematical import to Galton's 1885 work. In 1888, Galton noted that r measures the closeness of the "co-relation," and suggested that r could not be greater than 1 (although he had not yet recognized the idea of negative correlation). Seven years later, Pearson (1895) developed the mathematical formula that is still most commonly used

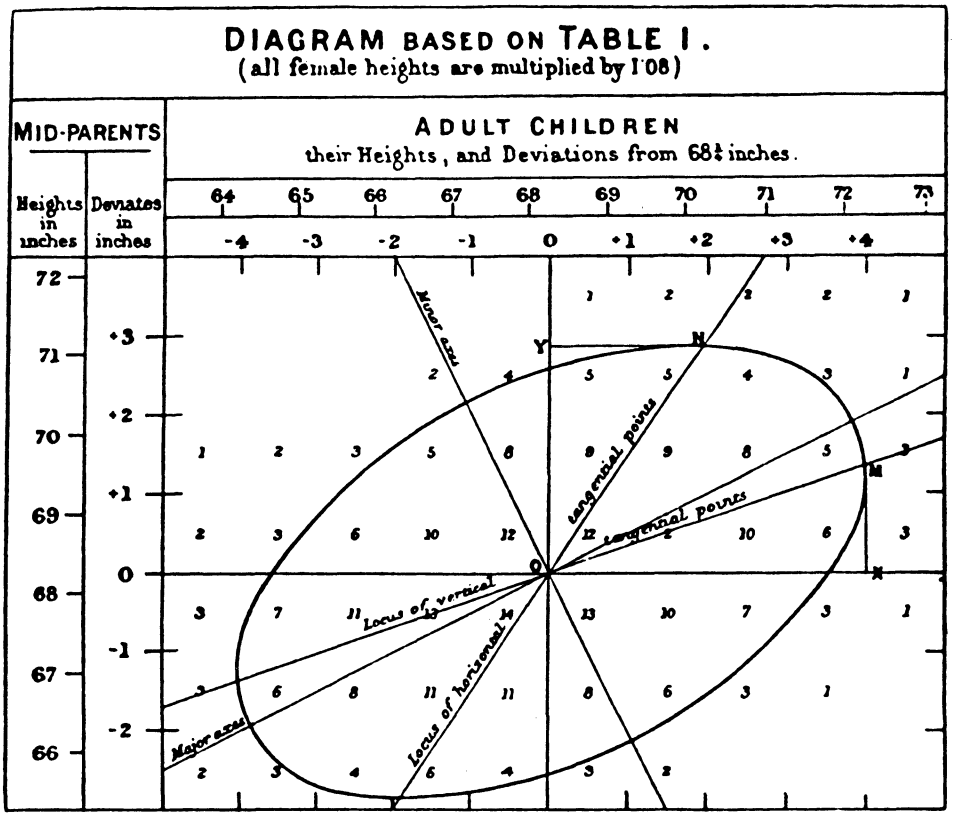

Figure 1. The First Bivariate Scatterplot (from Galton 1885).

Table 1. Landmarks in the History of Correlation and Regression

Date	Person	Event
1823	Carl Friedrich Gauss, German mathematician	Developed the normal surface of N correlated variates.
1843	John Stuart Mill, British philosopher	Proposed four canons of induction, including concomitant variation.
1846	Auguste Bravais, French naval officer and astronomer	Referred to "une correlation," worked on bivariate normal distribution.
1868	Charles Darwin, Galton's cousin, British natural philosopher	"All parts of the organisation are . . . connected or correlated."
1877	Sir Francis Galton, British, the first biometrician	First discussed "reversion," the predecessor of regression.
1885	Sir Francis Galton	First referred to "regression." Published bivariate scatterplot with normal isodensity lines, the first graph of correlation. "Completed the theory of bi-variate normal correlation." (Pearson 1920)
1888	Sir Francis Galton	Defined r conceptually, specified its upper bound.
1895	Karl Pearson, British statistician	Defined the (Galton-) Pearson product-moment correlation coefficient.
1920	Karl Pearson	Wrote "Notes on the History of Correlation."
1985		Centennial of regression and correlation

to measure correlation, the Pearson product-moment correlation coefficient. In historical perspective, it seems more appropriate that the popular name for the index *should* be the Galton–Pearson r. The important developments in the history of correlation and regression are summarized in Table 1.

By now, a century later, contemporary scientists often take the correlation coefficient for granted. It is not appreciated that before Galton and Pearson, the only means for establishing a relationship between variables was to educe a *causative* connection. There was no way to discuss—let alone measure—the association between variables that lacked a cause–effect relationship. Today, the correlation coefficient—and its associated regression equation—constitutes the principal statistical methodology for observational experiments in many disciplines. Carroll (1961), in his presidential address to the Psychometric Society, called the correlation coefficient "one of the most frequently used tools of psychometricians . . . and perhaps also one of the most frequently misused" (p. 347). Factor analysis, behavioral genetics models, structural equations models (e.g., LISREL), and other related methodologies use the correlation coefficient as the basic unit of data.

This article focuses on the Pearson product-moment correlation coefficient. Pearson's r was the first formal correlation measure, and it is still the most widely used measure of relationship. Indeed, many "competing" correlation indexes are in fact special cases of Pearson's formula. Spearman's rho, the point-biserial correlation, and the phi coefficient are examples, each computable as Pearson's r applied to special types of data (e.g., Henrysson 1971).

Our presentation will have a somewhat didactic flavor. On first inspection, the measure is simple and straightforward. There are surprising nuances of the correlation coefficient, however, and we will present some of these. Following Pearson, our focus is on the correlation coefficient as a computational index used to measure bivariate association. Whereas a more statistically sophisticated appreciation of correlation demands attention to the sampling model assumed to underlie the observations (e.g., Carroll 1961; Marks 1982), as well as understanding of its extension to multiple and partial correlation, our focus will be more basic. First, we restrict our primary interest to bivariate settings. Second, most of our interpretations are distribution free, since computation of a sample correlation requires no assumptions about a population (see Nefzger and Drasgow 1957). For consideration of the many problems associated with the inferential use of r (restriction of range, attenuation, etc.) we defer to other treatments (e.g., Lord and Novick 1968). We present 13 different ways to conceptualize this most basic measure of bivariate relationship. This presentation does not claim to exhaust all possible interpretations of the correlation coefficient. Others certainly exist, and new renderings will certainly be proposed.

1. CORRELATION AS A FUNCTION OF RAW SCORES AND MEANS

The Pearson product-moment correlation coefficient is a dimensionless index, which is invariant to linear transformations of either variable. Pearson first developed the mathematical formula for this important measure in 1895:

$$r = \frac{\Sigma(X_i - \bar{X})\,(Y_i - \bar{Y})}{[\Sigma(X_i - \bar{X})^2\,\Sigma(Y_i - \bar{Y})^2]^{1/2}}. \qquad (1.1)$$

This, or some simple algebraic variant, is the usual formula found in introductory statistics textbooks. In the numerator, the raw scores are centered by subtracting out the mean of each variable, and the sum of cross-products of the centered variables is accumulated. The denominator adjusts the scales of the variables to have equal units. Thus Equation (1.1) describes r as the centered and standardized sum of cross-product of two variables. Using the Cauchy–Schwartz inequality, it can be shown that the absolute value of the numerator is less than or equal to the denominator (e.g., Lord and Novick 1968, p. 87); therefore, the limits of ± 1 are established for r. Several simple algebraic transformations of this formula can be used for computational purposes.

2. CORRELATION AS STANDARDIZED COVARIANCE

The covariance, like the correlation, is a measure of linear association between variables. The covariance is defined on the sum of cross-products of the centered variables, unadjusted for the scale of the variables. Although the covariance is often ignored in introductory textbooks, the variance (which is not) is actually a special case of the covariance—that is, the variance is the covariance of a variable with itself. The covariance of two variables is unbounded in infinite populations, and in the sample it has indeterminant bounds (and unwieldy interpretation). Thus the covariance is often not a useful descriptive measure of association, because its value depends on the scales of measurement for X and Y. The correlation coefficient is a rescaled covariance:

$$r = s_{XY}/s_X s_Y, \tag{2.1}$$

where s_{XY} is the sample covariance, and s_X and s_Y are sample standard deviations. When the covariance is divided by the two standard deviations, the range of the covariance is rescaled to the interval between -1 and $+1$. Thus the interpretation of correlation as a measure of relationship is usually more tractable than that of the covariance (and different correlations are more easily compared).

3. CORRELATION AS STANDARDIZED SLOPE OF THE REGRESSION LINE

The relationship between correlation and regression is most easily portrayed in

$$r = b_{Y\cdot X}(s_X/s_Y) = b_{X\cdot Y}(s_Y/s_X), \tag{3.1}$$

where $b_{Y\cdot X}$ and $b_{X\cdot Y}$ are the slopes of the regression lines for predicting Y from X and X from Y, respectively. Here, the correlation is expressed as a function of the slope of either regression line and the standard deviations of the two variables. The ratio of standard deviations has the effect of rescaling the units of the regression slope into units of the correlation. Thus the correlation is a standardized slope.

A similar interpretation involves the correlation as the slope of the standardized regression line. When we standardize the two raw variables, the standard deviations become unity and the slope of the regression line *becomes* the correlation. In this case, the intercept is 0, and the regression line is easily expressed as

$$\hat{z}_Y = r\, z_X. \tag{3.2}$$

From this interpretation, it is clear that the correlation rescales the units of the standardized X variable to predict units of the standardized Y variable. Note that the slope of the regression of z_Y on z_X restricts the regression line to fall between the two diagonals portrayed in the shaded region of Figure 2. Positive correlations imply that the line will pass through the first and third quadrants; negative correlations imply that it will pass through the second and fourth quadrants. The regression of z_X on z_Y has the same angle with the Y axis that the regression of z_Y on z_X has with the X axis, and it will fall in the unshaded region of Figure 2, as indicated.

4. CORRELATION AS THE GEOMETRIC MEAN OF THE TWO REGRESSION SLOPES

The correlation may also be expressed as a simultaneous function of the two slopes of the unstandardized regression lines, $b_{Y\cdot X}$ and $b_{X\cdot Y}$. The function is, in fact, the geometric mean, and it represents the first of several interpretations of r as a special type of mean:

$$r = \pm\sqrt{b_{Y\cdot X} b_{X\cdot Y}}. \tag{4.1}$$

This relationship may be derived from Equation (3.1) by multiplying the second and third terms in the equality to give r^2, canceling the standard deviations, and taking the square root.

There is an extension of this interpretation involving multivariate regression. Given the matrices of regression coefficients relating two sets of variables, $B_{Y\cdot X}$ and $B_{X\cdot Y}$, the square roots of the eigenvalues of the product of these matrices are the canonical correlations for the two sets of variables. These values reduce to the simple correlation coefficient when there is a single X and a single Y variable.

5. CORRELATION AS THE SQUARE ROOT OF THE RATIO OF TWO VARIANCES (PROPORTION OF VARIABILITY ACCOUNTED FOR)

Correlation is sometimes criticized as having no obvious interpretation for its units. This criticism is mitigated by squaring the correlation. The squared index is often called the coefficient of determination, and the units may be interpreted as proportion of variance in one variable accounted for by differences in the other [see Ozer (1985) for a discussion of several different interpretations of the coefficient of determination]. We may partition the total sum of squares for Y (SS_{TOT}) into the sum of squares due to regression (SS_{REG}) and the sum of squares due to error (SS_{ERR}). The variability in X accounted for by differences in Y is the ratio of SS_{REG} to SS_{TOT}, and r is the square root ojf that ratio:

$$r = \sqrt{\Sigma(Y_i - \hat{Y}_i)^2/\Sigma(Y_i - \bar{Y})^2} = \sqrt{SS_{REG}/SS_{TOT}}.$$

Equivalently, the numerator and denominator of this equation may be divided by $(N-1)^{1/2}$, and r becomes the square root of the ratio of the variances (or the ratio of the standard deviations) of the predicted and observed variables:

$$r = \sqrt{s^2_{\hat{Y}}/s^2_Y} = s_{\hat{Y}}/s_Y. \tag{5.1}$$

(Note that $s^2_{\hat{Y}}$ is a biased estimate of $\sigma^2_{\hat{Y}}$, whereas s^2_Y is unbiased.) This interpretation is the one that motivated Pearson's early conceptualizations of the index (see Mulaik 1972, p. 4). The correlation as a ratio of two variances may be compared to another interpretation (due to Galton) of the correlation as the ratio of two means. We will present that interpretation in Section 13.

6. CORRELATION AS THE MEAN CROSS-PRODUCT OF STANDARDIZED VARIABLES

Another way to interpret the correlation as a mean (see Sec. 4) is to express it as the average cross-product of the standardized variables:

$$r = \Sigma z_X z_Y / N. \quad (6.1)$$

Equation (6.1) can be obtained directly by dividing both the numerator and denominator in Equation (1.1) by the product of the two sample standard deviations. Since the mean of a distribution is its first moment, this formula gives insight into the meaning of the "product-moment" in the name of the correlation coefficient.

The next two portrayals involve trigonometric interpretations of the correlation.

7. CORRELATION AS A FUNCTION OF THE ANGLE BETWEEN THE TWO STANDARDIZED REGRESSION LINES

As suggested in Section 3, the two standardized regression lines are symmetric about either diagonal. Let the angle between the two lines be β (see Fig. 2). Then

$$r = \sec(\beta) \pm \tan(\beta). \quad (7.1)$$

A simple proof of this relationship is available from us. Equation (7.1) is not intuitively obvious, nor is it as useful for computational or conceptual purposes as some of the others. Its value is to show that there is a systematic relationship between the correlation and the angular distance between the two regression lines. The next interpretation—also trigonometric—has substantially more conceptual value.

8. CORRELATION AS A FUNCTION OF THE ANGLE BETWEEN THE TWO VARIABLE VECTORS

The standard geometric model to portray the relationship between variables is the scatterplot. In this space, observations are plotted as points in a space defined by variable axes. An "inside out" version of this space—usually called "person space"—can be defined by letting each axis represent an observation. This space contains two points—one for each variable—that define the endpoints of vectors in this (potentially) huge dimensional space. Although the multidimensionality of this space precludes visualization, the two variable vectors define a two-dimensional subspace that is easily conceptualized.

If the variable vectors are based on centered variables, then the correlation has a straightforward relationship to the angle α between the variable vectors (Rodgers 1982):

$$r = \cos(\alpha). \quad (8.1)$$

When the angle is 0, the vectors fall on the same line and $\cos(\alpha) = \pm 1$. When the angle is 90°, the vectors are perpendicular and $\cos(\alpha) = 0$. [Rodgers, Nicewander, and Toothaker (1984) showed the relationship between orthogonal and uncorrelated variable vectors in person space.]

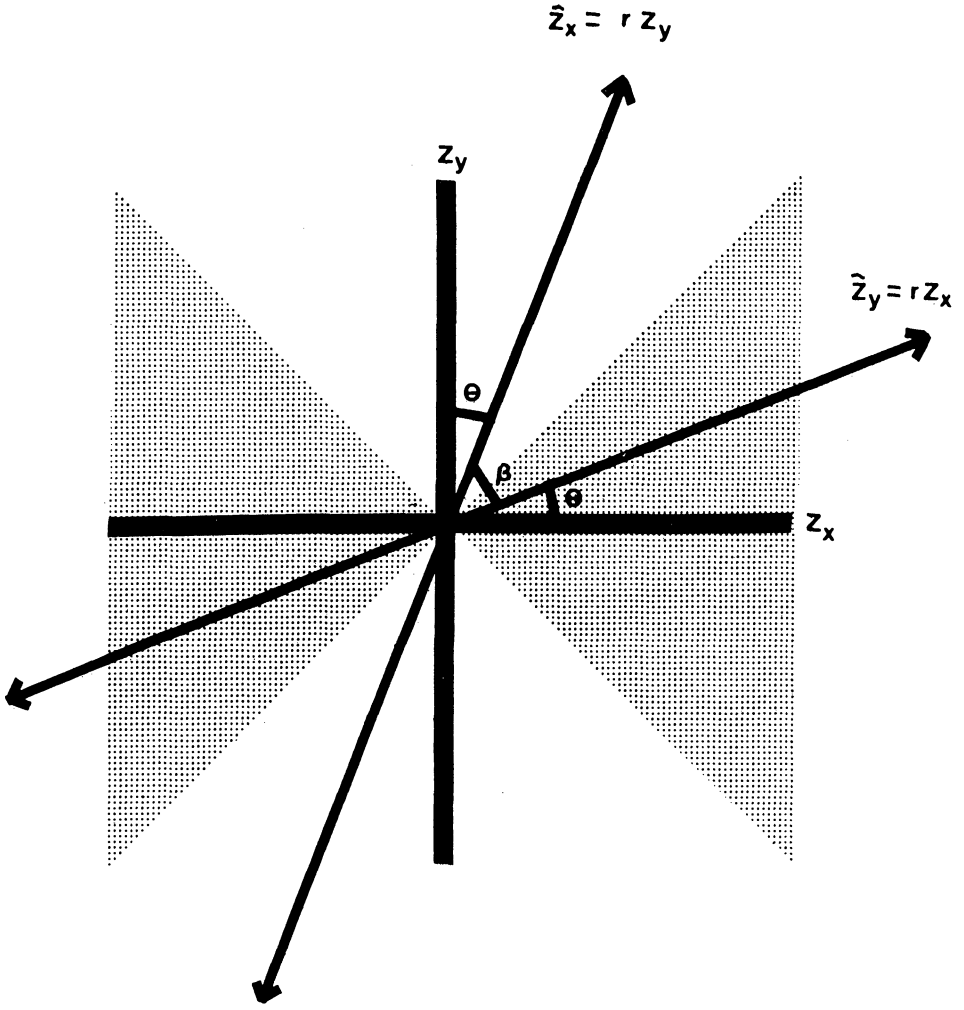

Figure 2. The Geometry of Bivariate Correlation for Standardized Variables.

Visually, it is much easier to view the correlation by observing an angle than by looking at how points cluster about the regression line. In our opinion, this interpretation is by far the easiest way to "see" the size of the correlation, since one can directly observe the size of an angle between two vectors. This inside-out space that allows r to be represented as the cosine of an angle is relatively neglected as an interpretational tool, however. Exceptions include a number of factor-analytic interpretations, Draper and Smith's (1981, pp. 201–203) geometric portrayals of multiple regression analysis, and Huck and Sandler (1984, p. 52). Fisher also used this space quite often to conceptualize his elegant statistical insights (see Box 1978).

9. CORRELATION AS A RESCALED VARIANCE OF THE DIFFERENCE BETWEEN STANDARDIZED SCORES

Define $z_Y - z_X$ as the difference between standardized X and Y variables for each observation. Then

$$r = 1 - s^2_{(z_Y - z_X)}/2. \tag{9.1}$$

This can be shown by starting with the variance of a difference score: $s^2_{Y-X} = s^2_X + s^2_Y - 2rs_Xs_Y$. Since the standard deviations and variances become unity when the variables are standardized, we can easily solve for r to get Equation (9.1).

It is interesting to note that in this equation, since the correlation is bounded by the interval from -1 to $+1$, the variance of this difference score is bounded by the interval from 0 to 4. Thus the variance of a difference of standardized scores can never exceed 4. The upper bound for the variance is achieved when the correlation is -1.

We may also define r as the variance of a sum of standardized variables:

$$r = s^2_{(z_Y + z_X)}/2 - 1. \tag{9.2}$$

Here, the variance of the sum also ranges from 0 to 4, and the upper bound is achieved when the correlation is $+1$. The value of this ninth interpretation is to show that the correlation is a linear transformation of a certain type of variance. Thus, given the correlation, we can directly define the variance of either the sum or difference of the standardized variables, and vice versa.

All nine of the preceding interpretations of the correlation coefficient were algebraic and trigonometric in nature. No distributional assumptions were made about the nature of the univariate or bivariate distributions of X and Y. In the final interpretations, bivariate normality will be assumed. We maintain our interest in conceptual and computational versions of r, but we base our last set of interpretations on this common assumption about the population distribution.

10. CORRELATION ESTIMATED FROM THE BALLOON RULE

This interpretation is due to Chatillon (1984a). He suggested drawing a "birthday balloon" around the scatterplot of a bivariate relationship. The balloon is actually a rough ellipse, from which two measures—h and H—are obtained (see Fig. 3). h is the vertical diameter of the ellipse at the center of the distribution on the X axis; H is the vertical range of the ellipse on the Y axis. Chatillon showed that the correlation may be roughly computed as

$$r = \sqrt{1 - (h/H)^2}. \tag{10.1}$$

He gave a theoretical justification of the efficacy of this rough-and-ready computational procedure, assuming both bivariate normality and bivariate uniformity. He also presented a number of examples in which the technique works quite well. An intriguing suggestion he made is that the "balloon rule" can be used to construct approximately a bivariate relationship with some specified correlation. One draws an ellipse that produces the desired r and then fills in the points uniformly throughout the ellipse. Thomas (1984) presented a "pocket nomograph," which was a $3'' \times 5''$ slide that could be used to "sight" a bivariate relationship and estimate a correlation based on the balloon rule.

11. CORRELATION IN RELATION TO THE BIVARIATE ELLIPSES OF ISOCONCENTRATION

Two different authors have suggested interpretations of r related to the bivariate ellipses of isoconcentration. Note that these ellipses are more formal versions of the "balloon" from Section 10 and that they are the geometric structures that Galton observed in his empirical data (see Fig. 1). Chatillon (1984b) gave a class of bivariate distributions (including normal, uniform, and mixtures of uniform) that have elliptical isodensity contours. There is one ellipse for every positive constant, given the population correlation. The balloon that one would draw around a scatterplot would approximate one of these ellipses for a large positive constant. If the variables are standardized, then these ellipses are centered on the origin. For $\rho > 0$, the major axes fall on the positive diagonal; for $\rho < 0$, the negative diagonal.

Marks (1982) showed, through simple calculus, that the slope of the tangent line at $z_X = 0$ is the correlation. Figure

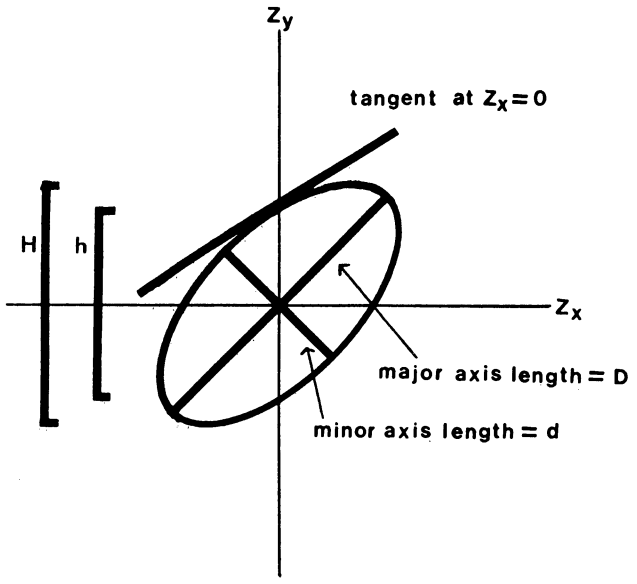

Figure 3. The Correlation Related to Functions of the Ellipses of Isoconcentration.

3 shows this tangent line, the slope of which equals r. When the correlation is 0, the ellipse is a circle and the tangent has slope 0. When the correlation is unity, the ellipse approaches a straight line that is the diagonal (with slope 1). Note that, since all of the ellipses of isoconcentration are parallel, the interpretation is invariant to the choice of the ellipse. It is also worth noting that the slope of the tangent line at $z_X = 0$ is the same as the slope of the standardized regression line (see Sec. 3).

Schilling (1984) also used this framework to derive a similar relationship. Let the variables be standardized so that the ellipses are centered on the origin, as before. If D is the length of the major axis of some ellipse of isoconcentration and d is the length of the minor axis, then

$$r = (D^2 - d^2)/(D^2 + d^2). \qquad (11.1)$$

These axes are also portrayed in Figure 3, and the interpretation is invariant to choice of ellipse, as before.

12. CORRELATION AS A FUNCTION OF TEST STATISTICS FROM DESIGNED EXPERIMENTS

The previous interpretations of r were based on quantitative variables. Our 12th representation of the correlation shows its relationship to the test statistic from designed experiments, in which one of the variables (the independent variable) is a categorical variable. This demonstrates the artificiality of the correlational/experimental distinction in discussing experimental design. In fact, Fisher (1925) originally presented the analysis of variance (ANOVA) in terms of the intraclass correlation (see Box 1978).

Suppose we have a designed experiment with two treatment conditions. The standard statistical model to test for a difference between the conditions is the two-independent-sample t test. If X is defined as a dichotomous variable indicating group membership (0 if group 1, 1 if group 2), then the correlation between X and the dependent variable Y is

$$r = t/\sqrt{t^2 + n - 2}, \qquad (12.1)$$

where n is the combined total number of observations in the two treatment groups. This correlation coefficient may be used as a measure of the *strength* of a treatment effect, as opposed to the *significance* of an effect. The test of significance of r in this setting provides the same test as the usual t test. Clearly, then, r can serve as a test statistic in a designed experiment, as well as provide a measure of association in observational settings.

In ANOVA settings with more groups or multiple factors, the extension of this relationship defines the *multiple* correlation coefficients associated with main effects and interactions in more complex experiments. For example, in a one-way ANOVA setting with k groups and a total of N subjects, the squared-multiple correlation between the dependent variable and the columns of the design matrix is related to the F statistic through the following formula (Draper and Smith 1981, p. 93): $R^2 = F(k - 1)/[F(k - 1) + (N - k)]$.

13. CORRELATION AS THE RATIO OF TWO MEANS

This is the third interpretation of the correlation involving means (see Secs. 4 and 6). It provides an appropriate conclusion to our article, since it was first proposed by Galton. Furthermore, Galton's earliest conceptions about and calculations of the correlation were based on this interpretation. An elaboration of this interpretation was presented by Nicewander and Price (1982).

For Galton, it was natural to focus on correlation as a ratio of means, because he was interested in questions such as, How does the average height of sons of unusually tall fathers compare to the average height of their fathers? The following development uses population rather than sample notation because it is only in the limit (of increasing sample size) that the ratio-of-means expression will give values identical to Pearson's r.

Consider a situation similar to one that would have interested Galton. Let X be a variable denoting mother's IQ, and let Y denote the IQ of her oldest child. Further assume that the means $\mu(X)$ and $\mu(Y)$ are 0 and that the standard deviations $\sigma(X)$ and $\sigma(Y)$ are unity. Now select some arbitrarily large value of X (say X_c), and compute the mean IQ of mothers whose IQ is greater than X. Let this mean be denoted by $\mu(X|X > X_c)$, that is, the average IQ of mothers whose IQ is greater than X_c. Next, average the IQ scores, Y, of the oldest offspring of these exceptional mothers. Denote this mean by $\mu(Y|X > X_c)$, that is, the average IQ of the oldest offspring of mothers whose IQ's are greater than X_c. Then it can be shown that

$$r = \frac{\mu(Y|X > X_c) - \mu_Y}{\mu(X|X > X_c) - \mu_X} = \frac{\mu(Y|X > X_c)}{\mu(X|X > X_c)}. \qquad (13.1)$$

The proof of (13.1) requires an assumption of bivariate normality of standardized X and Y. The proof is straightforward and entails only the fact that, for z_X and z_Y, r is the slope of the regression line as well as the ratio of these two conditional means.

Our example is specific, but the interpretation applies to any setting in which explicit selection occurs on one variable, which implicitly selects on a second variable. Brogden (1946) used the ratio-of-means interpretation to show that when a psychological test is used for personnel selection, the correlation between the test score and the criterion measure gives the proportionate degree to which the test is a "perfect" selection device. Other uses of this interpretation are certainly possible.

CONCLUSION

Certainly there are other ways to interpret the correlation coefficient. A wealth of additional fascinating and useful portrayals is available when a more statistical and less algebraic approach is taken to the correlation problem. We in no sense presume to have summarized all of the useful or interesting approaches, even within the fairly tight framework that we have defined. Nevertheless, these 13 approaches illustrate the diversity of interpretations available for teachers and researchers who use correlation.

Galton's original work on the correlation was motivated by a very specific biometric problem. It is remarkable that such a focused effort would lead to the development of what is perhaps the most broadly applied index in all of statistics. The range of interpretations for the correlation coefficient demonstrates the growth of this remarkable index over the past century. On the other hand, Galton and Pearson's index is surprisingly unchanged from the one originally proposed.

[*Received June 1987. Revised August 1987.*]

REFERENCES

Box, J. F. (1978), *R. A. Fisher: The Life of a Scientist*, New York: John Wiley.

Brogden, H. E. (1946), "On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency," *Journal of Educational Psychology*, 37, 65–76.

Carroll, J. B. (1961), "The Nature of the Data, or How to Choose a Correlation Coefficient," *Psychometrika*, 26, 347–372.

Chatillon, G. (1984a), "The Balloon Rules for a Rough Estimate of the Correlation Coefficient," *The American Statistician*, 38, 58–60.

——— (1984b), "Reply to Schilling," *The American Statistician*, 38, 330.

Cook, T. C., and Campbell, D. T. (1979), *Quasi-Experimentation*, Boston: Houghton Mifflin.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis*, New York: John Wiley.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh, U.K.: Oliver & Boyd.

Galton, F. (1885), "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15, 246–263.

Henrysson, S. (1971), "Gathering, Analyzing, and Using Data on Test Items," in *Educational Measurement*, ed. R. L. Thorndike, Washington, DC: American Council on Education, pp. 130–159.

Huck, S. W., and Sandler, H. M. (1984), *Statistical Illusions: Solutions*, New York: Harper & Row.

Lord, F. M., and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

MacKenzie, D. A. (1981), *Statistics in Britain: 1865–1930*, Edinburgh, U.K.: Edinburgh University Press.

Marks, E. (1982), "A Note on the Geometric Interpretation of the Correlation Coefficient," *Journal of Educational Statistics*, 7, 233–237.

Mulaik, S. A. (1972), *The Foundations of Factor Analysis*, New York: McGraw-Hill.

Nefzger, M. D., and Drasgow, J. (1957), "The Needless Assumption of Normality in Pearson's *r*," *The American Psychologist*, 12, 623–625.

Nicewander, W. A., and Price, J. M. (1982), "The Correlation Coefficient as the Ratio of Two Means: An Interpretation Due to Galton and Brogden," unpublished paper presented to the Psychometric Society, Montreal, Canada, May.

Ozer, D. J. (1985), "Correlation and the Coefficient of Determination," *Psychological Bulletin*, 97, 307–315.

Pearson, K. (1895), *Royal Society Proceedings*, 58, 241.

——— (1920), "Notes on the History of Correlation," *Biometrika*, 13, 25–45.

Rodgers, J. L. (1982), "On Huge Dimensional Spaces: The Hidden Superspace in Regression and Correlation Analysis," unpublished paper presented to the Psychometric Society, Montreal, Canada, May.

Rodgers, J. L., Nicewander, W. A., and Toothaker, L. (1984), "Linearly Independent, Orthogonal, and Uncorrelated Variables," *The American Statistician*, 38, 133–134.

Schilling, M. F. (1984), "Some Remarks on Quick Estimation of the Correlation Coefficient," *The American Statistician*, 38, 330.

Seal, H. L. (1967), "The Historical Development of the Gauss Linear Model," *Biometrika*, 54, 1–24.

Thomas, H. (1984), "Psychophysical Estimation of the Correlation Coefficient From a Bivariate Scatterplot: A Pocket Visual Nomograph," unpublished paper presented to the Psychometric Society, Santa Barbara, California, June.

Walker, H. M. (1929), *Studies in the History of Statistical Method*, Baltimore: Williams & Wilkins.

Election Recounting

BERNARD HARRIS*

1. INTRODUCTION

The purpose of this article is to provide a simple model for the calculation of the probability of success in reversing the result of a closely contested election by recounting of ballots. Technological changes render the model described here less relevant to actual political elections than has been the case. The consequences of these technological changes and their effect on modeling election recounting are described in Section 5. There I mention some other election recounting problems, which are also not covered by the simple model employed here. Nevertheless, the election described actually occurred, and I discussed the chance of reversing the outcome with the losing candidate.

In the anecdotal material that follows, the names of the actual participants are suppressed. Also, I have delayed writing this material for many years with the belief that the passage of time would make the facts concerning this election indistinguishable from many other elections with comparable outcomes.

2. THE ELECTION

About 500,000 votes were cast in the disputed election. Candidate A lost by a plurality of about 1,500 votes. That is, Candidate B received 50.15% of the votes cast. In view of the closeness of the outcome, Candidate A felt that a recount would be desirable and that there was an excellent chance that the outcome would be reversed. His campaign advisor, who was also the treasurer of the election campaign, said, "There are 2,000 election districts. We can win with a recount if we change just one vote in each district." State law required the candidate requesting a recount to guarantee

*Bernard Harris is Professor, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706. He thanks the referees for their careful perusal of this article and for their constructive comments.

References

The Balloon Rules for a Rough Estimate of the Correlation Coefficient
Guy Chatillon
The American Statistician, Vol. 38, No. 1. (Feb., 1984), pp. 58-60.
Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198402%2938%3A1%3C58%3ATBRFAR%3E2.0.CO%3B2-C

The Balloon Rules for a Rough Estimate of the Correlation Coefficient
Guy Chatillon
The American Statistician, Vol. 38, No. 1. (Feb., 1984), pp. 58-60.
Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198402%2938%3A1%3C58%3ATBRFAR%3E2.0.CO%3B2-C

A Note on a Geometric Interpretation of the Correlation Coefficient
Edmond Marks
Journal of Educational Statistics, Vol. 7, No. 3. (Autumn, 1982), pp. 233-237.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198223%297%3A3%3C233%3AANOAGI%3E2.0.CO%3B2-P

Notes on the History of Correlation
Karl Pearson
Biometrika, Vol. 13, No. 1. (Oct., 1920), pp. 25-45.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28192010%2913%3A1%3C25%3ANOTHOC%3E2.0.CO%3B2-6

LINKED CITATIONS

Linearly Independent, Orthogonal, and Uncorrelated Variables

Joseph Lee Rodgers; W. Alan Nicewander; Larry Toothaker

The American Statistician, Vol. 38, No. 2. (May, 1984), pp. 133-134.

Stable URL:

http://links.jstor.org/sici?sici=0003-1305%28198405%2938%3A2%3C133%3ALIOAUV%3E2.0.CO%3B2-W

Letters to the Editor

Mark F. Schilling; Guy Chatillon; C. Philip Cox; H. J. Keselman; Hanspeter Thoni; Charles J. Monlezun; David C. Blouin; Linda C. Malone

The American Statistician, Vol. 38, No. 4. (Nov., 1984), pp. 330-332.

Stable URL:

http://links.jstor.org/sici?sici=0003-1305%28198411%2938%3A4%3C330%3ALTTE%3E2.0.CO%3B2-4

Studies in the History of Probability and Statistics. XV: The Historical Development of the Gauss Linear Model

Hilary L. Seal

Biometrika, Vol. 54, No. 1/2. (Jun., 1967), pp. 1-24.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28196706%2954%3A1%2F2%3C1%3ASITHOP%3E2.0.CO%3B2-Z