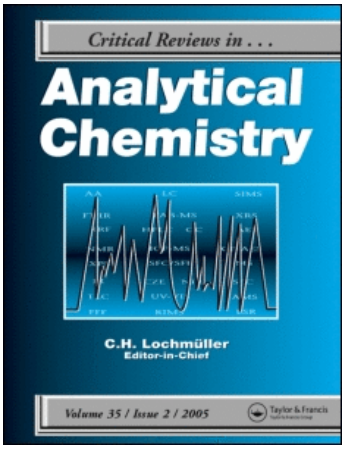


Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

The Correlation Coefficient: An Overview

A. G. Asuero[a]; A. Sayago[a]; A. G. González[a]
[a] Department of Analytical Chemistry, Faculty of Pharmacy, The University of Seville, Seville, Spain

To cite this Article Asuero, A. G. , Sayago, A. and González, A. G.(2006) 'The Correlation Coefficient: An Overview', Critical Reviews in Analytical Chemistry, 36: 1, 41 — 59
To link to this Article: DOI: 10.1080/10408340500526766
URL: http://dx.doi.org/10.1080/10408340500526766

Critical Reviews in Analytical Chemistry, 36:41–59, 2006

ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408340500526766

The Correlation Coefficient: An Overview

A. G. Asuero, A. Sayago, and A. G. González
Department of Analytical Chemistry, Faculty of Pharmacy, The University of Seville, Seville, Spain

Correlation and regression are different, but not mutually exclusive, techniques. Roughly, regression is used for prediction (which does not extrapolate beyond the data used in the analysis) whereas correlation is used to determine the degree of association. There situations in which the *x* variable is not fixed or readily chosen by the experimenter, but instead is a random covariate to the *y* variable. This paper shows the relationships between the coefficient of determination, the multiple correlation coefficient, the covariance, the correlation coefficient and the coefficient of alienation, for the case of two related variables *x* and *y*. It discusses the uses of the correlation coefficient *r*, either as a way to infer correlation, or to test linearity. A number of graphical examples are provided as well as examples of actual chemical applications. The paper recommends the use of *z* Fisher transformation instead of *r* values because *r* is not normally distributed but *z* is (at least in approximation). For either correlation or for regression models, the same expressions are valid, although they differ significantly in meaning.

Keywords multiple correlation coefficient, correlation coefficient, covariance, cause and effect inference, lineariy, significance tests

INTRODUCTION

Although the concepts of correlation and regression are intimately related, they are nevertheless different (1). Correlation may be described as the degree of association between two variables, whereas regression expresses the form of the relationship between specified values of one (the independent, exogenous, explanatory, regressor, carrier or predictor) variable and the means of all corresponding values of the second (the dependent, outcome, response variable, the variable being explained) variable. In general, we can say that the study of interdependence leads to the investigation of correlations (2), while the study of dependence leads to the theory of regression. When the x variable is a random covariate to the y variable, that is, x and y vary together (continuous variables), we are more interested in determining the strength of the linear relationship than in prediction, and the sample correlation coefficient, r_{xy} (r), is the statistics employed (3) for this purpose.

The Pearson (Product–Moment) correlation r was developed by Pearson (1896) and was based on the work of others, including Galton (1888), who first introduced the concept of correlation (4, 5). As a matter of fact, correlation charts, also known as scatter diagrams is one of the seven basic tools of statistical quality control (6). Empirical relationships can be used, i.e., to determine yield vs. conditions, so process optimization can be achieved (7), or in linear free-energy relationships (LFR), quantitative structure-activity relationships (QASR) and quantitative structure property (QSPR) relationships (8, 9). The correlation, however, is a concept of much wider applicability than just 2D scatterplots. There is also "multiple correlation," which is the correlation of multiple independent variables with a single dependent. Also there is "partial correlation," which is the correlation of one variable with another, controlling for a third or additional variables (10).

If the parameter estimates associated with any one factor in a multifactor design are uncorrelated with those of another, the experiment design is to said to be orthogonal. This is the basic principle of the orthogonal design, which often permits (11) simple formulas to be used to calculate effects, thus avoiding on this way tedious manual calculations. Correlation and covariance play a central role in clustering; correlation is used as such to measure similarity between objects (12). Factor analysis, behavioural genetic models, structural equations models and other related methodologies use the correlation coefficient as the basic unit of data (5, 13). Canonical correlation analysis is a way of measuring the linear relationship between two multidimensional variables (10).

There are a number of different correlation coefficients to handle the special characteristics of such types of variables as dichotomies, and there are other measurements of association for nominal and ordinal variables (and for time-series analysis

Address correspondence to A. G. Asuero, Department of Analytical Chemistry, Faculty of Pharmacy, The University of Seville, Seville 41012, Spain. E-mail: asuero@us.es

as well). The literature provides a variety of measures of dependence (i.e., Spearman's ρ, the point biserial correlation and the ϕ coefficient), which are either improved versions of the correlation coefficient or based on complementary different concepts (14).

SCATTERPLOTS

The mandatory first step in all data analysis is to make a plot of the data in the most illustrative way possible. A two-dimensional representation of n pairs of measurements (x_i, y_i) made on two random variables x and y, is known as a scatterplot. The first bivariate scatterplot (5) showing a correlation was given by Galton in 1885. Such plots are particularly useful tools in exploratory analysis conveying information about the association between x and y (15), the dependence of y on x where y is a response variable, the clustering of the points, the presence of outliers, etc.

Scatterplots are much more informative (16) than the correlation coefficient. This should be clear (17) from Figure 1. Each of the four data yields the same standard output from a typical regression program. Scatterplots can also be combined in multiple plots per page to help understanding higher-level structure in data set with more than two variables. Thus, a scatterplot matrix can be a better summary of data than a correlation matrix, since the latter gives only a single number summary of the linear relationship between variables, while each scatterplot gives a visual summary of linearity, nonlinearity, and separated points (16, 18).

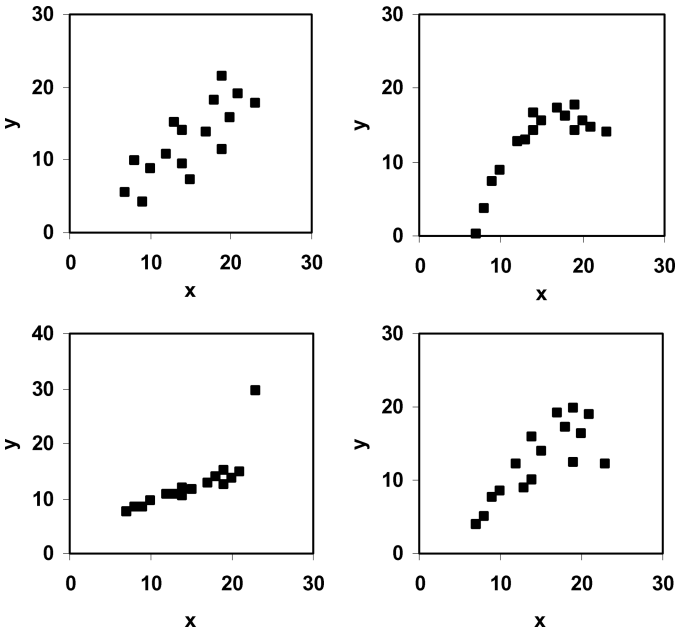


FIG. 1. Scatter diagrams of a four set of data showing the same $a_0 = 0.520$, $a_1 = 0.809$, $R^2 = 0.617$, $s_{y/x} = 3.26$, $N = 16$.

THE LEAST-SQUARES METHOD

Suppose that estimates a_0 and a_1 of the model parameters α_0 and α_1 are found by using the principle of least squares concerning the sum of squares of (weighted) vertical deviations (residuals) from the regression line. Heteroscedasticity is assumed, where the measuring quantity is determinable nor with constant (homoscedastic condition), but with different variance dependent on the size of the measuring quantity. The (weighted) least squares method is based on a number of assumptions (19), i.e., (i) that the errors, ε_i, are random rather than systematic, with mean zero and variances $\sigma_i^2 = \sigma^2/w_i$ (σ is a constant and w_i is the weight of point i) and follow a Gaussian distribution; (ii) that the independent variable, i.e., x, the abscissa, is known exactly or can be set by the experimenter either; (iii) the observations, y_i, are in an effective sense uncorrelated and statistically independent, i.e., for $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for i $\neq$ j, with means equal to their respective expectations or true values, $E\{y_i\} = \eta_i$; and (iv) that the correct weights, w_i are known. Normality is needed only for tests of significance and construction of confidence intervals estimates of the parameters. Formulae for calculating a_0 and a_1 and their standard errors by weighted linear regression are given in Table 1, where the analogy with simple linear regression (20) is evident.

Particular attention must be given on this respect to equations in which one variable is involved on both sides (21, 22). Then, the independent variable x is not an exact quantity and the independence of errors is not fulfilled. If for any reason the precision with which the x_i values are known is not considerably better than the precision of measurement of the y_i values, the statistical analysis based on the (ordinary) weighted least squares is not valid and it is necessary a more general approach (23). Weights in those cases are slope-dependent.

TABLE 1
Formulae for calculating statistics for weighted linear regression

Equation:	Slope
$\hat{y}_i = a_0 + a_1 x_i$	$a_1 = S_{XY}/S_{xx}$
Weights:	Intercept
$w_i = 1/s_i^2$	$a_0 = \bar{y} - a_1\bar{x}$
Explained sum of squares	Weighted residuals
$SS_{\mathrm{Reg}} = \sum w_i(\hat{y}_i - \bar{y})^2$	$w_i^{1/2}(y_i - \hat{y}_i)$
Residual sum of squares	Correlation coefficient
$SSE = \sum w_i(y_i - \hat{y}_i)^2$	$r = S_{XY}/\sqrt{S_{XX}S_{YY}}$
Mean	Standard errors
$\bar{x} = \sum w_i x_i / \sum w_i$	$s_{y/x}^2 = \frac{SSE}{n-2} = \frac{S_{YY} - a_1^2 S_{XX}}{n-2}$
$\bar{y} = \sum w_i y_i / \sum w_i$	$s_{a_0}^2 = s_{y/x}^2(\sum w_i x_i^2)/(S_{XX}\sum w_i)$
Sum of squares about the mean	$s_{a_1}^2 = s_{y/x}^2/S_{XX}$
$S_{XX} = \sum w_i(x_i - \bar{x})^2$	$\mathrm{cov}(a_0, a_1) = -\bar{x}s_{y/x}^2/S_{XX}$
$S_{YY} = \sum w_i(y_i - \bar{y})^2$	
$S_{XY} = \sum w_i(x_i - \bar{x})(y_i - \bar{y})$	

TABLE 2
Common data transformations

Transformation			Comments
Reciprocal	$1/y$		Linearizing data, particularly rate phenomena
Arcsine (angular)	$\arcsin\sqrt{y}$		Proportions ($0 > p > 1$)
Logarithmic	$\ln y$		Variance $\propto$ (mean)^2
Probability	probability		Percentage responding
Logistic (probit)	$\log(\frac{y}{1-y})$		Drug dose response curves and UV killing curves
Square root	$\sqrt{y}$		Counts, variance $\propto$ (mean)
Box Cox	$(y^p - 1)/p$	$p \neq 0$	Family of transformations for use when one has no prior knowledge of an appropriate transformation to use
	$\ln p$	$p = 0$	
Tukey	$\{p^\lambda - (1-p)\lambda\}/\lambda$		

HETEROSCEDASTICITY AND TRANSFORMATION

There are many examples and methods for heteroscedastic regression: those based on counting measurements and also photometric, chromatographic and capillary electrophoresis analysis under certain conditions (24, 25). With calibration over moderate-to-wide calibration ranges, the assumption of constant variance is almost always false (26). On the other hand, when we plot our initial results on a graph, it will be usually clear whether they best fit a linear relationship or a logarithmic relationship or; something else, like a sigmoid curve. We can analyse all these relationships if we transform the x and y values as appropriate so that the relationship between $x$and y become linear (Table 2). Nevertheless, when the outcome is not a direct experimental value but has been transformed, its error can vary substantially. When experimental data y are converted into transformed data Y for subsequent use in a least squares analysis, one should introduce a weighting factor given by (21)

$$w_i = 1 \Big/ \left(\frac{\partial Y}{\partial y}\right)^2. \qquad [1]$$

The objective of a transformation is to rescale the data so that the variability becomes constant, allowing the original unweighted least squares regression to be used. Therefore, it is often true that a variance-stabilizing transformation is also effective in transforming a skew, nonnormal variable into a reasonably symmetric and approximate normal one (27).

COEFFICIENT OF DETERMINATION R^2 AND CORRELATION COEFFICIENT r_{xy}

In addition to graphing, many numerical tools are available that you can use to determine how well a regression equation fits the data. A useful statistics to check is the sample coefficient of determination, R^2, of a linear regression fit (with any number of predictor). R^2 is equal to the ratio of the sum of squares accounted for by the regression (SS_{Reg}) to the total sum of squares of deviation about the mean (S_{YY}) for a model with constant term (homoscedastic case, $w_i = 1$)

$$R^2 = \frac{SS_{\mathrm{Reg}}}{S_{YY}} = \frac{S_{YY} - SSE}{S_{YY}} = 1 - \frac{SSE}{S_{YY}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad [2]$$

where $\hat{y}$ denotes the predicted value of y and as usual $\bar{y}$ is the mean of y's values and both summations are over $i = 1,2,\ldots n$. SSE is the residual sum of squares. In a model without constant term, $R^2 = 1 - SSE/SST$, where SST is the total sum of squares $\sum y^2$ (28). R^2 in Eq. (2) measures the proportion of total variation about the mean $\bar{y}$ explained by the regression. Thus, the large is R^2, the more is the total variation of $\bar{y}$ reduced by introducing the independent variable x. It is often expressed as a percentage by multiplying by 100. Since $0 \leq SSE \leq SS_{YY}$, it follows that $0 \leq R^2 \leq 1$. In fact, R is the correlation between y and $\hat{y}$

$$R = r_{y\hat{y}} = \frac{\sum (\hat{y}_i - \bar{\hat{y}}_i)(y_i - \bar{y})}{[\sum (y_i - \bar{\hat{y}}_i)^2]^{1/2} [\sum (y_i - \bar{y})^2]^{1/2}} \qquad [3]$$

and is usually called the multiple correlation coefficient. It is not appropriate to compare R^2 of different equations containing different numbers of coefficients derived from the same data set (29,30). In spite of this we continue to like R^2 as a useful thing to look at in a regression printout.

In the case of simple regression with constant term, the coefficient of determination equals the square of the correlation coefficient between x and y, which explain the notation. Then, for a straight line only it holds that (20)

$$r_{xy} = \pm\sqrt{R^2} = \sqrt{1 - \frac{S_{YY} - a_1^2 S_{XX}}{S_{YY}}} = a_1 \sqrt{\frac{S_{XX}}{S_{YY}}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}. \qquad [4]$$

A plus or minus sign is attached to this measure according to whether the slope, a_1, of the fitted regression line is positive or negative. The variability decomposition is depicted in Figure 2. If R^2 is the unity, all variation has been explained and we have a perfect fit with all the points lying on the regression line. If the coefficient is zero, the regression line does not explain anything; i.e., it is horizontal and y is not a function of x.

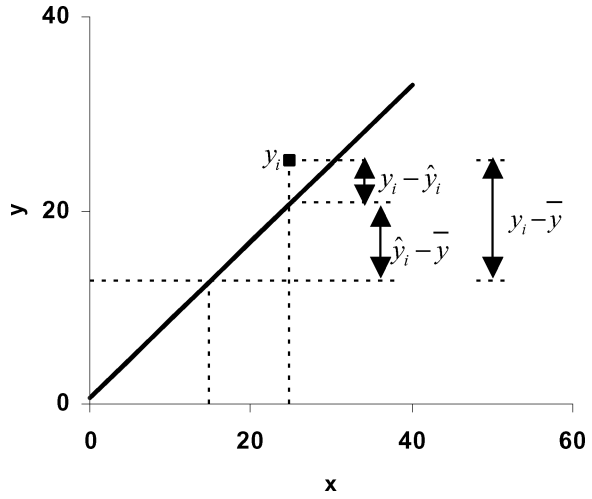


FIG. 2. Decomposition of variability in the least-squares straight line case.

Taking into account that $\hat{y}_i = a_0 + a_1 x$ and $\bar{\hat{y}}_i = a_0 + a_1 \bar{x} = \bar{y}$ if we substitute for $\hat{y}_i - \bar{\hat{y}}_i = a_1 (x - \bar{x})$ in Eq. (3) and cancel out at top and bottom we get with r_{xy} in Eq. (4). Equation (3) is true for any linear regression with any number of predictors whereas Eq. (4) holds only for the straight-line case. In more general regression problems, the regression coefficients are also related to correlation of the type r_{xy} but in a more complicated manner (3).

COVARIANCE

The covariance between two random variables x and y, with a joint normal distribution, is a measure of correlation of the fluctuation of two quantities and is defined as (31) the expected value of the product of the deviations of x and y from their expected values (true or population means). The sample covariance is given by

$$\mathrm{cov}(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}). \qquad [5]$$

It is a simple algebraic exercise to prove (31, 32) that it satisfies the so-called Schwartz inequality

$$\mathrm{cov}(x, y) \leq s_x s_y, \qquad [6]$$

which implies immediately that $r \leq 1$.

The covariance is a measure of the correlation between x and y. If two variables are related in a linear way, then the covariance will be positive or negative depending on whether the relationship has a positive or negative slope. If x and y are independent, i.e., not correlated, the covariance is zero. However, the converse is not necessarily true (33), for it is possible to construct examples of highly dependent random variables, often in a nonlinear way, whose covariance (correlation) is zero. Although the covariance is often ignored in introductory textbooks, the variance is the special case of the covariance of a random variable with itself. The square root of the variance is called the standard deviation (denoted by σ for the population and by s for the sample) and is always positive. Covariance have to be taken into account at least in cases where realistic uncertainty budgets have to be calculated or traceability chains have to be built up (34). Standard addition method, for example, inevitably leads to correlated data. Here, covariances must be taken into account (35, 36). The determination of the boiling point of water from measurements of its vapour pressure constitutes (37, 38) a dramatic example of the need to consider the covariance.

COVARIANCE AND CORRELATION

The covariance is often not a useful descriptive measure of association, because its value depends on the scales of measurements for x and y, and then it must be standardized before it can be used as a generally applicable measure of association. By dividing the sample covariance by the product of the sample standard deviation of x and y, s_x and s_y, respectively, we obtain (31) the sample correlation coefficient r_{xy}. A simpler formula can be used: r_{xy} is the covariance between two standardized variables z_x and z_y, and is independent of the scales chosen

$$r_y = \frac{\mathrm{cov}(x, y)}{s_x s_y} = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$
$$= \frac{1}{n-1} \sum z_x z_y. \qquad [7]$$

Also we get

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}. \qquad [8]$$

The part above the line in this equation is a measure of the degree to which x and y, vary together (using the deviations of each from the mean). The part below the line is a measure of the degree to which x and y varies separately. Eq. (8) describes r_{xy} as the centered and standardized sum of cross-product of two variables and allows the direct computational formula for r_{xy}, which automatically furnishes the proper sign (Table 3). The assumption $(x_1 - \bar{x})(y_1 - \bar{y}) \neq 0$ eliminates vertical and horizontal lines (Figure 3). Notice that r_{xy} has the same value whether n or $n - 1$ is chosen as the common divisor for $\mathrm{cov}(x, y)$, s_x^2 and s_y^2.

TABLE 3
Signs of the differences with respect to the mean values on quadrant location (49)

Quadrant	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
I Upper right	+	+	+
II Upper left	−	+	−
III Lower left	−	−	+
IV Lower right	+	−	−

The denominator will always be positive (unless all of the x's or all the y's are equal.

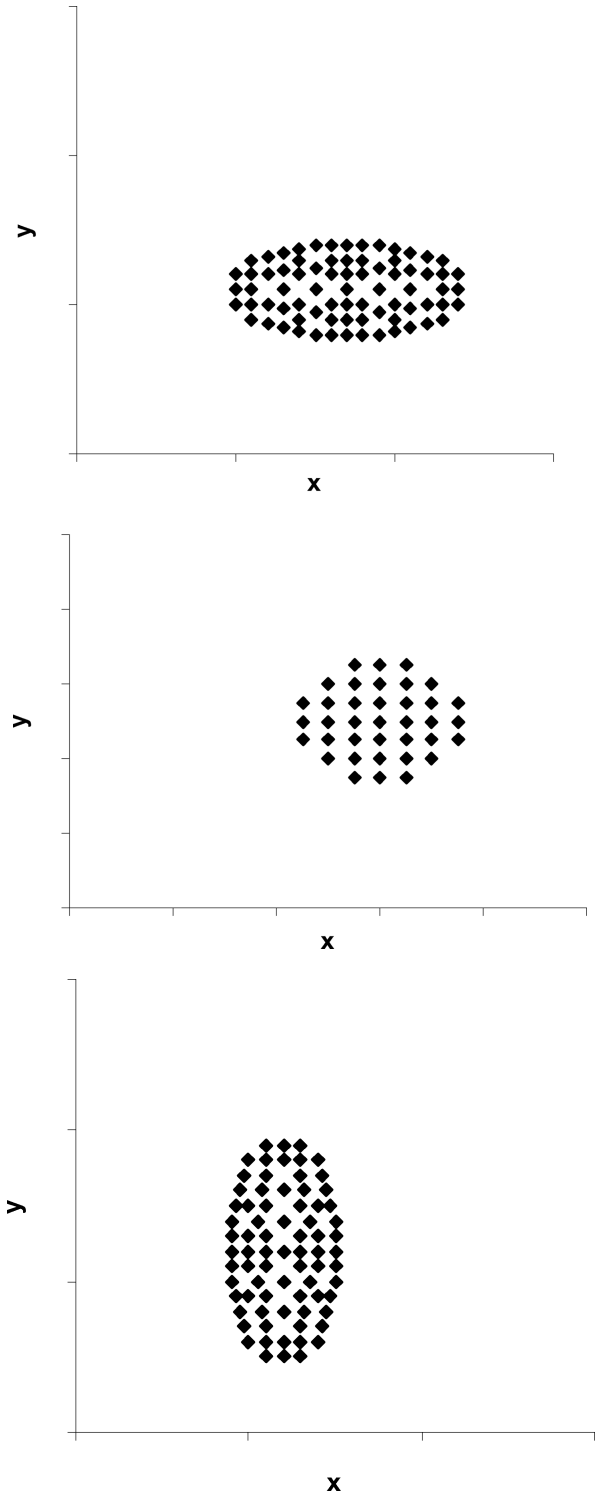


FIG. 3. Some examples of linear independence, $r = 0$, and functional dependence.

We shall call r without subscript in that follows. Although the signs of the sample correlation and the sample covariance are the same, r is ordinarily easier to interpret because its magnitude is bounded. Thus the interpretation of correlation as a measure of relationships is usually more tractable than that of the covariance, and different correlations are more easily (39) compared.

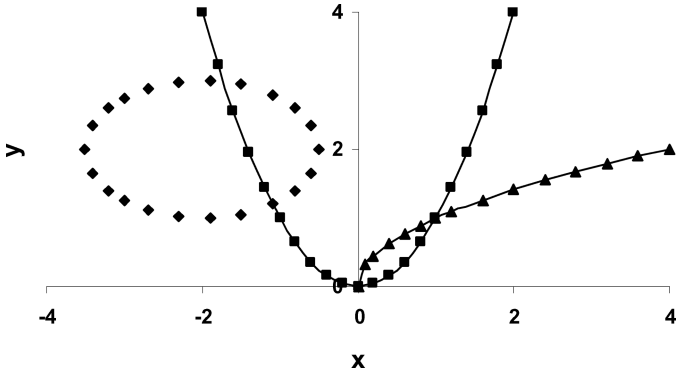


FIG. 4. Linear independence and functional dependence: elipse (■) $(x-a)^2+(y-b)^2=k$, and non linear relationships $y=x^2$ (parabole) and $y=\sqrt{x}$ (fractionary power).

CORRELATION AND NONLINEAR ASSOCIATION

The quantities $\mathrm{cov}(x, y)$ and r do not, in general, convey all there is to know about the association between two variables. Nonlinear association can exist that are not revealed by this descriptive statistics (40). Data may be highly correlated but have a zero linear correlation coefficient, as it occurs (41) with the function $y=x^2$ (Figure 4) for which $r=0$. However, if we restrict x values in Figure 4 to positive values, the correlation coefficient is 0.963. The correlation coefficient of (x_i, y_i) points for $y=\sqrt{x}$ in Figure 4 is 0.966 instead. Then, the correlation coefficient may remain less than unity, although the relationship between x and y is rigorously functional one. Plenty of physical laws, for example, lead to quadratic relation (31) of the form $y=a_0+a_1x+a_2x^2$. It should be noted that r examines only possible linear relationships between sets of continuous, normally distributed data. Other mathematical relationships (log, log/linear, exponential, etc.) between data sets exist which require either the use of another correlation testing method or that one or more of the data sets be transformed (Table 2) so that they are of linear nature (42).

CORRELATION AND HETEROSCEDASTICITY

Minimal measurement error is assumed since low reliability attenuates the correlation coefficient (43). The conventional test of correlation coefficient is derived under the assumption that x and y are independent. An implication of this assumption is that the error term, when predicting y from x is homoscedastic (44). Otherwise the correlation coefficient is a misleading average of points of higher and lower correlation (43). Imperfect validity of one or both variables in the correlation degrades the apparent relationship between them (45). If variance is truncated or restricted in one or both variables due, for instance, to poor sampling, this can also lead to attenuation of the correlation coefficient. Heteroscedasticity plays a role when applying a conventional method aimed at establishing whether two variables

are dependent. This problem may be substantially corrected by using the modified percentile bootstrap method. Readers interested in this issue can refer to reference (44).

CORRELATION AND OUTLIERS

Covariance and correlation coefficient can be very sensitive to "wild" observations ("outliers") and may indicate association when, in fact, little exists. Suspect observations must be accounted for by correcting obvious recording mistakes and by taking actions consistent with the identified causes (40). Note that over a few high or low points can have a large effect on the value of r; therefore it is useful to inspect a plot of the data and be sure that the data covers the range in a fairly uniform way. Depending upon where the outlier falls r may be increased or decreased. In fact, one point, properly placed, can cause the correlation coefficient to take on virtually any value between -1 and $+1$ (Figure 5A, B), so care must be taken when interpreting the value of r (44). However, remove points of influence (leverage, bias and outlying points) only if a reason can be found for their aberrant behaviour. The smaller the sample size the greater the effect of the outlier. Problems with outliers can be minimized by making duplicate measurements, carefully inspecting and plotting the data at the time it is collected and retesting discrepant results while the specimen are still available (46).

As examples of strategies where we guard against both unusual x values and y values we get the Winsorized correlation coefficient (44), which compensates for this by setting the tail values equal to a certain percentile value. Then, the standard correlation formula is applied. Spearman's rho converts the observations to so-called ranks.

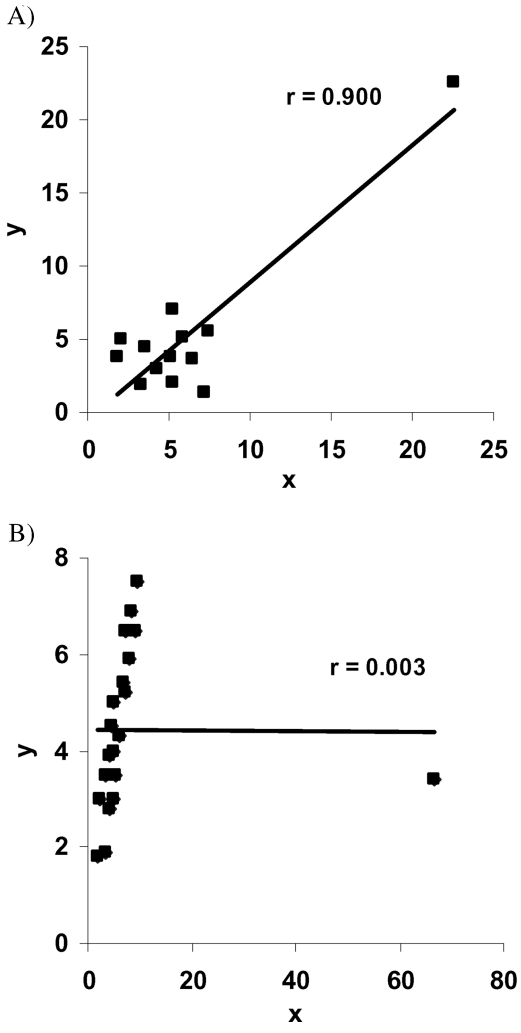


FIG. 5. A simple aberrant point can alter r by a large amount. The solid lines are least squares regression line. A: $r = 0$ without the extreme value. B: $r = 0.917$ without the extreme value.

CORRELATION, STRATIFICATION AND INTERVAL PROPERTY REQUERIMENT

Sometimes, the scatter diagrams do not reflect satisfactorily the relationship between variables (47). In Figure 6A data corresponding to random variables raw matter and its resistance are compiled as such. The plot of points is a random cloud with no visible pattern. In Figure 6B, data were classified according to the place where the raw matter was acquired, identified with different symbols. Then, when it is possible to stratify one of the variables it is recommended to elaborate diagram such as, by using different colours or symbols. It is possible even the contrary phenomenon to occur: correlation may not be detected (47) if data are stratified, but detected when we ignore it. When experimental data are obtained in controlled production conditions, they may not reflect the existence of relationship in spite that theoretically a variable is influenced by the other (Figure 7). Thus, correct interpretation of a correlation coefficient requires the assumption that both variables, x and y, meet the interval property requirement (44) of their respective measurement systems.

r VALUES

The basis of the correlation coefficient concept can be seen if two straight lines at right angles, and parallel to the axes, are drawn to the point representing the mean values x and y, as shown in Figure 8, where the dissolution speed of a streptomycin powder (a rapid dissolution being appreciated by the patient) is depicted versus the density of streptomycin solution previous to

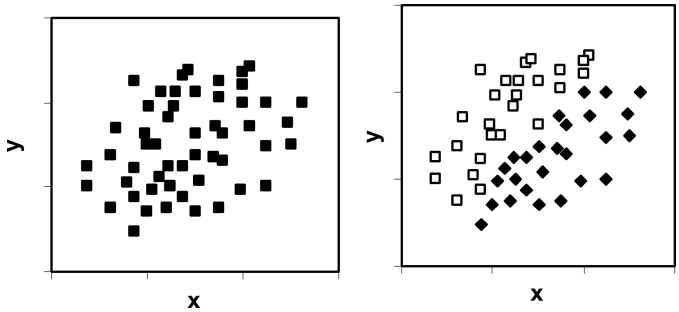


FIG. 6. A: overall scatter diagram. B: scatter diagram with stratified data.

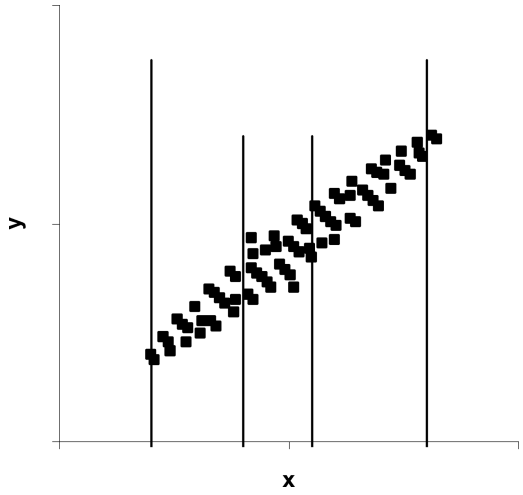


FIG. 7. A restricted range value may mask a linear relationship.

the dry step (48). If y is positively related to x, the majority of the points will be located in the $++$ upper right and $--$ lower left quadrants, if they are negatively related, the points will lie in the $-+$ upper left and $+-$ lower right quadrants, and if they are unrelated, the points will uniformly located in all four areas. When $r = 0$, the points scatter widely about the plot, the majority falling roughly in the shape of a circle. As the linear relationship increases, the circle becomes more and more elliptical in shape until the limiting case is reached ($r = 1$ or $r = -1$) and all the points fall on a straight line.

What constitutes a satisfactory correlation coefficient is dependent on the purpose of which is to be used, and on the nature of raw data. Table 4 provides a rule-of-thumb scale for evaluating the correlation coefficient. The greater n is, the lower the acceptable correlation coefficient. Correlations even as high as 0.6 do not look that different from correlations of 0.1; i.e., they do not mean much if the goal is to predict individual values of one variable from the other (49). For a linear regression between two physical properties involving five pairs of results, a correlation coefficient in excess of 0.990 would be sought, but In the quantitative structure-activity relationships (QSAR) of medicinal chemistry, coefficients of around 0.95 are often quoted (50). In some procedures, for example cluster analysis, considerably lower correlation coefficients can be considered important.

Atomic absorption spectrophotometry (AAS) and differential pulse stripping voltammetry (DPSV) as independent methods

TABLE 4
Strength of correlation (46)

Size of r	Interpretation
0.90 to 1.00	Very high correlation
0.70 to 1.89	High correlation
0.50 to 0.69	Moderate correlation
0.30 to 0.49	Low correlation
0.00 to 0.29	Little if any correlation

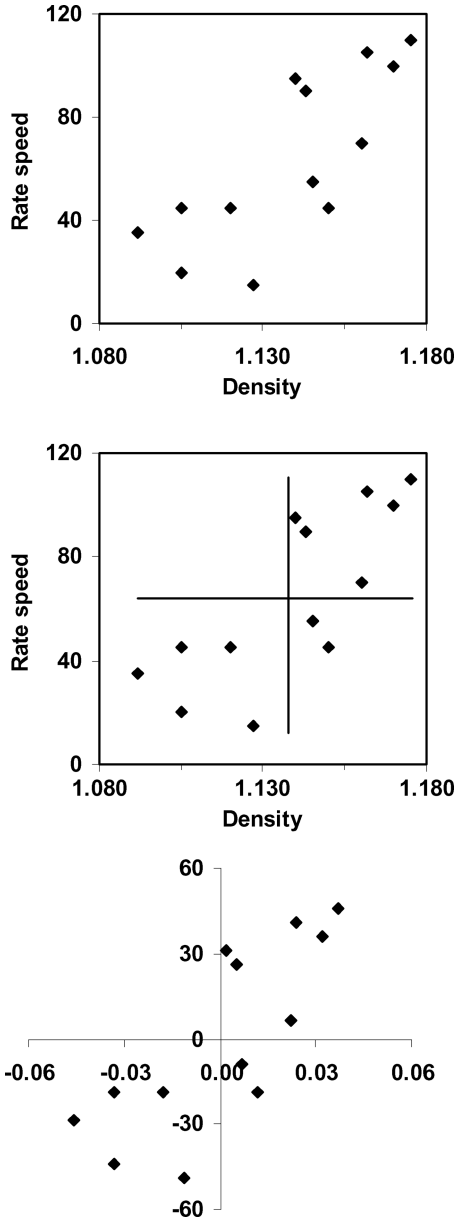


FIG. 8. Dissolution rate of a streptomycin power. A: scatter diagram. B: axis at $\bar{x}$ and $\bar{y}$ included. C: centered data.

were used in connection with an appropriate digestion or pretreatment procedure for the analysis of trace and ultratrace of metals in environmental and biological matrices (51). Some of the results obtained are summarized in Figure 9. The various correlation diagrams depicted should give the reader an intuitive feeling for the concept of correlation. Notice that the plots show increasing scatter as the r value decreases toward 0.

From Eq. (8) we derive a simple computational formula, namely

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{\sum x_i y_i - \frac{\sum (x_i y_i)^2}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}. \quad [9]$$

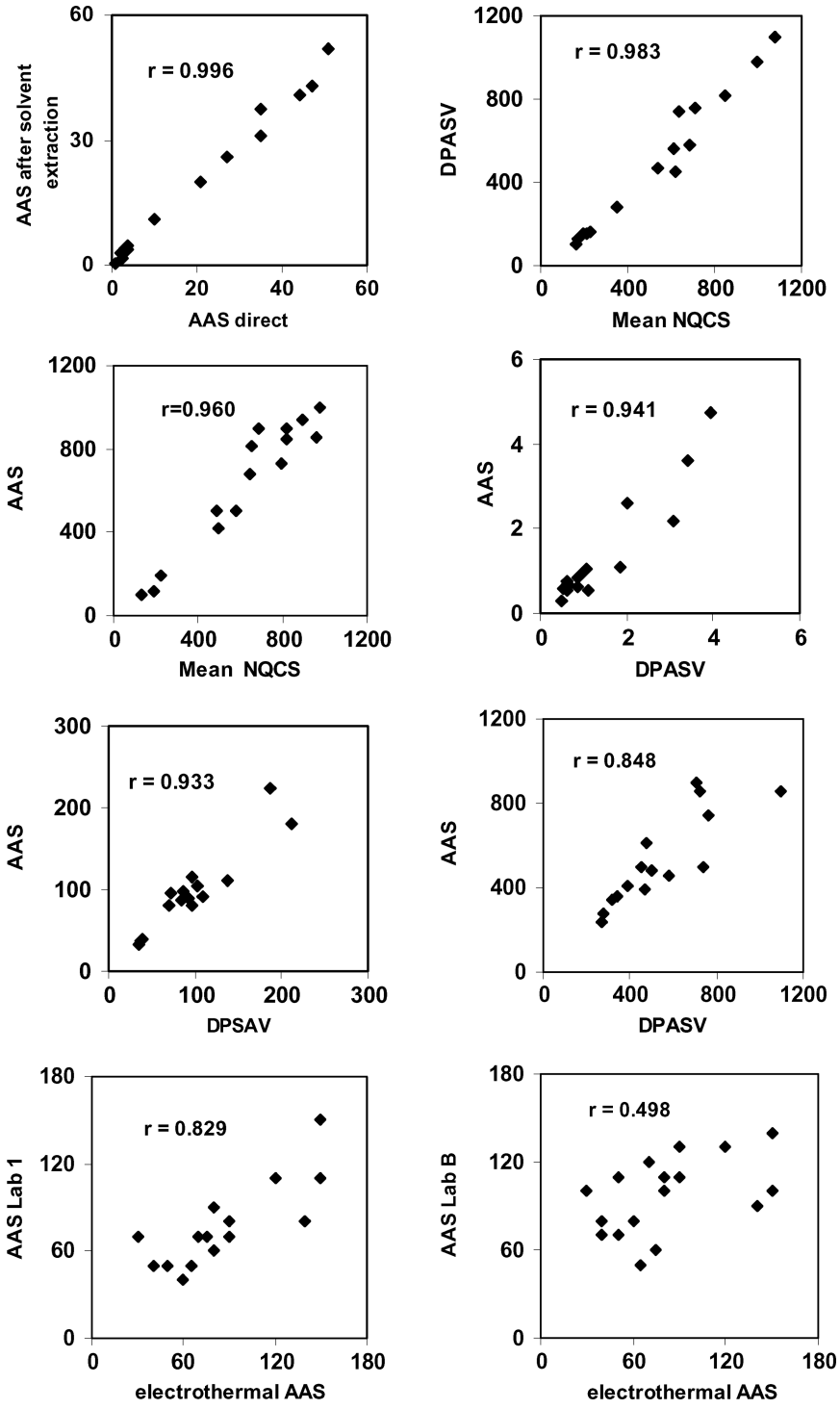


FIG. 9. Method comparison for the detection of trace levels of cadmium and lead in biological fluids (in ng/ml) by atomic absorption spectrometry (AAS) and differential pulse stripping anodic voltammetry (DPSAV). NQCS: British National Quality Control Scheme (51).

One advantage of this equation over Eq. 8 is the fact that rounding off is not necessary until the square root and the final division are taken. However, the form of Eq. (9) is unstable. It has the property that it suffers (52) from subtractive cancellation for data set with small coefficient of variation. A method designed for hand calculation with a few observations, each of which is small in magnitude, is not appropriate for use (52–54) in a computer program, which may handle a great many observations whose magnitude may be large. In practical terms r is the ratio of two very similar numbers. The numerators and denominators in Eq. (9) must therefore be calculated using a large number of significant figures, and the value of r should be rounded (often to two, three or four decimal places) at the end of the calculation. Eq. (9) then requires double precision or better. If this precaution is not taken, r values > 1 can easily be obtained in error (55). There is no point in quoting r to more than three or at most four decimal places, unless one is absolutely sure that the algorithm is not the limiting factor (56). Note that r according Eq. (9) is invariant to any linear transformation of either variable, i.e., to both shift of the intercept and change of scaling (57). However, it is affected by rotation of coordinates.

In summary, the following factors (44) influence the value of r: (i) the magnitude of the residuals; (ii) the slope of the line around which points are clustered; (iii) outliers; (iv) curvature; (v) a restriction range. So, if we are told r, and nothing else, we cannot deduce much about the details of how x and y are related. If the data do not have a normal distribution either or both variables can be transformed, a non-parametric correlation coefficient can be calculates. People perceives association not as proportional to the correlation coefficient but as proportional to $1 - \sqrt{1 - r^2}$ (49).

COEFFICIENT OF MULTIPLE CORRELATION R, CORRELATION COEFFICIENT r, AND COEFFICIENT OF ALIENATION k

The correlation coefficient, r, is a standardized index for which the value does not depend on the measurement scales of the variables. Its values lies (58) in the range $(-1, 1)$, and its squared value describes (59) the proportional reduction in variability of one variable when the other is held constant. It is worth noting that since for any r^2 other than 0 or 1, $|r^2| < r$; r may give the impression of a closer relationship (60) between x and y than does the corresponding r^2. Therefore, although the correlation coefficient is by itself an important measure of relationship between the variables, it is R squared that permits comparison of the strengths of relationships.

The reasons for making a distinction (41, 5) between r and R are that i) r is a measure of association between two random variables, whereas R is a measure between a random variable y and its prediction $\hat{y}$ from a regression model; ii) R is always well defined, regardless of whether the independent variable assumed to be random or fixed. In contrast, calculating the correlation between a random variable, y, and a fixed predictor variable x, that is, a variable that is not considered random, makes no sense. Because r^2 gives the proportion of variance that is common between the two variables x and y, the uncommon or unique variance is the remainder, and this is known as the coefficient of nondetermination and is usually symbolized as $k^2 = 1 - r^2$. Some statisticians refer to (61) the coefficient of alienation k, which indicates the degree of lack of relationship.

POPULATION AND SAMPLE CORRELATION COEFFICICIENT

The population analogue of r, i.e., the value of r if all subjects could be measured, is typically labelled ρ. It can be shown that the expected value and variance of r are defined approximately by (62)

$$E(r) = \rho\left(1 - \frac{1-\rho^2}{2n} + \cdots\right) \approx \rho \qquad [10]$$

$$\sigma^2(r) = \frac{(1-\rho^2)^2}{n}\left(1 + \frac{11\rho^2}{2n} + \cdots\right) \approx \frac{(1-\rho^2)^2}{n} \qquad [11]$$

respectively, provided that n is not too small. The Pearson correlation coefficient of linear correlation r has a complicated distribution involving special functions (59); with $\rho \neq 0$ and $n > 3$ may be expressed in the form

$$p(r) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}} \times \sum_{i=0}^{\infty}\left[\Gamma\left(\frac{n+i-1}{2}\right)\right]^2 \frac{(2\rho r)^i}{i!} \qquad [12]$$

for $-1 \leq r \leq 1$, $-1 \leq \rho \leq 1$; $= 0$, elsewhere.

As ρ approaches $+1$ or -1, the sampling variance decreases, so that when ρ is either at $+1$ or -1, all sample values equal the parameter and the sample variance is zero. The shape becomes increasingly normal with large values of n, and becomes increasingly skewed with increasing $|\rho|$. The sample variance and thus the significance test, depend upon the size of the population correlation and the sample size. When $\rho = 0$, r is symmetrically distributed about zero, and the mean of the sampling distribution does equal the parameter. As ρ increases from zero (becomes positive), the sampling distribution becomes negatively skewed. As ρ becomes negative, the sampling distribution becomes positively skewed.

Table 5 gives the 5% significance values for varying numbers of points. The significance of i.e., $r = 0.9$, is much greater when the sample size is very large than when the sample size is very small. For a small sample size, the alignment of the (x_i, y_i) data points along a straight line may be fortuitous. However, when many data points lie along a straight line, the case becomes much more convincing or significant. Care should be exercised when using significance tables, since some tables are one-tails and others are two-tailed (63). A one-tailed test should be used when only a direct relationship or only an inverse relationship between the x and y values is of importance, while a two-tailed test should be used whenever both direct and inverse relationships between

TABLE 5
Significance values of r
for varying number of (x, y) points

	5% significance value for Pearson's r	
n	Two tailed	One tailed
3	0.997	
4	0.950	
5	0.878	0.805
6	0.811	0.729
7	0.754	0.669
8	0.707	0.621
9	0.666	0.582
10	0.632	0.549
11	0.602	0.521
12	0.576	0.497

the x and y values are equally important. Look up r in the table (ignoring + or − sign). If our calculated r-value exceeds the tabulated value at $p = 0.05$, then the correlation is significant. The number of degrees of freedom is two less than the number of points on the graph. The correlation coefficient r must lie in the range $0 \leq r \leq 1$, but in practice, because of random errors, $0 < r < 1$.

CAUSE AND EFFECT INFERENCE FROM THE CORRELATION COEFFICIENT?

If two random variables x and y are statistically independent, their correlation coefficient is zero. However, the converse is not true; i.e., if $r = 0$, this does not necessarily imply that x and y are statistically independent (64). The correlation coefficient is thus an estimate of association between the variables and is valid only when the observations are randomly drawn. Many statistical software packages include a program for such calculation and the correlation coefficient r, is routinely printed out in connection with other statistical parameters.

The correlation coefficient provides information about random error, but r cannot be easily interpreted (65, 66) and therefore it is of no practical use in statistical analysis of comparison data. Caution must be exercised in drawing inferences about cause and effect (41) from correlation coefficients. The correlation coefficient is often (67) misunderstood. A positive correlation simply means that y is believed to increase when x increases. However, it must not be considered necessarily to indicate a causal relationship. There must be something that causes both to change. One should be keenly aware of the common occurrence of spurious correlations due to indirect causes or remote mechanisms (68).

Perhaps, the best known example where overwhelmingly strong evidence of statistical correlation was adduced is in the classic studies of Fisher (69, 70) on the relationship between cigarette smoking and the incidence of lung cancer. The incidence of lung cancer depends on many different factors; there is no true relationship between the number of cigarettes smoked and the date of incidence of the disease. Smoking is simply one factor, an important one, which affects the chance of incidence.

In either case the determination of causation involves a scientific study of the subject, possibly using additional experimental data, statistics being merely one of the more powerful tools in arriving at the right answer. Sociologists have developed a branch of correlational analysis, called path analysis, precisely to determine causation from correlation (49).

CRUDE TEST FOR LINEARITY?

The correlation coefficient, r, has often been used as a crude test for linearity on the grounds (68) that a linear calibration function nearly always gives a high correlation coefficient. However, the converse is not true: a correlation coefficient close to unity does not necessarily indicate a linear calibration function. Moreover, the numerical value of r cannot be interpreted in terms of degree of deviation from linearity, and so, the experimenter should be very careful when drawing conclusions on the basis of these deceptively simple numbers (67, 72–73). In a residual analysis, the difference for each data point between the true y value as determined from the best fit are plotted for each x value of the data points. A residual plot is the best indicator of goodness of fit of experimental data (74) giving useful information to validate the chosen regression model. The residual plot can be used to check whether the underlying assumptions, like normality of the residuals and homoscedasticity, are met as for evaluating the goodness of fit of the regression model (75). Even rather, poor calibration graphs, i.e., with significant y-direction errors, will have r values close to 1 (or −1), values of $|r| <$ about 0.98 being unusual. Even worse, points that clearly lie on a gentle curve can easily give high values of $|r|$.

Many analysts depend entirely (76) on the use of r^2 (or r) value between 0.999 and 1.000 as an acceptability criterion. This is well known to be inadequate and many chemometrics experts have expressed concern that publications are still accepted with this minimum data. By itself, the correlation coefficient gives only a relative idea of the linearity inherent in a particular data set (53) and supplementary criteria are needed. Data sets in which non-linear deviation is concentrated (77) in 1 or 2 areas of a graph (e.g., at the upper and lower ends) and data sets that are subject to slight but overall persistent curvature are prime examples of this problem. Very large percentage errors at the lower end of concentration range can coexist (76) with acceptable correlation r^2 and are grossly underestimated by confidence limits from an analysis of the errors in slope and intercept.

Carbon dioxide is a material of interest as supercritical solvent and has been used for, e.g., removing caffeine from coffee. The data for its vapour pressure as a function of temperature are by no means linear (78, 79). We may expect, however, $\ln P$ versus $1/T$ to be linear (Clausius–Clapeyron equation) (Figure 10A). Similar results to those depicted in Figure 10A

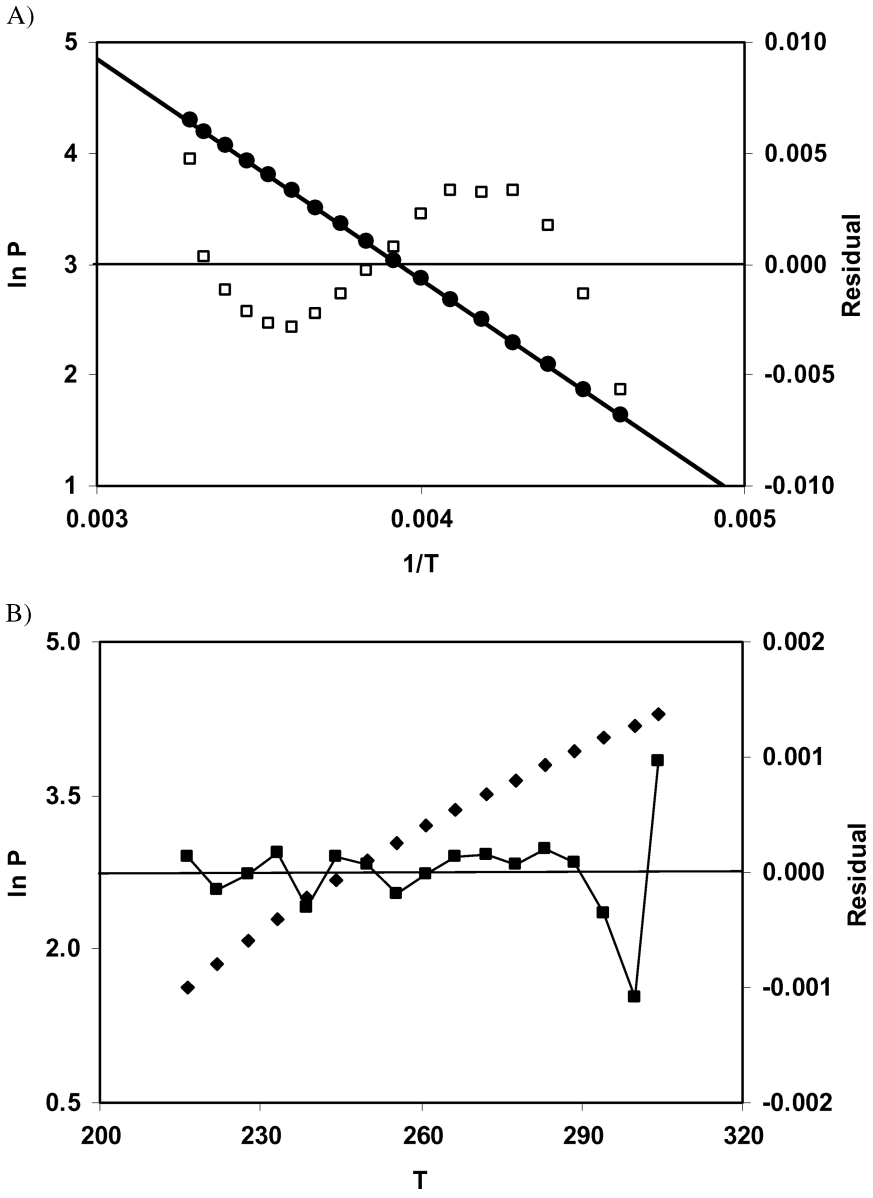


FIG. 10. Vapor pressure of carbon dioxide as a function of temperature according to the Clausius-Clapeyro equation (A) and a more complicated model (B).

were obtained, however, when a weighting factor of $w_i = P_i^2$ is applied on the basis of the transformation being used (21). The residual plot can be incorporated to the line resulting from the least squares fit of the model to the data to show the extent of the agreement between data and model. Results (Figure 10A) led to a 0.999988 76 correlation coefficient for the fit (Excel 2002). This near perfect fit was in fact, a very bad one in terms of the potential quality of the fit as indicated by the residual pattern. The error is not in the data, but in the model. A more general form of that equation is

$$\ln P = A + \frac{B}{T} + C \ln T + DT + ET^2. \qquad [13]$$

The results obtained in this case (multiple linear regression analysis) are greatly superior to the linear equation, the residuals being randomly scattered (Figure 10B).

A Gamma-ray energy versus channel number calibration (80) led to a 0.999999 correlation coefficient for the fit (given by a computer program limited to six decimal digits). This near perfect fit was in fact, a very bad one in terms of the potential quality of the fit (residual pattern). Although r was > 0.999 for a HPLC method for mannitol (internal standard) based on the measure of the peak area ratio versus mannitol concentration for standards, the plot indicates (81) deviations from linearity at low and high concentrations. In

consequence, analysts should avoid being misled by the correlation coefficient.

Actually this quantity is intended as a measure of statistical relationships and therefore it has little to do with functional relations such as calibration curves; it ought not to even appear (71).

CORRELATION COEFFICIENT AND SLOPE

Correlation is simply a scaled (by the ratio of the spread of the x_i divided by the spread of the y_i) version of the slope (82), i.e., the slope estimate multiplied by a factor to keep r always between -1 and $+1$

$$r = r_{xy} = a_1 \frac{s_x}{s_y} = r_{yx} = a_1' \frac{s_y}{s_x} \qquad [14]$$

(a_1' is the slope of the regression line of x on y). The level of significance of correlation given by the correlation coefficient is the same (83) as the level of significance of the slope of the regression line given by the t test of $t = a_1/s_{a_1}$. In consequence, standard errors of the parameters are the first thing to look at in a regression printout, in order to see if all parameters are significant, and second at the residuals of the regression to see if these are random. Only if one has more than one regression on more than one model to the same data should they—third—compare r-values, or better z-values, as shown in the following. We can note that r is symmetrical with respect to x and y being also given by a geometric mean (7)

$$r = \sqrt{a_1 a_1'}. \qquad [15]$$

THE FISCHER z-TRANSFORMATION FOR r

With a small-size sample and a relatively close correlation, the distribution of the correlation coefficients substantially differs from normal. What we require (84) is a transformation to a measurement scale that will give: (a) homogeneous variance; (b) at least approximate normality; (c) non-constrained response values; with d) a simple form for the expected value of the response variable; and (e) independence of the observed response values. Still another alternative for the test of significance when n is only moderately large (>10) uses the Fisher's z-transformation to associate each measured r with a corresponding z(3).

Let g represent ρ expressed in a transformed scale defined by

$$g = f(\rho). \qquad [16]$$

By applying the law of random error propagation (31) we get

$$\sigma_g^2 = \left(\frac{\partial f(\rho)}{\partial \rho}\right)^2 \sigma_\rho^2. \qquad [17]$$

As we wish to select the function $f(\rho)$ in such a way that σ_g^2 be a constant C^2, then from Eq. (16), the transformation of scale to achieve homocedasticity is given (85) by the differential

$$df(\rho) = \frac{C d\rho}{s_\rho} = \frac{C\sqrt{n}}{1-\rho^2} d\rho, \qquad [18]$$

which upon integration led to

$$g = f(\rho) = C\sqrt{n} \int \left(\frac{1}{1-\rho} - \frac{1}{1+\rho}\right) d\rho$$
$$= C\sqrt{n} \ln\left(\frac{1+\rho}{1-\rho}\right) + K, \qquad [19]$$

where K is an arbitrary constant (usually for $C = 1/\sqrt{n}$ and $K = 0$). The function

$$E(z) = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) \qquad [20]$$

is known as Fischer z transformation for the correlation coefficient, which has a variance of

$$\sigma_z^2 = \frac{1}{n-3}. \qquad [21]$$

For sample values, z, s_z^2 and r, replaces $E(z)$, σ_z^2, and ρ, respectively, in Eqs. (20) and (21).

Relative to the correlation coefficient z has a simpler distribution; its variance is more nearly independent of the corresponding population parameter; and it converges more quickly to normality (86). The approximation would be helpful only when $\mu = E(z)$ is large compared with σ. This does often happen in problems of physical interest, i.e., in the context of calibration and parameter estimation studies. The inclusion of further terms of the expansion for the mean and variance of z (in Eqs. (10) and (11)) increases the accuracy of the approximations considerably, and the values using the second approximation for z are closer to the exact values than any (87).

Table 6 gives a number of r values and the corresponding z values. Note that if the correlation is negative, the z value should be negative: $\tanh^{-1}(r) = -\tanh^{-1}(-r)$. Fisher z values are much practical, because basically, count the nines of r for you. Besides, they tend to have a normal distribution and are much easier to work with.

The computation of the correlation coefficient distribution according to Eq. (12) can be easily carried out with a pocket calculator (88). We have used the algorithm proposed in the Gunther paper (88) to write a program in FORTRAN and the corresponding results for $\rho = 0.999$ and $n = 12$ were collected from $r = \rho - 1.96(1-\rho^2)/\sqrt{n}$ to $r = \rho + 1.96(1-\rho^2)/\sqrt{n}$ with

TABLE 6
Some selected r values and the corresponding z values

r	z	r	z
0.300	0.310	0.999	3.800
0.500	0.549	0.9999	4.952
0.700	0.973	0.99999	6.103
0.900	1.472	0.999999	7.254
0.997	3.250	0.9999999	8.406
0.998	3.453	0.99999999	9.557

a step suitable to obtain about 100,000 points. The obtained distribution is skewed (Figure 11A). By applying the Fisher transformation

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} \qquad [22]$$

the distribution tends to a normal one with population mean given by Eq. (20) and population standard deviation given by Eq. (21). The skewness close to zero (0.00063) and the kurtosis very near to three (2.9955) agree well with the normal distribution (62). This finding can be proved by Monte Carlo simulation (100,000 trials) using Crystal Ball software (89). The corresponding outputs are included in the Excel file (Figure 11A).

The same r and z distributions were obtained for 365 samples instead of 100,000 in order to represent the possible r values obtained during 1 year from daily calibration. In this later case, taking into account the lesser number of iterations, 365, Monte Carlo simulation being carried out by Latin hypercube sampling. Results are depicted in Figure 11B. The skewness in this latter case is 0.05 and the kurtosis 3.13.

CONFIDENCE INTERVAL AND TESTS OF SIGNIFICANCE FOR CORRELATION COEFFICIENT

The degree of uncertainty of the correlation coefficient can be estimated (90, 91) if x and y have a joint bivariate normal distribution. Confidence intervals may be constructed in the usual way by constructing a confidence interval in the transformed scale and transforming then back into the original scale. We can construct a 95% confidence interval for z as being from

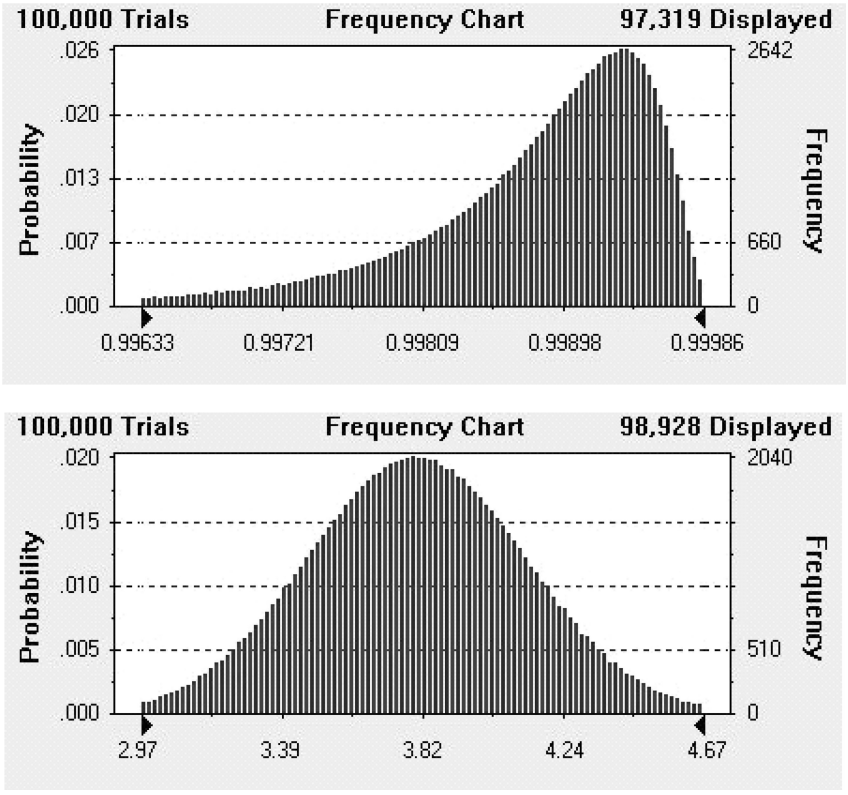
$$z_1 = z - 1.96/\sqrt{n-3} \quad \text{to} \quad z_2 = z + 1.96/\sqrt{n-3}. \qquad [23]$$

For the 90 % confidence interval, the standard error is multiplied by 1.645 and for 99% by 2.576. We back transform the above values to get a confidence interval for the population correlation coefficient r as

$$\frac{e^{2z_1}-1}{e^{2z_1}+1} \quad \text{to} \quad \frac{e^{2z_2}-1}{e^{2z_2}+1}. \qquad [24]$$

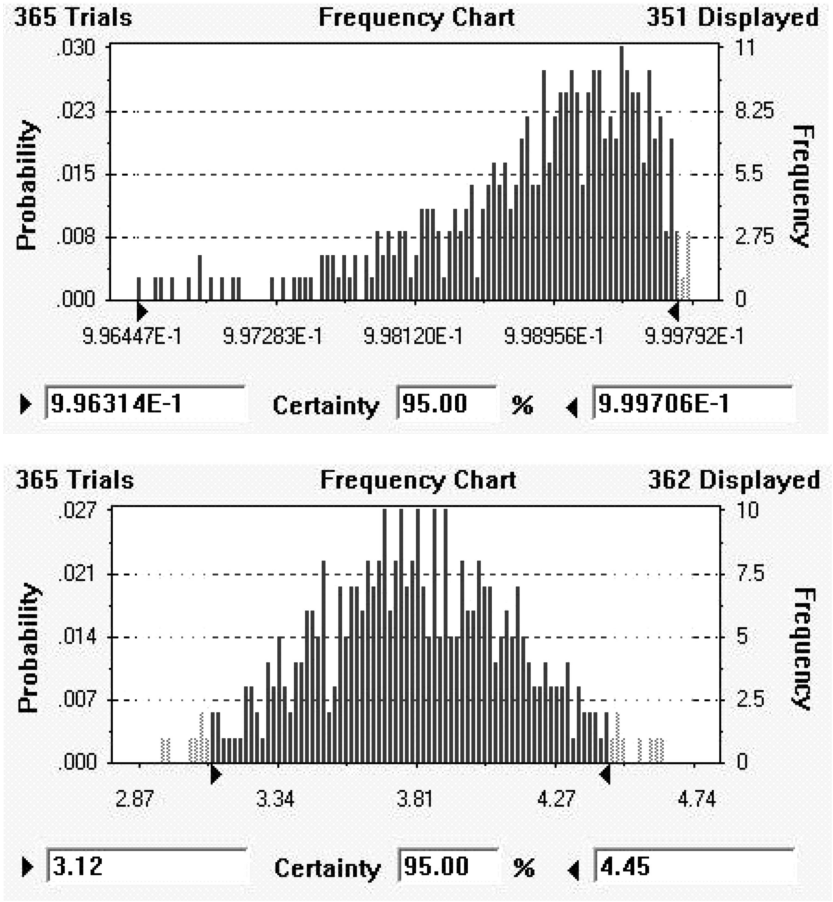
Example 1

The experimental and calculated values of $\ln(1/IC_{50})$ have been determined (92) for a series of 42 1,2-diarylimidazole



(A)

FIG. 11. (A) Simulation run of 100000 r's from a bivariate population having a theoretical correlation coefficient of 0.999 (n = 12) together with their corresponding z distribution values. (B) Simulation run of 365 r's from a bivariate population having a theoretical correlation coefficient of 0.999 (n = 12) together with their corresponding z distribution values. *(Continued)*



(B)

FIG. 11. (Continued)

derivatives with cyclooxygenase inhibitory activities in a QSAR analysis. Molar concentration causing 50% inhibition of enzyme was expressed as IC_{50}. For the data obtained, the correlation coefficient was estimated as 0.896. Then, $z = 1.4516$, z_2=1.7655, z_1=1.1377, and the 95% confidence interval for the correlation coefficient is 0.813 to 0.943. Notice that the confidence limits in the above example are not spaced equally on each side of the observed value. That happens with non-normally distributed statistics like the correlation coefficient. Most other statistics are normally distributed, so the observed value falls in the middle of the confidence interval.

Often, it is necessary to determine whether one correlation is significantly different from another ($H_0 : \rho_1 = \rho_2$), the difference being computed between different-sized random samples. To perform this test, convert the two correlations to z-scores (z_1 and z_2) and estimate the standard error of the difference between the two correlations as

$$s = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}, \qquad [25]$$

where n_1 and n_2 are the sample sizes of the two independent samples. Divide the difference between the two z-scores by the standard error to yield a normal curve deviate (expected value equal to 0 and variance equal to 1).

$$U(z) = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}. \qquad [26]$$

If the z value for the difference computed is 1.96 or higher, the difference in the correlations is significant at the 0.05 level.

Example 2

The calibration curve in Figure 12A, shows fluorescence readings of a series of standards (93). The calibration line with r= 0.995 21, is noticeable curved. In order to check for systematic deviations between data and model, it is recommended that a plot be made of the resulting residuals (ε) against the y values as shown in Figure 12A). The pattern observed in the residual plot shows the equation being fitted is inadequate and should possibly contain higher order terms to accommodate the curvature. One way to handle a curved calibration line is to fit

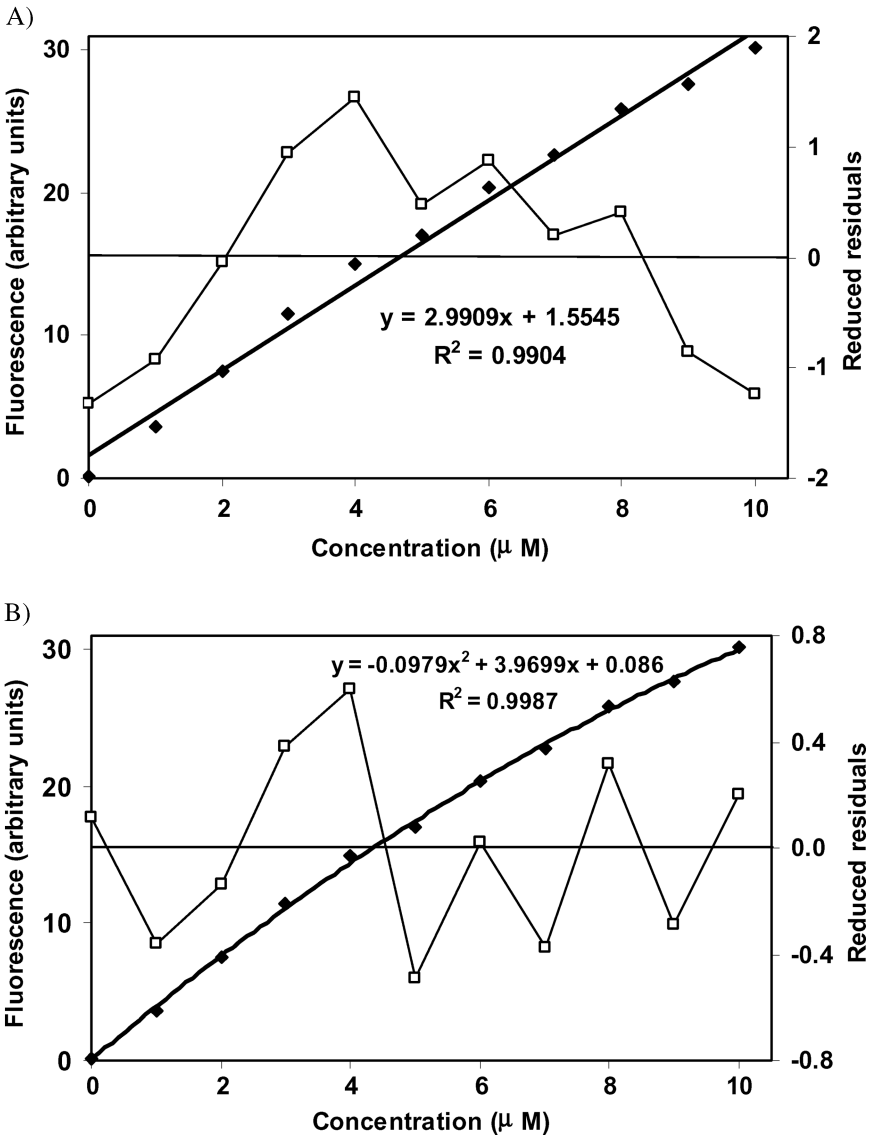


FIG. 12. Fluorescence readings of a series of standards. A: $y = a_0 + a_1 x$ model. B: $y = a_0 + a_1 x + a_2 x^2$ model. Reduce residual = $\varepsilon / s_{y/x}$.

the line to a power series. A quadratic equation (Figure 12B) is usually sufficient to fit: the scatter above and below the zero line is about the same in this case.

Example 3

We wish to test if there is a significant difference between the correlations, 0.769 and 0.711, corresponding to the resistance versus raw matter stratification example of Figure 5A,B. Then, $z_1 = 1.018$, $z_2 = 0.889$, and $U(z) = 0.437$, much smaller than 1.96, and thus not significant at the 0.05 level.

CORRELATION IN MULTIVARIATE ANALYSIS

In multivariate experimentation (e.g., with two variables) it is often desirable to choose one's experiments (e.g., the concentration of two reactants in a kinetic experiment) such as there is no correlation in the design. This is the basic principle of experimental design (94). All in all, when the experimenter plans his experiments, he achieves desired results in a more economical manner (95). A convenient way of summarizing a large number of correlation coefficients is to put them in a single table, called a correlation matrix. The correlation of any variable with itself is necessarily 1. Thus the diagonals of the matrix are the unity. As the correlation coefficient is non-directional, $r_{kj} = r_{jk}$. So, the correlation matrix is symmetrical around the diagonal, and only $n^2 - n - (n-1)/2 = n(n-1)/2$ terms need to be computed (96). The correlation matrix plays a central role in multivariate statistics and is used when dealing with distances and display methods (97). Correlations, especially in many variables, can become a

very messy subject. For this reason, it is highly desirable where possible to choose the variables in a multi-dimensional problem in such a way that the errors are normal and uncorrelated.

CONCLUDING REMARKS

Two scatterplots with the same statistical information can appear different because our ability to process and recognize patterns depends on how the data are displayed (98, 99). Correlation is usually applied to relationship of continuous variables, and is best visualized as scatterplot or correlation diagram. The correlation coefficient r has been interpreted in great many ways (5). The meaning of this concept is not easy to grasp (100). Regression and correlation are very closely related. In fact the t-test of the null hypothesis of zero correlation is exactly equivalent to that for the hypothesis of zero slope in regression analysis—the two values are identical (80). We note, however, that a measure of statistical relationship, such as a correlation coefficient should never be used to deduce a causal connection; our ideas on causation must come from outside statistics. Statistical calculations that neglect correlations often result in incorrect results and erroneous conclusions (41).

The correlation coefficient in its many forms has become the workhorse of quantitative research and analysis (101). And well it should be, for our empirical knowledge is fundamentally of covarying things. The interpretations of the r statistic, however, can be completely meaningless if the joint distribution of the variables x and y is too different from a binormal distribution (102). When the joint distribution of random variables is not normal and the sample contains strong outliers, the normalized transformation is not valid and the correlation coefficient is not suitable for expressing a stochastic association. We can then use various robust estimates of correlation coefficients, which apply robust estimates of parameters of location, spread and covariance (103).

Although correlation is a symmetric concept of two variables, this is not the case for regression where we distinguish a response from an explanatory variable. When two variables are functionally related (104, 105), it is meaningless to calculate a r. This will often be apparent from a causal look at a graph of the data. It is when the data scatters markedly and when the graph is hard to draw that the r may begin to take importance. In fitting functional models, values of r and R^2 close to $+1$ or -1 do provide (105) an aura of respectability, but not much else. In addition, although the correlation coefficient is conceptually simple and attractive, and is frequently used as a measure of how well a model fits a set of data, it is not, by itself, a good measure of the factors as they appear in the model (11, 106), primarily because it does not take into account (107–110) the degrees of freedom.

A high value of r is thus seen to be no guarantee at all that a straight line rather than a curve, is appropriate for a given calibration plot. It is for this reason that the plot must always be inspected visually. It is surprising, given these obvious inadequacies, that the r, had been used so frequently to assess the linearity of calibration graphs and other plots. There are many warnings in the literature (55, 111–113) on the dangers of doing this. In short, the r value is in reality not a measure of model adequacy (74) and then, r can, or has to be replaced by other statistical tools such as rank correlation, regression analysis and specific tests of linearity. Residual analysis on this respect is a good way to decide if a linear fit was a good choice (10).

In summary, the r parameter is of limited value due to that r values cannot be meaningfully compared if they are obtained from curves based on different standards, curve shape can change without necessarily affecting r, r values do not yield quantitative comparison of data quality, even if the same standards are used and the r value does not indicate whether the chosen mathematical model (e.g., a first-or second-order equation) adequately fits the data. In addition, a correlation coefficient does not give indication of the error associated (75) with an individual measurement.

When r is tested to be not equal to zero the use of the Fisher transformation produces a statistic that is distributed normally. This transformation is referred to as Fisher's z transformation; that is, the z-statistic. Fisher z-values are much more practical than r values because they tend to have a normal distribution and are therefore much easier to work with. Apart from its normalizing properties, the great practical advantage of the z transformation is that it gives a variance which, to order $(n-3)^{-1}$ is independent of r.

REFERENCES

1. W. G. Warren, Correlation or regression: bias or precision. *Applied Statistic* 20 (1971):148–164.
2. B. A. Moore, Correlation and regression analysis: applications to the analysis of chemical data. *Analytical Proceedings* 17 (1980):124–127.
3. S. Aknazarova and V. Kafarov, *Experiment Optimization in Chemistry and Chemical Engineering* (Moscow: Mir Publishers, 1982), 123–127.
4. F. Galton, Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London* 45 (1888):135–145.
5. J. L. Rodgers and W. A. Nicewander, Thirteen ways to look at the correlation coefficient. *The American Statistician* 42 (1998):59–66.
6. R. A. Nadkarni, The quest for quality in the laboratory. *Analytical Chemistry* 63 (1991):675A–682A.
7. ISO 9004-4:1993, *Quality Management and Quality System Elements. Part 4: Guidelines for Quality Improvement* (Geneva: ISO, 1993).
8. *IUPAC Compendium of Chemical Terminology*, 2nd ed. (1997).
9. A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics* (Dordrecht: Kluwer, 2003), 79.
10. N. Draper and H. Smith, *Applied Regression Analysis*, 3th ed. (New York: Wiley, 1998).
11. S. N. Deming and S. L. Morgan, *Experimental Design, a Chemometric Approach* (Amsterdam: Elsevier, 1987), 143–145.
12. R. de Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 50 (2000):1–18.

13. T. P. E. Auf der Heyde, Analyzing chemical data in more than two dimensions. A tutorial on factor and cluster analysis. *Journal of Chemical Education* 67 (1990):461–469.
14. R. A. Yafle, Common correlation and reliability analysis with SPSS for Windows (2003) ⟨http://www.nyu.edu/its/socsci/Docs/correlate.html⟩.
15. C. Weihs, Multivariate exploratory data-analysis and graphics—a tutorial. *Journal of Chemometrics* **7** (1993):305–340.
16. R. D. Cook and S. Weisberg, *An Introduction to Regression Graphics* (New York: Wiley, 1994) 10–12.
17. R. Tomassone, E. Lesquoy, and C. Millier, *La Regression, Nouveaux regards sur une ancienne méthode statistique* (Paris: Masson, 1983), 21–23.
18. G. I. Ouchi, *Personal Computers for Scientists, A Byte at a Time* (Washingon: American Chemical Society, 1987), 111.
19. A. Sayago, M. Boccio, and A. G. Asuero, Fitting straight lines with replicated observations by linear regression: the least squares postulates. *CRC Critical Reviews in Analytical Chemistry* 34 (2004):39–50.
20. A. G. Asuero and A. G. González, Some observations on fitting a straight line to data. *Microchemical Journal* 40 (1989):216–225.
21. R. de Levie, When, why, and how to use weighted least squares. *Journal of Chemical Education* 63 (1986):10–15.
22. O. Exner, How to get wrong results from good experimental data: a survey of incorrect applications of regression. *Journal of Physical Organic Chemistry* 10 (1997):797–813.
23. J. Riu and F. X. Rius, Univariate regression models with error in both axes. *Journal of Chemometrics* 9 (1995):343–362.
24. M. A. Castillo and R. C. Castells, Initial evaluation of quantitative performance of chromatographic methods using replicates at multiple concentrations. *Journal of Chromatography A*, 921 (2001):121–133.
25. K. Baumann, Regression and calibration for analytical separation techniques, Part II Validation, weighted and robust calibration. *Process Control Quality* 10 (1997):75–112.
26. S. N. Kelkar and J. J. Bzik, Calibration of analytical instruments. Impact of nonconstant variance in calibration data. *Analytical Chemistry* 72 (2000):4762–4765.
27. C. K. Bayne and I. B. Rubin, *Practical Experimental Designs and Optimization Methods for Chemistry* (Deerfield Beach, FL: VCR Publishers, 1986).
28. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection* (New York, Wiley, 1987), 42.
29. O. Exner and K. Zvára, Coefficient of determination in some atypical situations: use in chemical correlation analysis. *Journal of Physical Organic Chemistry* 12 (1999):151–156.
30. T. O. Kvalseth, Cautionary note about R^2. *The American Statistician* 39 (1985):279–285.
31. A. G. Asuero, G. González, F. de Pablos, and J. L. Gómez Ariza, Determination of the optimum working range in spectrophotometric procedures. *Talanta* 35 (1988):531–537.
32. J. R. Taylor, An introduction to error analysis. The study of uncertainties in physical measurements (Oxford: University Science Books, 1982), 185–186.
33. K. A. Brownlee, *Statistical Methodology in Science and Engineering,* 3th ed. (Malabar, Fl: Robert G. Krieger Publishing Co., 1984), 77–80.
34. W. Bremser and W. Hässelbarth, Shall we consider covariance? *Accreditation and Qualily Assurance* 3 (1998):106–110.
35. C. Salter, Error using the variance-covariance matrix. *Journal of Chemical Education* 57 (2000):1239–1243.
36. G. R. Bruce and P. S. Gill, P.S., Estimates of precision in a standard addition analysis. *Journal of Chemical Education* 76 (1999):805–807.
37. E. F. Meyer, A note on covariance in propagation of uncertainty. *Journal of Chemical Education* 74 (1997):1339–1340.
38. R. de Levie, *Advanced Excel for Scientific Data Analysis* (Oxford: Oxford University Press, 2004), 91–92.
39. M. Bookbinder and K. J. Panosian, Using the coefficient of correlation in method-comparison studies. *Clinical Chemistry* 33 (1987):1170–1176.
40. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis,* 3rd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1992), 11–12.
41. D. E. Sands, Correlation and covariance. *Journal of Chemical Education* 54 (1977):90–94.
42. S. C. Gad, *Statistics and Experimental Design for Toxicologists* (Boca Raton, FL: CRC Press, 1999), 49–51, 119.
43. G. D. Garson, Correlation (1999) ⟨http://www2.chass.ncsu.edu/garson/pa765/correl.htm⟩.
44. R. R. Wilcox, *Fundamentals of Modern Statistical Methods* (New York: Springer-Verlag 2001).
45. W. G. Hopkins, A new view of statistics, correlation coefficient (2004). ⟨http://www.sportsci.org/resource/stats/correl.html⟩.
46. M. F. Zady, Correlation and simple least squares regression (October 2000). ⟨http://www.westgard.com/lesson44.htm⟩.
47. J. Lawson and J. Erjavec, *Modern Statistics for Engineering and Quality Improvement* (New York: Brooks Cole, 2000).
48. J. Phillippe, *Les Méthodes Statistiques en Pharmacie et en Chimie* (Paris: Masson, 1967).
49. G. E. Dallal, Correlation coefficient (2003) ⟨http://www.tufts.edu/∼gdallal/corr.htm⟩.
50. N. A. Armstrong and K. C. James, *Pharmaceutical Experimental Design and Interpretation* (London: Taylor and Francis, 1996), 59.
51. M. Stoeppler, P. Valenta, and H. W. Nürnberg, Application of independent methods and standard materials–effective approach to relible trace and ultratrace analysis of metals and metalloids in environment and biological studies. *Fresenius Zeitschrif für analytische Chemie. Anal. Chem*. 297 (1979):22–34.
52. M. G. Cox and P. M. Harris, Design and use of reference data sets for testing scientific software. *Analytical Chimica Acta* 380 (1999):339–351.
53. P. M. Wanek, R. E. Whippe, T. E. Fickies, and P. M. Grant, Inaccuracies in the calculation of standard-deviation with electronic calculators. *Analytical Chemistry* 54 (1982):1877–1878.
54. B. D. McCullough and B. Wilson, On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistical Data Analysis* 31 (1999):27–37.
55. J. N. Miller, Basic statistical methods for analytical chemistry. Part 2. Calibration and regression methods. A review. *Analyst* 116 (1991):3–14.
56. P. C. Meier and R. E. Zünd, *Statistical Methods in Analytical Chemistry*, 2nd ed. (New York: Wiley, 2000), 92–94.
57. N. L. Johnson and F. C. Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences,* Vol. I, 2nd ed. (New York: Wiley, 1977), 455–456.

58. G. G. Koch, A basic demonstration of the [−1, 1] range for the correlation coefficient. *The American Statistician* 39 (1985):201–202.
59. T. O. Kvalseth, Cautionary note about R^2. *The American Statistician* 30 (1985):279–285.
60. J. Neter, W. Wasserman, and M. H. Kutner, *Applied Linear Models*, 2nd ed. (Homewood, II: Richard D. Irving, 1989), 102.
61. L. L. Havilcek and R. D. Crain, *Practical Statistics for the Physical Sciences* (Washington DC: ACS, 1988), 91–92.
62. M. Meloun, J. Militki, and F. Forina, *Chemometrics for Analytical Chemistry*. *Volume 2. PC-Aaided Regression and Related Methods* (New York: Ellis Horwood, 1994), 197–198.
63. J. D. Lee and T. D. Lee, *Statistics and Computer Methods in BASIC* (New York: Van Nostrand Reinhold, 1982) 81–83.
64. J. Mandel, *The Statistical Analysis of Experimental Data* (New York: Wiley-Interscience, 1964).
65. A. H. Reed, Misleading correlations in clinical application. *Clinica Chimica Acta* 40 (1972):266–268.
66. J. O. Westgard and M. R. Hunt, Use and interpretation of common statistical tests in method comparison studies. *Clinical Chemistry* 19 (1973):49–57.
67. J. L. Auget and J. Maccario, Mesusages du coefficient de correlation lineaire dans la comparaison de deux methods. *European Reviews of Biomedical Technology* 3 (1981):187–192.
68. S. Bolton, *Pharmaceutical Statistics, Practical and Clinical Applications*, 3rd ed. (New York: Marcel Dekker, 1994), 252.
69. R. A. Fisher, Lung cancer and cigarettes. *Nature* 182 (1958):108–109.
70. R. A. Fisher, Cancer and smooking. *Nature* 182 (1958):596–597.
71. Analytical Methods Committee, Is my calibration linear? *Analyst* 119 (1994):2363–2366.
72. R. Boque, F. X. Rius, and D. L. Massart, Straight line calibration—something more than slopes, intercepts and correlation coefficients. *Journal of Chemical Education* 70 (1993):230–232.
73. J. V. Loco, M. Elskens, C. Croux, and H. Beernaert, Linearity and calibration curves: use and misuse of the correlation coefficient. *Accreditation and Quality Assurance* 7 (2002):281–285.
74. J. R. J. Belloto and T. D. Sokolovski, Residual analysis in regression. *American Journal of Pharmaceutical Education* 49 (1985):295–303.
75. M. Thompson, Regression methods in the comparison of accuracy. *Analyst* 107 (1982):1169–1180.
76. M. Mulholland and P. B. Hibert, Linearity and the limitations of least squares calibration. *Journal of Chromatography A* 762 (1997):73–82.
77. D. L. MacTaggart and S.O. Farwell, Analytical uses of linear regression. Part I. Regression procedures for calibration and quantitation. *Journal of the Association of Official Analytical Chemists* 75 (1992):594–607.
78. J. H. Noggle, *Practical Curve Fitting and Data Analysis* (Englewood Cliffs, NJ: Prentice-Hall, 1993), 34–36.
79. *Matheson Gas Data Book* (East Rutherford, NJ: The Matheson Co., 1961).
80. L. A. Currie, Approach to accuracy in analytical chemistry, in *Treatise on Analytical Chemistry,* Part 1, Vol. 1, 2nd ed., eds I. M. Kolthoff, P. J. Elving (New York: Wiley, 1978) 95–242.
81. J. M. Green, A practical guide to analytical method validation. *Analytical Chemistry* 66 (1996): 305A–309A.
82. D. C. Crocker, *How to Use Regression Analysis in Quality Control* (Milwaukee: ASQC, 1990), 21.
83. J. B. Kennedy and A. M. Neville, *Basic Statistical Methods for Engineers and Scientists*, 2nd ed. (New York: Harper, 1976), 301.
84. R. F. Bartlett, Linear modelling of Pearson's product moment correlation coefficient: An application of Fisher's z-transformation. *The Statistician* 42 (1993):43–53.
85. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis* (New York: Wiley Classic Library Edition, 2000), 276–277.
86. C. F. Bond, and K. Richardson, Seeing the Fischer z-transformation. *Psychometrika* 69 (2004):291–304.
87. B. I. Harley, Some properties of an angular transformation for the correlation coefficient. *Biometrika* 43 (1956):219–224.
88. W. C. Günther, Desk calculation of probabilities for the distribution of the sample correlation coefficient. *The American Statistician* 31 (1977):45–48.
89. Crystal Ball 2002. Standard edition, Decisioneering, Inc. ⟨http://www.crystallball.com⟩.
90. A. R. Henderson, Chemistry with confidence: should *clinical chemists* require confidence intervals for analytical and other data? *Clinical Chemistry* 39 (1993):929–935.
91. J. Workman and H. Mark, Comparison of goodness of fit statistics for linear regression, Part I. *Spectroscopy* 19(9) (2004):38–41.
92. J.-X. Lü, Q. Shen, J.-H. Jiang, G.-L. Shen, and R.-Q. Yu, QASR analysis of cyclooxygenase inhibitor using particle swarm optimisation and multiple linear regression.*Journal of Pharmaceutical and Biomedical Analysis* 35 (2004):679–687.
93. J. N. Miller, II; Is it a straight line? *Spectroscopy International* 3(4) (1991):41–43.
94. J. A. Cornell, *How to Apply Response Surface Methodology* (Milwaukee: American Society for Quality Control, ASQC, 1990), 41.
95. L. Lyons, *Statistics for Nuclear and Particle Physicist* (Cambridge: Cambridge University Press, 1986), 61.
96. C. Chatfield and A. J. Collins, *Introduction to Multivariate Analysis* (London: Chapman and Hall, 1980), 26.
97. D. L. Massart and L. Kaufman, *The Interpretation of Analytical Chemistry Data by the Use of Cluster Analysis* (New York: Wiley, 1983), 15–17.
98. J. M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey, *Graphical Methods for Data Analysis* (Belmont, CA: Wadsworth International Group, 1983), 77–79.
99. J. L. Anscombe, Graphs in statistical analysis. *The Statistician* 27 (1973):17–21.
100. D. R. Brillinger, Does anyone know when the correlation coefficient is useful? A study of the times of extreme river flows. *Technometrics* 43 (2001):266–273.
101. R. J. Rummel, Understanding Correlation (Honolulu: University of Hawaii), Chapter 1, p 1, 1976. ⟨http://www.mega.nu:8080/ampp/rummel/uc.htm⟩.
102. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed. (Cambridge: Cambridge University Press, 1999), 636–639.
103. Statistica, *Volume I: General Conventions & Statistics I*, 2nd ed. (Tulsa, OK: StatSoft, 1998), 1450–1453.

104. M. G. Natrella, *Experimental Statistics, NBS Handbook* 91 (Washington DC: NBS, U.S. Government Printing Office, 1963), 5–6.
105. J. S. Hunter, Calibration and the straight line: Current statistical practices. *Journal of Association of Official Analytical Chemists* 64 (1981):574–583.
106. R. de Levie, Two linear correlation coefficients. *Journal of Chemical Education* 80 (2003):1030–1032.
107. R. Plesch, Der Korrelations Koeffizient—Prügröber der Analytik? *GIT Fachz. Lab.* 26 (1982):1040–1044.
108. M. D. VanArendonk, R. K. Skogerboe, and C. L. Grant, Correlation coefficients for evaluation of analytical calibration curves. *Analytical Chemistry* 53 (1981):2349–2350.
109. P. F. Tiley, The misuse of correlation coefficients. *Chemistry in Britain* 21 (1985):162–163.
110. C. K. Hancock, Some misconceptions of regression analysis in physical organic chemistry. *Journal of Chemical Education* 42 (1965):608–609.
111. W. H. Davies and W. A. Pryor, Measures of goodness of fit in linear free energy relationships. *Journal of Chemical Education* 53 (1976):285–287.
112. Analytical Methods Committee, Uses (proper and improper) of correlation coefficients. *Analyst* 113 (1988):1469—1471.
113. W. Huber, On the use of the correlation coefficient r for testing the linearity of calibration functions. *Accreditation and Quality Assurance* 9 (2004):726–727.