Estimating Correlations from Single-Sample Distributions of Measurements

Eugene K. Harris

The correlation coefficient between two variables (e.g., blood constituents) in normal individuals (or in patients) is usually estimated from measurements of both variables in single samples from the individuals studied. This coefficient is supposed to represent a correlation between the two variables within the average individual. However, unless all the subjects have identical mean values, the "single-sample" coefficient also reflects the correlation of mean values across individuals, generally different from the average intra-individual correlation. Even when both intra- and interindividual correlations are equal, the "single-sample" coefficient underestimates the true intra-individual correlation unless corrected for measurement errors. Using data from normal individuals, two examples are presented: in one case (calcium–total protein), the corrected single-sample coefficient appears to be an unbiased estimate of the desired correlation, whereas in the other example (albumin–globulins), it is quite misleading. These examples support the need for multiple, independent samples from each individual to assure a valid estimate of the average intra-individual correlation between two variables.

Additional Keyphrases: calcium–total protein and albumin–globulin correlations as examples • intra- and interindividual variation • discriminant function analysis • biological and analytical variance

Correlations between clinical variables, in patients or normal volunteers, are almost always estimated from "single-sample" distributions of observations—that is, one pair of measurements of the variables $x_1$ and $x_2$ in each of a large number of persons. The statistical characteristics of a single-sample distribution of one variable have been shown (1) to depend on the distributions of intra-individual means and variances across the population of individuals sampled. The characteristics of bivariate and multivariate single-sample distributions, in particular the correlations between variables, are also affected by intra- and interindividual variations. This report will attempt to describe some of these effects and their implications for the study of such correlations.

The variance of a single-sample distribution of one variable is the sum of the average intra-individual variance and the variance among individual means. Similarly, the covariance between two variables, estimated from a single pair of measurements on each individual, can be shown to be the sum of two terms: the average intra-individual covariance and the covariance between individual mean values across the population. Therefore, the "expected" (or true mean) value of a correlation coefficient from paired single-sample measurements may be written,

$$
\rho_{12} = E \text{Cov}(x_1, x_2) + \text{Cov}(\mu_{1i}, \mu_{2i})/[(E \sigma_{1}^{2} + \text{Var} \mu_{1i}) (E \sigma_{2}^{2} + \text{Var} \mu_{2i})]^{1/2} 
$$ (1)

where the subscript $i$ refers to an individual; $E$ signifies expected value; $\sigma_{1}^{2}$ and $\sigma_{2}^{2}$ represent the variances of $x_1$ and $x_2$ within the $i$th individual; $\mu_{1i}$ and $\mu_{2i}$ denote the means of these variables in the $i$th individual, and Cov and Var are abbreviations of covariance and variance.1

Thus, a "single-sample" correlation coefficient contains both inter- and intra-individual effects. It approximates a weighted average of the correlation among individual mean values [say, $\rho(\mu_{1i}, \mu_{2i})$] and the average intra-individual correlation coefficient $E \rho_{12}$.2 The weight applied to the first of these correlations is $(\text{Var} \mu_{1i} \text{Var} \mu_{2i})^{1/2}$, and to the second, $(E \sigma_{1}^{2} E \sigma_{2}^{2})^{1/2}$. In practice, however, the single-sample correlation coefficient is probably never interpreted as a weighted average of "within" and "between" correlations. It is more likely to be viewed (implicitly, at least) as an estimate of $E$$

---

1 Equation 1 is only approximate because a sample correlation coefficient is a ratio of sample covariance to sample variances, and the expected value of an observed ratio is not exactly equal to the ratio of the expected values of numerator and denominator. In practical applications to clinical studies, the approximation should be within ±0.1. This has been confirmed in data available to me.

2 I assume, of course, that real differences exist among individuals with respect to the variables under study, so that $\text{Var} \mu_{1i}$ and $\text{Var} \mu_{2i}$ are greater than zero. Equation 1 will be an exact weighted average of the two correlations only if $E \sigma_{1}^{2}/\text{Var} \mu_{1i} = E \sigma_{2}^{2}/\text{Var} \mu_{2i}$, that is, if the ratio of intra- to interindividual variance is the same for both variables. Otherwise, the single-sample coefficient will be less than the weighted average. However, even if one variance ratio is up to 4 times larger than the other, the resulting downward bias will not exceed 5 percent. Past experience with common blood tests (2, II, Table 5) indicates that only rarely will the variance ratios of two tests differ by this high a factor. Nevertheless, this mathematical bias adds to the difficulties of interpreting single-sample correlation coefficients.
\( p(x_1, x_2) \) alone—that is, as a measure of the relationship between two variables within the average individual under study. Indeed, it would seem that no useful clinical purpose is served by computing the correlation coefficient unless such interpretation is valid.

Unfortunately, the quantity \( E \rho(X_1, X_2) \) is only part of the expression actually being estimated, nor is there any a priori reason why this correlation and the interindividual correlation \( \rho(\mu_1, \mu_2) \) should be equal, or even close in value. Figures 1a–c illustrate three possible situations involving a sample of four individuals, each of whom has provided a series of paired values of \( x_1 \) and \( x_2 \) lying within the dashed curves. The large dots represent the true means \( \mu_1 \), \( \mu_2 \). In Figure 1a, the intra-individual correlations vary widely and their average would be close to zero, but the correlation among individual mean values is strongly positive. In Figure 1b, both the average intra-individual correlation and the interindividua correlation are positive and probably close in value. In Figure 1c, intra-individual correlations are all strongly negative, while the interindividual correlation is poor but in the same direction. This figure could equally well have been drawn with the correlations in opposite direction.

Even when the two “true” correlations are equal, a single-sample correlation coefficient will be biased because of the attenuating effect of analytic “noise” on the estimate of \( E \rho(x_1, x_2) \) contained implicitly in the single-sample coefficient. (This is in addition to the downward bias introduced by the mathematical factor mentioned in Footnote 2.) The extent of this attenuation because of “noise” depends on the ratio of analytic variance to total intra-individual variance: analytic plus biological (see examples below). The estimated value of \( \rho(\mu_1, \mu_2) \), also contained within the single-sample coefficient, is unaffected by analytic variance, assuming that the latter is independent of the true value of the variable being measured. This may be seen more clearly when a series of paired measurements of \( x_1 \) and \( x_2 \) in each individual are analyzed, as in the examples below. Then the interindividual correlation \( \rho(\mu_1, \mu_2) \) may be separately estimated by a ratio of covariance and variance components from which analytic variance has already been deducted. Thus, analytic variance reduces the expression of \( E \rho(x_1, x_2) \), but not of \( \rho(\mu_1, \mu_2) \), within the single-sample correlation coefficient. The single-sample coefficient, a hybrid of the two correlations, is not a reliable estimator of either one.

Now, if \( E \rho(x_1, x_2) \) is in fact equal to \( \rho(\mu_1, \mu_2) \), then it can be shown by a simple algebraic argument based on Equation 1 that a correction for attenuation will make the single-sample correlation coefficient a closer estimator of \( E \rho(x_1, x_2) \). However, the hypothesis that the two correlations are equal cannot be tested when only one sample is obtained from each subject.

### Examples and Discussion

The following examples, based on a study of 68 normal individuals (2), may help to clarify these ideas. Blood samples were collected weekly for 8–12 weeks and sera analyzed in duplicate immediately after collection. The data used here are weekly means of duplicates. From such multi-sample data, separate estimates of interindividual and average intra-individual correlation coefficients may be obtained. To estimate the former quantity, analyses of covariance and variance “among” and “within” individuals are performed with respect to the two selected variables. Components of covariance and variance among individuals are obtained in a standard manner by adjusting the covariance and variances of observed mean values for the average values of these quantities within individuals (2, II). In the earlier report, the estimated component of variance among individuals (estimating Var \( \mu_1 \)) was labeled \( s_{\Omega 2}^2 \). Similarly, the estimated component of covariance among individuals may be denoted by \( s_{\Omega 2}^2 \). The correlation \( \rho(\mu_1, \mu_2) \) may then be estimated by the ratio \( r(\mu_1, \mu_2) = s_{\Omega 2}^2 / s_{\Omega 2}^2 \). An initial estimate of \( E \rho(x_1, x_2) \) is best obtained by computing a separate correlation coefficient from the \( n \) pairs of values for each individual and averaging these individual coefficients. This estimate, say \( r(x_1, x_2) \), remains biased downward by analytic “noise.”

Table 1 presents the following correlations linking

<table>
<thead>
<tr>
<th></th>
<th>Interindividual ( r(\mu_1, \mu_2) )</th>
<th>Av Intra-individual ( r(x_1, x_2) )</th>
<th>Single sample ( r_{12} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca–total protein</td>
<td>.63</td>
<td>.24</td>
<td>.42</td>
</tr>
<tr>
<td>Albumin–globulin</td>
<td>-.26</td>
<td>-.37</td>
<td>-.32</td>
</tr>
</tbody>
</table>

---

82 CLINICAL CHEMISTRY, Vol. 19, No. 1, 1973
calcium with total protein and albumin with globulins: (a) the estimated correlation among true mean values $\rho_1$, (b) the average observed intra-individual correlation coefficient $\rho(x_1, x_2)$, and (c) an estimate ($r_{12}$) of the expected single-sample correlation, computed from Equation 1. Globulins were calculated as total protein less albumin (and should properly be called nonalbumin protein). In computing these statistics, each individual's correlation coefficient, variance, or covariance, was weighted by the number of samples collected. In Figure 2, the 68 observed mean values of calcium and total protein are plotted against each other. Figure 3 shows the same graph for albumin and globulins.

These two examples show quite different patterns of correlation, beyond the expected difference in sign. The calcium–total protein correlations, given the attenuating influence of analytic variance on the estimate $\gamma(x_1, x_2)$, appear consistent with the general pattern illustrated in Figure 1b. The question then arises whether in this case $E \rho(x_1, x_2) = \rho_1$, with analytic variance the sole, or at least the only significant, factor accounting for the difference between $\gamma(x_1, x_2)$ and $\rho_1$. If this hypothesis is correct, then one should be able to estimate $E \rho(x_1, x_2)$ from the single-sample correlation by adjusting the latter upwards through division by an "attenuation factor" less than unity. Such a factor has appeared in the statistical literature in one form or another for at least 50 years [see (4) for a clinical application]. It may be expressed here as

\[
\text{"a.f."} = [(1 - s_{A1}^2/s_{T1}^2)(1 - s_{A2}^2/s_{T2}^2)]^{1/2}
\]

where $s_{A1}^2$ denotes estimated analytic variance and $s_{T1}^2$ denotes the variance observed in single-sample measurements.

The variance $s_{T1}^2$ includes both intra- and interindividual variances, i.e., $E s_{T1}^2 = E s_{A1}^2 + \text{Var} \mu_1$. With multisample data, these quantities are estimated by analysis of variance within and among individuals, as described in (2, II). Following the notation used there, $s_{T1}^2 = s_{P1}^2 + s_{G1}^2$, where $s_{P1}^2$ is the average observed intra-individual variance, including both analytic and intrapersonal variation, while $s_{G1}^2$, defined above, is the variance component estimate of Var $\mu_1$. Analytic variance represents a combination of day-to-day variance ($s_{f1}^2$) and variance between duplicates ($s_{D1}^2$). The values of these variances are given in (2, I and II). Then $s_{A1}^2 = s_{f1}^2 + s_{G1}^2/2$, since the data used here are the means of duplicates. All of these statistics needed for the present examples are collected in Table 2.

In the calcium–total protein example, if the single-sample correlation coefficient $r_{12} = .42$ is divided by its attenuation factor, $[(1 - .0061/.013)(1 - .034/ .245)]^{1/2} = .68$, an adjusted correlation of .62 is obtained, almost identical to the observed (unattenuated) value of .63 for $\rho(x_1, x_2)$, which by hypothesis is also an estimate of the expected intra-individual correlation $E \rho_1(x_1, x_2)$. Moreover, when the average observed intra-individual correlation $\gamma(x_1, x_2) = .24$ is divided by a corresponding intra-individual attenuation factor, $[(1 - s_{A1}^2/s_{P1}^2)(1 - s_{A2}^2/s_{P2}^2)]^{1/2} = .37$, we obtain the value .65, which is the most direct estimate of $E \rho_1(x_1, x_2)$ provided by these data. The agreement among the three estimates (.62, .63, .65) is remarkable considering the fact that all depend on ratios of sample variances and covariances.4

Thus, the relationship between calcium and total protein appears to be a true example of the special case, $E \rho_1(x_1, x_2) = \rho_1$. The correlation coefficient computed from single samples per individual

---

4It is well known that total calcium and total protein in serum are positively correlated. A negative correlation between serum albumin and globulins has been reported by Fryce (3).
and corrected for attenuation because of analytic variation becomes in this case a valid estimate of the expected intra-individual correlation.

The albumin–globulin relationship is much less favorable. Assuming that a negative intra-individual correlation does exist, \(^6\) the overall pattern resembles that in Figure 1c. Unknown to the investigator with single-sample data, his observed correlation coefficient is depressed in numerical value by a relatively low interindividual correlation \(\rho(\mu_2, \mu_2)\). The correction for attenuation that can be applied (given only one sample per individual studied) is too weak to improve the observed correlation so that it may become a valid estimate of intra-individual correlation. It must remain an underestimate. This is because, as noted before, the quantity \(s_T^2\) in the single-sample attenuation factor includes both inter- and intra-individual variance. The proper factor would involve intra-individual variance alone, but this cannot be determined from single-sample statistics.

In the albumin–globulin example, the attenuation factor for the single-sample correlation is \([\{1 - .024/1.121\} - .030/2.30]^{1/2} = .83\). Dividing the uncorrected value of \(r_{12}(-.32)\) by this factor yields an adjusted single-sample correlation of -.39. This result, however, appears to be much less in numerical value than the true intra-individual correlation coefficient. The raw estimate of the latter quantity, \(\gamma(x_1, x_2) = -.37\), after adjustment by the intra-individual attenuation factor, \([\{1 - .024/.051\} - .030/0.093]^{1/2} = .60\), becomes -.62, about 60% larger numerically than the corrected single-sample coefficient.

This example confirms that when the true inter-individual correlation \(\rho(\mu_1, \mu_2)\) is weaker than the average intra-individual correlation, the correlation coefficient estimated from single-sample data will underestimate the latter even when the single-sample estimate is “corrected” for attenuation due to analytic variance. The opposite situation is shown in Figure 1a. Here, the single-sample correlation coefficient will be pulled upward by the influence of a relatively strong interindividual correlation. When the true average intra-individual correlation is close to zero, any upward adjustment of the observed correlation by an attenuation factor will only serve to increase the distance between the estimate and the quantity we hope it is estimating.

**Final Remarks**

A correlation coefficient calculated from single measurements of \(x_1\) and \(x_2\) in each of a number of individuals has been shown to be an invalid estimate of the true average intra-individual correlation, even if corrected for attenuation because of analytic “noise,” except under a special hypothesis that cannot be tested when only single-sample data are available. Using data from a study of multiple samples per individual, we have presented evidence that in one case—the correlation of calcium with total protein—this hypothesis probably holds, while in another case—albumin vs. globulins—it almost certainly does not.

In general, then, the single-sample correlation coefficient cannot be considered a reliable quantitative measure of the strength of a relationship between two variables within an individual. Perhaps, however, it has never been relied upon as a quantitative measure but only as a qualitative guideline. That is, perhaps the rule of thumb has been: if \(r_{12}\) is less than, say \([0.8]\) (certainly, if less than \([0.6]\)), the relationship between the variables is too tenuous to be worth further consideration. This is a good rule to follow if one wants to be sure to reject weak correlations. Since neither the intra- nor the interindividual correlation can exceed unity, an observed single-sample correlation of \([.8]\) undoubtedly reflects a high degree of both kinds of correlation, and, in particular, implies the situation shown in Figure 1b, which favors the single-sample correlation coefficient, corrected for attenuation.

However, in a case like the albumin–globulin comparison, a rule that recommends ignoring all but the highest observed correlation coefficients will fail to detect a strong intra-individual correlation that has been attenuated by analytic variance and depressed by the influence of a weak correlation among individual mean values. In general, such a rule seems a poor substitute for estimating average intra-individual correlation directly by computing a correlation.

---

\(^6\) That the average observed correlation between albumin and globulin concentrations in these data must be negative is apparent from the statistics in Table 2, where the average variance \((s_T^2)\) of total protein is less than the sum of the \(s_1^2\) values for albumin and globulins. If the albumin-globulin correlation were positive or zero, the value of \(s_T^2\) for total protein would equal or exceed the sum of the constituent variances. Those individuals with correlations between the average value (-.37) and +1 showed about 17 percent greater standard deviation in total protein, on the average, than did individuals with correlations between -.37 and -1. Perhaps a homeostatic constraint on variability in total serum protein concentration within many normal individuals underlies the average negative correlation observed between albumin and globulins. If so, then an estimate of average intra-individual correlation between these chief constituents of serum protein, free of any effect of interindividual correlation of mean values, would certainly be desirable.

---

**Table 2. Variances Needed for Calculating Attenuation Factors**

<table>
<thead>
<tr>
<th></th>
<th>Calcium (mmol/liter)</th>
<th>Total protein</th>
<th>Albumin (g/100ml)</th>
<th>Globulins</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_P^2)</td>
<td>.0078</td>
<td>.089</td>
<td>.051</td>
<td>.093</td>
</tr>
<tr>
<td>(s_C^2)</td>
<td>.0050</td>
<td>.156</td>
<td>.070</td>
<td>.137</td>
</tr>
<tr>
<td>(s_T^2)</td>
<td>.0128</td>
<td>.245</td>
<td>.121</td>
<td>.230</td>
</tr>
<tr>
<td>(s_1^2)</td>
<td>.0061</td>
<td>.034</td>
<td>.024</td>
<td>.030a</td>
</tr>
</tbody>
</table>

\(^a\) The estimated analytic variance of globulin concentrations is given by, \(s_2^2(\text{Alb}) + s_2^2(\text{T. Prot.}) = 2 s_2(\text{Alb})(\text{T. Prot.})\), where \(r\) denotes the correlation between analytic errors in albumin and total protein determinations. Since these determinations were carried out together on the Auto-Analyzer, some correlation may be expected. From 884 concurrent serum pool analyses (2), this correlation was estimated as \(r = .52\). Therefore, the analytic variance of globulins was estimated as \(.024 + .034 - 2(0.52)(0.024) = .030\).
coefficient between the variables of interest for each individual in the study. Of course, this requires a series of independent measurements of each individual under controlled conditions. Moreover, because such correlations will need correction for attenuation, the analytic variance in each variable should be kept small relative to the average total intra-individual variance (i.e., the average sum of analytic and intrapersonal variances) since the ratio of these two quantities will define the attenuation factor. This problem of analytic variance is naturally most acute in variables under close homeostatic control. And so we return again, by a new route, to the salient arguments of the earlier series of reports from which these data were selected, namely, the merit of distinguishing between intra- and interindividual variation in clinical studies, the need for multiple samples per individual in order to obtain separate estimates of these two sources of variation, and the importance of minimizing analytic variance to ensure the reliability of such estimates.

One last note: The defects of single-sample data in measuring physiological relationships do not detract from their possible usefulness in discriminating among different diseases in a patient population. Discriminant function analysis assumes that the individuals under study come from specific, previously defined categories. It then tries to separate these categories statistically by analysis of the means of variables measured on a subset of individuals from each category. The purpose is to derive a reliable mathematical rule for deciding in which category to place a new patient. Multiple samples per individual are not required to obtain such a rule, since the fact that \( \text{Var} \mu_1 \) is greater than zero for any variable within a category (i.e., that the individuals in a category are not perfectly homogeneous) can be accepted so long as this variance is small relative to the variance between categories. If this is not the case, the discriminant function will simply not work reliably in practice and will soon be abandoned. Discriminant function analysis thus serves a more pragmatic purpose than does the estimation of intra-individual correlation. It is more concerned with defining the boundaries separating groups of individuals than with measuring how much two variables affect each other within one individual.

References