PSYCHOMETRIKA—VOL. 13, NO. 4
DECEMBER, 1948

# THE ROLE OF CORRELATION IN ANALYSIS OF VARIANCE

CLYDE H. COOMBS*
UNIVERSITY OF MICHIGAN

A test of the significance of a row or column agent in an analysis of variance may be expressed in the form of correlation between the agent and the variate. A test of the significance of interaction variance may be expressed in the form of correlation between the agents. These expressions are principally of theoretical interest in that the degree of significance in an $F$ test or the value of a correlation coefficient may be controlled at will, or inadvertently, within certain limits.

In a recent article Peters (1) discussed the interpretation of interaction variance as correlation. There are a number of further interrelations to be pointed out and some interesting interpretations to be made. Peters' article just begins the study of the role of correlation in analysis of variance and does not emphasize some of the characteristics of both correlation and analysis of variance which are pointedly revealed by their joint study. The significance of this study lies not in the practical application of some of the formulas developed but rather in the design of research involving analysis of variance or correlational analysis.

## *I. Effectiveness of an Agent and Intraclass Correlation*

Consider the simple case of an analysis of variance into three components. The levels of one agent are designated $1, 2, \cdots, i, \cdots, n$ and those of the other agent $1, 2, \cdots, j, \cdots, m$. With a single observation (or mean of observations) at each joint level $ij$, the observations may be represented by a table of $n$ rows and $m$ columns.

We shall proceed first to get an expression for the product-moment correlation between pairs of columns (or rows) [equation (10) or (16) and (17).] Let the mean of all observations be designated $M$ and the score in any cell be designated $X_{ij}$.

$$M = \frac{\sum_i \sum_j X_{ij}}{n\,m}. \tag{1}$$

*This study was made possible by the Bureau of Psychological Services, Institute for Human Adjustment, Horace H. Rackham School of Graduate Studies, University of Michigan.

Designating the mean of any row as $M_i$, then

$$M_i = \frac{\sum\limits_j X_{ij}}{m}. \tag{2}$$

The variance of a distribution of a sum of scores is

$$\begin{aligned}\sigma^2_{(X_{i1}+X_{i2}+\ldots+X_{im})} = \sigma^2_{X_{i1}} + \sigma^2_{X_{i2}} + \cdots + \sigma^2_{X_{im}} + 2r_{X_{i1}X_{i2}}\sigma_{X_{i1}}\sigma_{X_{i2}} \\ + \cdots + 2r_{X_{i,m-1}X_{im}}\sigma_{X_{i,m-1}}\sigma_{X_{im}}\ .\end{aligned} \tag{3}$$

Assuming homogeneity of variance from column to column and that the correlations between pairs of columns are equal, then

$$\sigma^2_{(X_{i1}+X_{i2}+\ldots+X_{im})} = m\,\sigma_j{}^2[1 + (m-1)r_j]\,, \tag{4}$$

where $\sigma_j{}^2$ represents the variance of any column and $r_j$ the product-moment correlation between any pair of columns.

The variance of the means of rows may be expressed as:

$$\sigma^2_{M_i} = \frac{1}{m^2}\sigma^2_{(X_{i1}+X_{i2}+\ldots+X_{im})} = \frac{\sigma_j{}^2}{m}[1 + (m-1)r_j]. \tag{5}$$

But

$$\sigma^2_{M_i} = \frac{\sum\limits_i (M_i - M)^2}{n} \tag{6}$$

and

$$\sigma_j{}^2 = \frac{\sum\limits_i (X_{ij} - M_j)^2}{n}. \tag{7}$$

But assuming homoscedasticity or homogeneity of variance, a better estimate of $\sigma_j{}^2$ may be obtained by combining the within columns sums of squares and dividing by the total number of cases. Hence, (7) becomes

$$\sigma_j{}^2 = \frac{\sum\limits_j \sum\limits_i (X_{ij} - M_j)^2}{n\,m}. \tag{8}$$

Substituting (6) and (8) in (5),

$$m^2 \sum_i (M_i - M)^2 = \sum_j \sum_i (X_{ij} - M_j)^2[1 + (m-1)r_j]. \tag{9}$$

Solving for $r_j$ ,

$$r_j = \frac{m^2 \sum_i (M_i - M)^2 - \sum_j \sum_i (X_{ij} - M_j)^2}{(m-1) \sum_j \sum_i (X_{ij} - M_j)^2}. \tag{10}$$

A single score in this design may be described alternatively as:

$$X_{ij} - M = (X_{ij} - M_j) + (M_j - M) \tag{11}$$

or

$$X_{ij} - M = (M_i - M) + (M_j - M) + d_{ij}. \tag{12}$$

Squaring and summing equation (11) over the rows and columns:

$$\sum_i \sum_j (X_{ij} - M)^2 = \sum_i \sum_j (X_{ij} - M_j)^2 + n \sum_j (M_j - M)^2. \tag{13}$$

Similarly for equation (12):

$$\begin{aligned} \sum_i \sum_j (X_{ij} - M)^2 = m \sum_i (M_i - M)^2 \\ + n \sum_j (M_j - M)^2 + \sum_i \sum_j d_{ij}^2. \end{aligned} \tag{14}$$

Subtracting equation (13) from (14) and rearranging terms:

$$\sum_i \sum_j (X_{ij} - M_j)^2 = m \sum_i (M_i - M)^2 + \sum_i \sum_j d_{ij}^2. \tag{15}$$

Substituting (15) in (10) and simplifying:

$$r_j = \frac{m(m-1) \sum_i (M_i - M)^2 - \sum_i \sum_j d_{ij}^2}{m(m-1) \sum_i (M_i - M)^2 + (m-1) \sum_i \sum_j d_{ij}^2}. \tag{16}$$

In a similar manner the equation for the product-moment correlation between rows ($r_i$) may be obtained and is as follows:

$$r_i = \frac{n(n-1) \sum_j (M_j - M)^2 - \sum_i \sum_j d_{ij}^2}{n(n-1) \sum_j (M_j - M)^2 + (n-1) \sum_i \sum_j d_{ij}^2}. \tag{17}$$

Our next step will be to express $r_j$ and $r_i$ in a form which will indicate their identity with Fisher's intraclass correlation [equations (24) and (25)], and then point out the well-known relation of $F$ and intraclass correlation [equations (28) and (29).] Let us designate by $V$ the mean sum of squares in analysis of variance. Then the mean sum of squares for rows is:

$$V_i = \frac{m \sum_i (M_i - M)^2}{n - 1}. \tag{18}$$

Then

$$(n-1)(m-1)V_i = m(m-1)\sum_i (M_i - M)^2. \tag{19}$$

Similarly, the mean sum of squares for columns is:

$$V_j = \frac{n \sum_j (M_j - M)^2}{m - 1} \tag{20}$$

and

$$(m-1)(n-1)\ V_j = n(n-1)\sum_j (M_j - M)^2, \tag{21}$$

and designating the error or remainder sum of squares $V_e$, then

$$V_e = \frac{\sum_i \sum_j d_{ij}^2}{(n-1)(m-1)} \tag{22}$$

and

$$(n-1)(m-1)V_e = \sum_i \sum_j d_{ij}^2. \tag{23}$$

Substituting (19) and (23) in (16):

$$r_j = \frac{(n-1)(m-1)V_i - (n-1)(m-1)V_e}{(n-1)(m-1)V_i + (n-1)(m-1)^2 V_e},$$

which simplifies to

$$r_j = \frac{V_i - V_e}{V_i + (m-1)V_e}. \tag{24}$$

Similarly, substituting (21) and (23) in (17) and simplifying, we have:

$$r_i = \frac{V_j - V_e}{V_j + (n-1)V_e}. \tag{25}$$

These last two equations will be immediately recognized as the same as the equation for Fisher's unbiased estimate of the intraclass correlation.

Solving equations (24) and (25) for $V_e$, we have, respectively,

$$V_e = \frac{V_i(1 - r_j)}{1 + (m-1)r_j}, \tag{26}$$

$$V_e = \frac{V_j(1-r_i)}{1+(n-1)r_i}. \tag{27}$$

From (26), rearranging terms

$$\frac{V_i}{V_e} = \frac{1+(m-1)r_j}{1-r_j} = F, \tag{28}$$

the $F$ being Snedecor's $F$ for testing the significance of the agent which varies from row to row.

Correspondingly, the test for the significance of the column agent is obtained from (27):

$$\frac{V_j}{V_e} = \frac{1+(n-1)r_i}{1-r_i} = F. \tag{29}$$

Hence, if

$$F = \frac{V_i}{V_e} \quad \text{or} \quad \frac{V_j}{V_e}$$

is found to be non-significant, then $r_j$ or $r_i$, respectively, does not differ significantly from zero. If, on the other hand, the $F$ is found to be significant, then it will sometimes be interesting and meaningful to express the relation as an intraclass correlation coefficient and thereby have an indication of the degree of the relationship.

Negative intraclass correlation will reveal itself in that $V_i$ (or $V_j$) will be less than $V_e$. In a two-component analysis of variance with only one agent there is merely a $V_B$ and a $V_W$ representing, respectively, the estimate of universe variance with the agent varying and an estimate of universe variance with the agent fixed. If $V_B < V_W$, then the agent is generating negative intraclass correlation.

It should be further pointed out here that it is not the *interaction variance* which is being interpreted as correlation, but the effect of an agent on the variate is being described or measured by means of the correlation coefficient. As a matter of fact, in the three-component analysis discussed here, it is implicitly assumed that the interaction variance is an estimate of error variance only and, hence, that there is no correlation between the two agents. That interaction variance is a function of error and *correlation between agents* is the thesis of section IV of this paper.

An expression indicating the effect on residual variance of correlation between columns or correlation between rows may be readily obtained as follows:

Solving equation (16) for $m \sum_i (M_i - M)^2$, we have

$$m \sum_i (M_i - M)^2 = \frac{1 + (m-1) r_j}{(m-1)(1-r_j)} \sum_i \sum_j d^2 , \tag{30}$$

and similarly from equation (17)

$$n \sum_j (M_j - M)^2 = \frac{1 + (n-1) r_i}{(n-1)(1-r_i)} \sum_i \sum_j d^2 . \tag{31}$$

Substituting equation (30) and (31) in (14) and rearranging the terms, we have

$$\sum_i \sum_j d^2 = \frac{r_i r_j \sum_i \sum_j (X-M)^2}{(1-r_i) r_j + (1-r_j) r_i + (n-1)(1-r_i) r_i r_j + (m-1)(1-r_j) r_i r_j + r_i r_j} , \tag{32}$$

which, in the simplest case of two rows and two columns ($n = m = 2$), becomes

$$\sum_i \sum_j d^2 = \frac{(1-r_i)(1-r_j) \sum_i \sum_j (X-M)^2}{3 - r_i - r_j - r_i r_j} . \tag{33}$$

It is apparent that if $r_i = r_j = 0$, then one-third of the total sum of squares in a three-component analysis is residual and none of the $F$'s would be significant.

## II. *The* t*-test and Intraclass Correlation*

In those instances in which an agent is given only two values, the $F$ test corresponds to the conventional application of the *t-test* to the significance of differences between two means.

Solving equation (28) for $r_j$ in terms of $F$, we have:

$$r_j = \frac{F-1}{F+m-1} . \tag{34}$$

It has been shown (2) that

$$F = t^2 \tag{35}$$

in the case of two groups, and hence any $t$-test may be converted into an intraclass correlation.

Substituting (35) in (34),

$$r_j = \frac{t^2 - 1}{t^2 + m - 1}. \tag{36}$$

This formula, however, is not normally to be recommended unless the limitations of the correlation so obtained are very clearly understood and interpreted. These limitations are not peculiar to this correlation but pertain to those secured from formulas (24) and (25) or any other correlation including those obtained in the usual manner rather than via analysis of variance. Analysis of variance merely brings it home with more force. The value of the correlation between two variates is a function of the proportion of the variance of either associated with a common agent. The variance which an agent contributes to a variate is, of course, a direct function of its own variance. But the *proportion* of variance contributed by an agent is a function not only of the variance it contributes but a function of the variance being contributed by all other agents simultaneously. Hence, if the correlation in the universe is significantly non-zero, the actual value of the correlation secured on a sample can be manipulated within certain limits by controlling the variability of the agent in relation to the other agents in the universe.

This is of considerable significance to the use of correlation as a descriptive statistic and to the design of studies involving analysis of variance. In the opinion of the writer, in either of the above cases the degree of significance or relation of an agent to a variate is best tested or described when the agent in question and all other agents affecting the variate are permitted full normal variation.

### III. $\varepsilon^2$ *and Intraclass Correlation*

In Section I the relation between $F$ and intraclass correlation was investigated. The relation between $\varepsilon^2$ and $F$ is known and given by the equation (2):

$$\varepsilon^2 = \frac{(K-1)\,F - (K-1)}{(K-1)\,F + N - K}. \tag{37}$$

From these two relations we are able to express the functional relation of $\varepsilon^2$ and the intraclass correlation. As a result of agent $B$ we have an $r_i$ given by (29). Substituting for $F$ in (37), and rearranging terms:

$$\varepsilon^2 = \frac{n(m-1)r_i}{(m\,n-1) - (n-1)r_i}, \tag{38}$$

where the $K$ of (37) is the same as $m$ in our notation and the $N$ of (37) is given by $nm$ in our notation.

## IV. *Application and Interpretation*

A recent study by Siegel and Stuckey (3) will be used here to show the application and interpretation of some of these formulas. Very briefly, this study consisted of observing the amount of water drunk and the amount of food eaten after each of four successive six-hour intervals for each of sixteen rats. Analyses of variance were made of the water intake and of the food intake. The results are reproduced in Tables I and 2 below.

TABLE 1
Water Intake

| Source | Sum of Squares | df | Mean Square | $F$ | Intraclass Correlations |
|---|---|---|---|---|---|
| time | 968.00 | 3 | 322.67 | 76.3* | .825 |
| animals | 82.72 | 15 | 5.51 | 1.30 | .07 |
| residual | 190.34 | 45 | 4.23 | | |
| total | 1241.06 | 63 | | | |

*Significant at 1% level.

TABLE 2
Food Intake

| Source | Sum of Squares | df | Mean Square | $F$ | Intraclass Correlations |
|---|---|---|---|---|---|
| time | 490.51 | 3 | 163.50 | 60.1* | .787 |
| animals | 13.94 | 15 | 0.93 | 2.92* | —.197 |
| residual | 122.40 | 45 | 2.72 | | |
| total | 626.85 | 63 | | | |

*Significant at 1% level.

The last column of these tables contains the intraclass correlation obtained by use of equations (24) and (25). Let us consider the interpretation of the correlation .825 for water intake and the agent *time*. Here we have four classes or families, the observations on sixteen rats for each of four intervals of time. This high intraclass correlation signifies that the amount of water intake for the various members of a family tends to be much more similar than one would expect from chance. In other words, "time" is a significant agent in the amount of water intake—there are certain periods of the day or night when the animals are more inclined to drink than at other times.*

*It should be made clear that the interpretations being made here are literal interpretations of the data and of the results of the formulas applied. That these results might be explained in terms of controls present or absent in the conduct of this experiment is not the concern of the present writer.

Similarly, if we observe the high intraclass correlation of .787 for food intake with the agent *time* we may make a similar interpretation: that during certain intervals of the day and night the animals will be more or less likely to be eating.

Some interesting relations are also revealed by the intraclass *r's* for animals in these two studies. In the case of water intake, a non-significant correlation of +.07 is found for animals. Here a class or family consists of the four successive observations of a single animal's water intake. An intraclass $r$ of zero indicates that the magnitude of one of the observations in a class has no relation to the magnitude of the others.

On the other hand, we find a significant negative intraclass correlation of —.20 for the food intake of an animal. This indicates that the four successive observations of the amount of food eaten are more *different* than would be expected by chance.

### V. *Interaction and Correlation Between Agents*

Let us consider next the description of interaction in terms of correlation between the agents. Let us first consider a three-component analysis with two agents, A and B, with agent A partitioned over $n$ rows and agent $B$ partitioned over $m$ columns with no replication.

A single score may then be represented as comprising the general mean plus a contribution from each of the agents and an error increment ($\varepsilon$), thus:

$$X = M + a + b + \varepsilon \,. \tag{39}$$

Hence,

$$x = X - M = a + b + \varepsilon \,. \tag{40}$$

Squaring, summing and dividing by the number of observations, we have

$$\sigma_x^2 = \sigma_a^2 + \sigma_b^2 + \sigma_\varepsilon^2 + 2\, r_{ab}\, \sigma_a\, \sigma_b \,, \tag{41}$$

assuming the errors to be uncorrelated with the agents A and B and letting $r_{ab}$ be the correlation between agents.

In a four-component analysis there is replication of the design with, say, $p$ observations in each group. Equation (41) is, then, the variance of the means of the *mn* groups. Multiplying (41) through by *mnp* we get the total sum of squares between groups, which may be written, in row and column notation, as

$$\begin{aligned} p \sum_j \sum_i (M_{ij} - M)^2 = m\, p \sum_i (M_i - M)^2 + n\, p \sum_j (M_j - M)^2 \\ + m\, n\, p\, \sigma_\varepsilon^2 + 2\, p\, r_{ij} [m\, n \sum_i (M_i - M)^2 \sum_j (M_j - M)^2]^{1/2}. \end{aligned} \tag{42}$$

The first two terms on the right-hand side of equation (42) when divided by the appropriate number of degrees of freedom give, respectively, the mean sum of squares for the row agent and the column agent ($V_i$ and $V_j$). The third and fourth terms together give the interaction sum of squares, thus:

$$p \sum_j \sum_i d_{ij}^2 = m\, n\, p\, \sigma_\varepsilon^2 \\ + 2\, p\, r_{ij} [m\, n \sum_i (M_i - M)^2 \sum_j (M_j - M)^2]^{1/2}. \tag{43}$$

The fourth component of the analysis is the independent estimate of sampling error coming from the within group sum of squares. Table 3 contains the analytical expressions for the various elements of a four-component analysis.

TABLE 3

| Source | $\Sigma$ of Squares | df | M $\Sigma$ of Sq. |
|---|---|---|---|
| Rows | $m\, p \sum_i (M_i - M)^2$ | $n - 1$ | $V_i = \dfrac{m\, p\, \Sigma (M_i - M)^2}{n - 1}$ |
| Cols. | $n\, p \sum_j (M_j - M)^2$ | $m - 1$ | $V_j = \dfrac{n\, p\, \Sigma (M_j - M)^2}{m - 1}$ |
| ixj | $p \sum_i \sum_j d^2{}_{ij}$ | $(n - 1)(m - 1)$ | $V_I = \dfrac{p\, \Sigma\, \Sigma\, d^2}{(n - 1)(m - 1)}$ |
| error | $\sum_i \sum_j \sum_h (X_{hij} - M_{ij})^2$ | $nm(p - 1)$ | $V_\varepsilon = \dfrac{\Sigma\, \Sigma\, \Sigma (X_{hij} - M_{ij})^2}{n\, m(p - 1)}$ |
| Total | $\sum_j \sum_i \sum_h (X_{hij} - M)^2$ | $nmp - 1$ | $V_t$ |

The first term on the right-hand side of equation (43), $mnp\sigma_\varepsilon^2$, is that part of the interaction sum of squares which is attributable to sampling error. In $V_\varepsilon$ we have an independent estimate of the sampling variance of the universe. Hence, dividing $mnp\sigma_\varepsilon^2$ by its appropriate degrees of freedom and *assuming that these two estimates are equal,* we have

$$\frac{m\, n\, p\, \sigma_\varepsilon^2}{(m - 1)(n - 1)} = V_\varepsilon, \tag{44}$$

or

$$m\, n\, p\, \sigma_\varepsilon^2 = (m - 1)(n - 1) V_\varepsilon. \tag{45}$$

Also, as may be seen from Table 3,

$$p\sum_j\sum_i d_{ij}^2=(n-1)(m-1)V_i, \tag{46}$$

$$m\sum_i(M_i-M)^2=\frac{n-1}{p}V_i, \tag{47}$$

$$n\sum_j(M_j-M)^2=\frac{m-1}{p}V_j. \tag{48}$$

Substituting from equations (45), (46), (47), and (48) into equation (43) we have

$$r_{ij}=\frac{V_I-V_\varepsilon}{2\sqrt{\dfrac{V_i\,V_j}{(n-1)(m-1)}}}. \tag{49}$$

It is obvious that if the interaction mean sum of squares is equal to the within group sum of squares, the correlation between agents is zero. It is apparent here, then, that the $V_e$ of equation (22) in a three-component analysis is an estimate of error variance only when the agents are uncorrelated.

Dividing equation (49) through by $V_\varepsilon$ and writing the $F$ tests as

$$F_I=\frac{V_I}{V_\varepsilon},\quad F_i=\frac{V_i}{V_\varepsilon},\quad \text{and}\quad F_j=\frac{V_j}{V_\varepsilon},$$

(49) becomes

$$r_{ij}=\frac{F_I-1}{2\sqrt{\dfrac{F_i\,F_j}{(n-1)(m-1)}}}. \tag{50}$$

It is apparent also from these equations that a mean interaction sum of squares less than the mean error sum of squares signifies a negative correlation between agents.

The generalization of these expressions to higher-order interactions should be pursued. It may be that higher-order interactions will require the postulation of a "mutual" correlation as a correlation between any number of agents.

## REFERENCES

1. Peters, C. C. Interaction in analysis of variance interpreted as intercorrelation. *Psychol. Bull.*, 1944, 41, 287-99.
2. Peters, C. C. and Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill, Inc., 1940, p. 353.
3. Siegel, Paul S. and Stuckey, Helen L. The diurnal course of water and food intake in the normal mature rat. *J. comp. physiol. Psychol.*, 1947, 40, 365-70.