CORRELATION BY THE METHOD OF FACTORS.

BY A. A. KRISHNASWAMI AYYANGAR.

Received February 14, 1937.
(Communicated by Prof. K. B. Madhava, M.A., A.I.A.)

§ 1. Any correlation table which gives the frequencies \( A_{pq} \) for the simultaneous occurrence of paired values \( (p, q) \) of the variates \( (x, y) \) may be algebraically represented by the series:

\[
f(x, y) = \sum A_{pq} x^p y^q,
\]

which may be called a correlation series if \( \sum A_{pq} = 1 \). This enables us to obtain simple formulae for the mean, and the standard deviations of the \( x \)- and \( y \)-distributions and also the correlation between \( x \) and \( y \). Further, as will be pointed out presently, if a correlation series can be factorised into other such series, we can express the statistical constants associated with the composite series in terms of those of the component series. Several known results are easily derived from this principle.

Adopting the usual notation for the constants \( \bar{x}, \sigma_x, \rho_{xy} \) (which represent the mean, standard deviation, product moment about the mean) we easily find:

\[
\begin{align*}
\bar{x} &= \sum_{x=y} f_x, \quad \bar{y} = \sum_{x=y} f_y, \\
\sigma_x^2 &= \frac{\partial}{\partial x} \left( x \frac{\partial f}{\partial x} \right)_{x=y=1} - (f_{xx} + f_x - f_x^2) \\
\end{align*}
\]

\[
\begin{align*}
\sigma_y^2 &= (f_{yy} + f_y - f_y^2), \\
\rho_{xy} &= \left( \frac{\partial^2 f}{\partial x \partial y} \right)_{x=y=1} - \bar{x} \bar{y} = (f_{xy} - f_x f_y)
\end{align*}
\]

Cor. : When \( f(x, y) \) satisfies the differential equation \( f_{xy} - f_x f_y = 0 \), the coefficient of correlation is 0.

§ 2.1. Lemma: If \( f(x, y) = \prod_{i=1}^n [f_i(x, y)] t_i \), then

(i) \( \frac{1}{f} \cdot f_x = \sum_{i=1}^n t_i f_i f_{lx} \);

(ii) \( \frac{1}{f} \cdot f_{xy} - \frac{1}{f^2} \cdot f_{x} \cdot f_{y} = \sum_{i=1}^n \left[ \frac{t_i}{f_i} \cdot f_{lx} - \frac{t_i}{f^2} \cdot f_{ix} \cdot f_{iy} \right] \);

(iii) \( \frac{1}{f} \cdot f_{xx} - \frac{1}{f^2} \cdot f_{x}^2 = \sum_{i=1}^n t_i \left[ \frac{1}{f_i} \cdot f_{lx} - \frac{1}{f^2} \cdot f_{ix} \right] \).

These results follow readily from logarithmic partial differentiation.
§ 2·2. If the functions are all correlation series the \(f\)'s become all equal to 1, when \(x = 1 = y\) and we have, from (A) and the above lemma:

1. \[\bar{x} = \Sigma t_i \bar{x}_i, \quad \bar{y} = \Sigma t_i \bar{y}_i\]
2. \[\sigma_x^2 = \Sigma t_i \sigma_{x_i}^2, \quad \sigma_y^2 = \Sigma t_i \sigma_{y_i}^2\]
3. \[p_{xy} = \Sigma t_i \bar{p}_{xy_i}\]

where \(\bar{x}_i, \bar{y}_i, \sigma_{x_i}, \sigma_{y_i}, \bar{p}_{xy_i}\) refer to the series \(f_i(x, y)\) and the corresponding symbols without the \(i\)-subscripts refer to \(f(x, y)\). If \(\gamma_i\) be the coefficient of correlation associated with \(f_i(x, y)\), since \(\bar{p}_{xy_i} = \gamma_i \sigma_{x_i}, \sigma_{y_i}\), we have the following theorem which expresses the correlation associated with a series in terms of those of its factors.

**Theorem:** If \(\sigma_{x_i}, \sigma_{y_i}, \gamma_i\) be the standard deviations and the coefficient of correlation associated with the series \(f_i(x, y)\), then the coefficient of correlation associated with the function \[\prod_{i=1}^{n} \{f_i(x, y)\}^k\] is

\[\sum_{i=1}^{n} \gamma_i \sigma_{x_i} \sigma_{y_i} \sqrt{\sum_{i=1}^{n} \sigma_{x_i}^2 \cdot \sum_{i=1}^{n} \sigma_{y_i}^2}\]

**Cor.:** The correlation coefficient is invariant when \(t_i\) is changed to \(kt_i\) \((i = 1, 2, \ldots, n)\), i.e., \(f(x, y)\) and \(\{f(x, y)\}^k\) have the same correlation.

§ 3. The results given above have several applications to well-known data. We will be content with pointing out a few.

(i) Let \(f(x, y) = P + Qx\) where \(P + Q = 1\).

Here \(\bar{x} = Q, \sigma_{x}^2 = Q - Q^2 = PQ, \sigma_{y}^2 = 0\); hence for the binomial distribution \((P + Qx)^n\), the mean and the standard deviations are respectively \(nP\) and \(\sqrt{nQP}\).

(ii) Let \(f_1(x, y) = Q + Px, f_2(x, y) = Q + Py\), and \(f_3(x, y) = Q + Pxy\), where \(P + Q = 1\).

Then \(\bar{x}_1 = P, \bar{y}_1 = 0, \sigma_{x_1}^2 = PQ, \sigma_{y_1}^2 = 0, \gamma_1 = 0\);

\(\bar{x}_2 = 0, \bar{y}_2 = P, \sigma_{x_2}^2 = 0, \sigma_{y_2}^2 = PQ, \gamma_2 = 0\);

\(\bar{x}_3 = P, \bar{y}_3 = P, \sigma_{x_3}^2 = PQ, \sigma_{y_3}^2 = PQ, \gamma_3 = 1\).

Hence, for the function \[(Q + Px)^{t_1} (Q + Py)^{t_2} (Q + Pxy)^{t_3}\], we have

\(\bar{x} = P(t_1 + t_3), \quad \bar{y} = P(t_2 + t_3), \quad \sigma_{x}^2 = (t_1 + t_3)PQ, \sigma_{y}^2 = (t_2 + t_3)PQ, \gamma = t_3/\sqrt{(t_1 + t_3)(t_2 + t_3)}\).

In particular, if \(t_1 = t_2\), then \(\gamma = t_3/(t_1 + t_3)\).

These results have an important significance and are related to the following theorem:—