A Formula in the Theory of Correlation

By A. E. Hope.

The formula in question gives the correlation between two variables for a number $n_1 + n_2$ of individuals when the ordinary parameters, including correlations, are available for the separate groups of n_1 and n_2 individuals; and it is extended to more than two groups. It is possible, even probable, that the formula is known, but the writer has not found any direct references to it; in any case its practical value is such that there can be no harm in drawing attention to it.

In the first place there is the well known formula connecting the standard deviation σ with the corresponding root-mean-square deviation s about a "provisional mean" differing by h units from the true mean. It is

$$s^2 = \sigma^2 + h^2. \tag{1}$$

Next, suppose two groups of n_1 and n_2 individuals measured respectively with regard to some character x, the means being m_1 and m_2. Let m be the mean of the whole group $n = n_1 + n_2$. Then

$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}. \tag{2}$$

Let σ_1, σ_2 be the standard deviations of the two groups, σ that of the whole group. Then

$$\begin{aligned} n_1 \sigma_1^2 &= \Sigma (x - m_1)^2 \\ &= \Sigma (x - m + m - m_1)^2, \end{aligned}$$

and similarly for $n_2 \sigma_2^2$.

The "total" squared deviation of the group $n_1 + n_2$ is

$$(n_1 + n_2)\sigma^2,$$

derived partly from σ_1^2 and partly from σ_2^2, these, however, being modified by being taken from m as origin. Hence we have

$$(n_1 + n_2)\sigma^2 = n_1 s_1^2 + n_2 s_2^2, \tag{3}$$

where

$$\begin{aligned} s_1^2 &= \sigma_1^2 + (m - m_1)^2, \\ s_2^2 &= \sigma_2^2 + (m - m_2)^2, \text{ by (1).} \end{aligned}$$

From (2) we find

$$m - m_1 = n_2 (m_2 - m_1)/(n_1 + n_2), \qquad m - m_2 = n_1 (m_1 - m_2)/(n_1 + n_2).$$

https://doi.org/10.1017/S1757748900002310 Published online by Cambridge University Press

Now writing $m_1 - m_2 = h$, we obtain from (3)

$$(n_1 + n_2)\,\sigma^2 = n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + \frac{n_1\,n_2^2\,h^2}{(n_1 + n_2)^2} + \frac{n_1^2\,n_2\,h^2}{(n_1 + n_2)^2}$$

$$= n_1\,\sigma_1^2 + n_2\,\sigma_2^2 + n_1\,n_2\,h^2/(n_1 + n_2), \tag{4}$$

which gives σ^2 in terms of known quantities.

Again, let m_{1x} and σ_{1x} denote the mean and s.d. of group n_1 in regard to x, with m_{2x} and so on for the group n_2, and similar notation with regard to y. Let r_1 and r_2 be the correlation coefficients of x and y in the groups, r that in the whole group; m_x, m_y the "total" means, σ_x, σ_y the corresponding s.d.'s; h_x the difference of means of groups in x; h_y similarly; Σ_1, Σ_2 summations over the groups, Σ the "total" summation. Then

$$r = \frac{\Sigma\,(x - m_x)\,(y - m_y)}{(n_1 + n_2)\,\sigma_x\,\sigma_y}. \tag{5}$$

Now

$$\Sigma_1\,(x - m_x)\,(y - m_y) = \Sigma_1\,(x - m_{1x} + m_{1x} - m_x)\,(y - m_{1y} + m_{1y} - my)$$

$$= \Sigma_1\,(x - m_{1x})\,(y - m_{1y}) + n_1\,n_2^2\,h_x\,h_y/(n_1 + n_2)^2,$$

after a little working. Adding this to the corresponding result involving Σ_2, we derive

$$\Sigma\,(x - m_x)\,(y - m_y) = \Sigma_1\,(x - m_{1x})\,(y - m_{1y}) + \Sigma_2\,(x - m_{2x})\,(y - m_{2y}) + \frac{n_1\,n_2}{n_1 + n_2}\,h_x\,h_y, \tag{6}$$

or, by definition,

$$(n_1 + n_2)\,r\sigma_x\,\sigma_y = n_1\,r_1\,\sigma_{1x}\,\sigma_{1y} + n_2\,r_2\,\sigma_{2x}\,\sigma_{2y} + \frac{n_1\,n_2}{n_1 + n_2}\,h_x\,h_y. \tag{7}$$

The values of σ_x, σ_y being already given in (4), we have finally

$$r = \frac{n_1\,r_1\,\sigma_{1x}\,\sigma_{1y} + n_2\,r_2\,\sigma_{2x}\,\sigma_{2y} + n_1\,n_2\,h_x\,h_y/(n_1 + n_2)}{[n_1\,\sigma_{1x}^2 + n_2\,\sigma_{2x}^2 + n_1\,n_2\,h_x^2/(n_1 + n_2)]^{\frac{1}{2}}\,[n_1\sigma_{1y}^2 + n_2\,\sigma_{2y}^2 + n_1\,n_2\,h_y^2/(n_1 + n_2)]^{\frac{1}{2}}}, \tag{8}$$

which is the formula required.

The structure of the formula appears most clearly in the extension to any number of pooled groups, which is not difficult to establish. It is then

$$r = \frac{\Sigma n_i\,r_i\,\sigma_{ix}\,\sigma_{iy} + (\Sigma n_i\,n_j\,h_{ijx}\,h_{ijy})/\Sigma n_i}{[\Sigma\,n_i\,\sigma_{ix}^2 + (\Sigma n_i\,n_j\,h_{ijx}^2)/\Sigma n_i]^{\frac{1}{2}}\,.\,[\Sigma n_i\,\sigma_{iy}^2 + (\Sigma n_i\,n_j\,h_{ijy}^2)/\Sigma n_i]^{\frac{1}{2}}}.$$

https://doi.org/10.1017/S1757748900002310 Published online by Cambridge University Press

For practical workings with the formula (8) a useful variant is obtained by employing not standard deviations but "variances" V defined by

$$n\sigma^2 = c^2 V,$$

where c is the "class interval." We then have

$$n_1 \sigma_{1x}^2 = c_{1x}^2 V_{1x}, \text{ etc.},$$

$$n_1 r_1 \sigma_{1x} \sigma_{1y} = r_1 c_{1x} c_{1y} \surd (V_{1x} V_{1y}).$$

The reader may be left to make the substitution, which takes a specially useful form when, as is normally the case, the class intervals for both groups in x, as well as in y, are the same.

The Probability Distribution of a Bridge Hand

By J. B. Marshall.

The probability distribution of a bridge hand affords a good example of drawings without replacement from a limited stock.

Let n drawings be made from such a stock. Let p_{rs} and q_{rs} be the probabilities of success and failure after there have been r drawings with s successes, and let the probabilities in successive drawings be connected by the relation

$$p_{rs}\, q_{r+1,\,s+1} = q_{rs}\, p_{r+1,\,s}. \tag{1}$$

[This relation is easily seen to hold in the case of a bridge hand. For if b is the number of cards left in the pack after r drawings, and if a is the number which will give a successful result, then

$$p_{rs} = a/b, \quad p_{r+1,\,s} = a/(b-1), \quad p_{r+1,\,s+1} = (a-1)/(b-1),$$

whence

$$\begin{aligned} p_{rs}\, q_{r+1,\,s+1} &= \frac{a}{b} \times \frac{(b-1)-(a-1)}{b-1} \\ &= \frac{b-a}{b} \times \frac{a}{b-1} \\ &= q_{rs}\, p_{r+1,\,s}.] \end{aligned}$$

Let us, in the usual manner, construct a generating function (G.F.) by introducing a variable t, the powers of which will enumerate

https://doi.org/10.1017/S1757748900002310 Published online by Cambridge University Press