35

THE DISTRIBUTION OF THE PARTIAL CORRELATION COEFFICIENT

1. The theoretical distribution

In ascertaining the exact distribution in random samples to which the correlation coefficient between two normally distributed variates is subject, the following geometrical device was utilised (1. 1915).

Let $x_1, x_2, \ldots\ldots, x_n$ represent the n values of one variate in the sample, and $y_1, y_2, \ldots\ldots, y_n$ the n values of the second variate; let $\bar{x}$ and $\bar{y}$ be the means; then the quantities $x - \bar{x}$, $y - \bar{y}$ may be regarded as coordinates in n-fold Euclidian space of two points P and Q. If O be the origin of coordinates the lengths OP and OQ will be proportional to the standard deviations of the two variates as estimated from the sample; upon the *directions* of OP and OQ will the correlational properties of the two variates depend. In particular it is easy to see that the correlation between x and y as estimated by the formula,

$$r = \frac{S(x - \bar{x})(y - \bar{y})}{\sqrt{S(x - \bar{x})^2 . S(y - \bar{y})^2}},$$

will be the cosine of the angle between OP and OQ.

It may be noted that if any pair of observations, say x_1 and y_1, be omitted, the effect of doing so will be to project the above mentioned figure upon a region at right angles to one of the axes of coordinates. Hence the distribution of r as obtained from the projected angle, will be the same as that for $(n - 1)$ observations taken at random.

The fact that r depends only on the angle between two *radii vectores* shows that it is invariant for orthogonal transformations of coordinates; we may therefore enunciate the following proposition. *If* $x'_1, x'_2, \ldots\ldots x'_n$ *be the coordinates of* P *with respect to any system of rectangular axes through* O, *and* $y'_1, y'_2, \ldots\ldots y'_n$ *the corresponding functions of* $y_1, y_2, \ldots\ldots y_n$, *then the correlation, in any sample, between* x' *and* y' *is equal to that between* x *and* y. Moreover, if $x_1, x_2, \ldots\ldots x_n$, be normally and independently distributed then also will $x'_1 \ldots\ldots x'_n$ be normally and independently distributed.

This proposition may be applied to a variety of problems. It leads directly to the solution of the distribution of the partial correlation coefficient. For let $z_1, z_2, \ldots\ldots z_n$ be the va-

lues of any third variate, and let R be the point whose coordinates are $z - \bar{z}$. Then the correlations of z with x and y are the cosines of the angles which OR makes with OP and OQ; that is of the sides of the spherical triangle determined by the radii OP, OQ and OR; but the partial correlations are the cosines of the angles of this triangle. So that the partial correlation between x and y, is the cosine of the angle between the projections of OP and OQ upon the region perpendicular to OR.

If therefore we choose a new set of orthogonal coordinates such that one of them lies along OR, it appears that the total correlation between x and y is still the correlation obtained from n independent pairs of values of normally distributed variates, but that the partial correlation is the correlation obtained from $(n-1)$ independent pairs of normally distributed variates. Consequently the random sampling distribution of the partial correlation obtained from n pairs of values, when one variate is eliminated, is the same as the random sampling distribution of a total correlation derived from $(n-1)$ pairs. By mere repetition of the above reasoning it appears that when s variates are eliminated the effective size of the sample is diminished to $(n-s)$. The investigations of YULE in 1907 (6) indicated that the probable error of a partial correlation should be the same as that for a total correlation derived from a sample of the same size. Our result may be regarded as confirming this approximation provided that the number of variates eliminated is a small fraction of the number of the sample.

2. EXPERIMENTAL EVIDENCE

J. W. BISPHAM has investigated experimentally the distribution of the partial correlation coefficient in three experiments. In the first of these (2. 1920) 1000 values were obtained with uncorrelated variates. The distribution of the variates was far from normal, and the conditions of sampling allowed no variation in the standard deviations. In spite of this, however, the distribution of the total correlations seem to agree with the theoretical expectation. In the second and third experiments (3. 1923), giving 200 and 100 values respectively, the conditions of sampling allow of a nearer approach to random normal samples; in these experiments the variates were highly correlated, though the partial correlation investigated is still nearly zero. The distribution of the total cor-

relations was shown to agree with the theoretical expectation, the latter being calculated by the laborious method developed in *Biometrika* (4. 1917) and not by the direct transformation since given in *Metron* (5. 1921). In all cases the size of the sample was 30; so that the variance of the true distribution of the partial correlations was 1/28, whereas that for the total correlation would be 1/29. To discriminate experimentally between these two values would need samples amounting to about 6000, whereas in all we have only 1300 observations.

BISPHAM only compares his results with the theoretical distribution of the total correlation coefficient, and finds the values to be effectively in agreement. This is to be expected. It may be noted, however, that in all these cases the observed standard deviation exceeds his expectation. A more refined test of the agreement may be made by noting that, if r is normally distributed with standard deviation σ, then, for a sample of n,

$$\frac{1}{\sigma^2} S \cdot (r - \bar{r})^2$$

is distributed in random samples in the χ^2 distribution corresponding to $(n-1)$ degrees of freedom; consequently its expected value is $(n-1)$. We give below the expected values, and those calculated from the observations taking $\frac{1}{\sigma^2}$ equal to 28 and to 29.

Experiment	1	2	3	Total	$\sqrt{2\chi^2} - \sqrt{2593}$
$28\,S(r-\bar{r})^2$	980.78	210.20	101.27	1292.25	—.104
$29\,S(r-\bar{r})^2$	1015.81	217.71	104.80	1338.41	+.816
Expectation	999	199	99	1297	

It is evident that the value calculated from the true distribution is somewhat nearer to the expectation than that derived on the supposition that the partial and the total correlations have the same distribution. The difference, however, cannot be regarded as significant. This may be directly tested by comparing the values of $\sqrt{2\chi^2}$ with $\sqrt{2n-1}$, where n is the expectation. The difference has a standard deviation of 1. The values do, however, show the advantage of the true variance.

The differences would, of course, be more strongly marked in smaller samples, and especially so if several variates were eliminated.

REFERENCES

1. R. A. Fisher (1915), *The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.* Biometrika, X pp. 507-521.
2. J. W. Bispham (1920), *An experimental determination of the distribution of the partial correlation coefficient in samples of thirty.* Royal Soc. Proc. A. XCVII, pp. 218-224.
3. J. W. Bispham (1923), *An experimental determination of the distribution of the partial correlation coefficient in samples of thirty.* Metron II. pp. 684-696.
4. *Co-operative Study* (1917), *On the distribution of the correlation coefficient in small samples.* Biometrika XI, pp. 328-413.
5. R. A. Fisher (1921), *On the "Probable Error" of a coefficient of correlation deduced from a small sample.* Metron I, Pt. 4, pp. 1-32.
6. G. Udny Yule (1907), *On the theory of correlation for any number of variables treated by a new system of notation.* Roy. Soc. Proc. A. LXXIX, pp. 182-193.