61

THE GENERAL SAMPLING DISTRIBUTION OF THE MULTIPLE CORRELATION COEFFICIENT

Author's Note (CMS 14.653a) *

The introductory section says almost all that need be said about this paper. In 1938 Bose and Roy found that Mahalanobis' D^2 statistic, used for measuring the "generalised distance" in multivariate analysis, was of the form of (C) as defined here. Both (A) and (C) will, I am sure, reappear in other problems, and their tabulation, perhaps in a form comparable with that of the limiting distribution (B) given in this paper, is much to be desired. Since (B) is the limiting form as $n_2 \to \infty$ of both (A) and (C), the simplest approach to the problem of tabulation would seem to be to add tables like that of (B) for $n_2 = 144, 36, 16, 9$, interpolation among which would make the general functions readily accessible for all reasonably large values of n_2. The distributions for small even values, if required, are easily obtained from the elementary cases set out in this paper.

The General Sampling Distribution of the Multiple Correlation Coefficient.

By R. A. FISHER, Sc.D., Rothamsted Experimental Station, Harpenden, Herts.

(Communicated by Sir James Jeans, Sec. R.S.—Received August 24, 1928.)

1. *Introductory.*

Of the problems of the exact distribution of statistics in common use that of the multiple correlation coefficient is the last to have resisted solution. It will be seen that the solution introduces an extensive group of distributions, occurring naturally in the most diverse types of statistical investigation, and which in their mere mathematical structure supply an extension in a new direction of the entire system of distributions previously obtained, which in their generality underlie the analysis of variance. The individual distributions of this system were in each case obtained by the exact investigation of a particular problem. It was realised only gradually that many of these distributions, disguised by the different notations appropriate to different problems, were in reality equivalent, and could be made available in practice by a single system of tables. The remaining cases, with the notable exception of the correlation coefficient, then fall into place as particular limiting forms of a single general distribution. As the practical utility of these earlier solutions depends greatly on a recognition of their place in a single system, a very brief account of their mutual relationship may be given.

The only statistic derived from samples of a continuous variate, of which the distribution was known before the present century, appears to be the arithmetic mean of a sample drawn from the normal distribution. In addition, however, two distributions which may be regarded as distributions of statistics had also been found, namely, Bernoulli's binomial distribution, and Poisson's series. Both of these distributions possess the property that the aggregate of the values of a sample is itself distributed in a distribution of the same type. In all three classical cases, therefore, the distribution of the statistic derived from a finite sample was known only by a mathematical simplification of this special type. In all other cases, approximations of unknown accuracy based on the use of the standard error and the assumption of normal distributions had perforce to be used.

In 1908 " Student "* attacked the problem of the distribution of the mean of a normal sample measured, as in practice it must usually be, in terms of the standard error as estimated from the same sample. He was thus incidentally led to the equally fundamental distribution of the variance of a normal sample. This latter, to which the general distribution of the analysis of variance degenerates when n_2 tends to infinity, is in reality equivalent to the distribution found by Pearson† in 1900 for the χ^2 measure of discrepancy developed for testing goodness of fit. From this " Student " was able to derive the exact distribution (the distribution of t) of the mean of a unique sample, which as subsequently appeared falls into the same system with $n_1 = 1$. The two principal limiting forms of the general distribution were thus known in 1908, and were available for practical application by means of Elderton's‡ and " Student's "* first tables.

In 1915 the distribution of the coefficient of correlation was obtained§ by a use of Euclidean hyper-space similar to that employed below. The same method served at the same time to put " Student's " results upon a rigorous basis. The distribution of the correlation coefficient stands outside the analysis of variance system, but, as will be seen in the present paper, it is brought into coherent connection with it by the distribution of the coefficient of multiple correlation. When, however, the corresponding distribution of the intraclass correlation was obtained||, the distribution found was of a new and different type, which, as subsequently appeared, was the general distribution of the analysis of variance, in which the variance is analysed into two parts representing that within and that between the classes or " fraternities " of which the data are composed. This was the first instance of the general distribution which from the notation there used is distinguished as the distribution of z.

The recognition of the fundamental importance of the two parameters, n_1 and n_2, which specify the numbers of degrees of freedom in the two estimates of variance to be compared, and the recognition of the distribution of χ^2 as equivalent to that of an estimate of variance led, in 1922 and the following two years,¶ to the demonstration that it is always the number of degrees of freedom

* ' Biometrika,' vol. 6, p. 1 (1908).

† ' Phil. Mag.,' vol. 50, p. 157 (1900).

‡ ' Biometrika,' vol. 1, p. 155 (1902).

§ Fisher, ' Biometrika,' vol. 10, p. 507 (1915).

|| ' Metron.,' vol. 1, No. 4, p. 1 (1921).

¶ ' J. R. Stat. Soc.,' vol. 85, p. 87 (1922); ' Economica,' vol. 3, p. 139 (1923); ' J. R. Stat. Soc.,' vol. 87, p. 442 (1924).

which is to be used in applying the test of goodness of fit. The further proof that the test is only valid when the methods of estimation employed have been *efficient*, binds the theory of goodness of fit closely to that of estimation in the development of which the exact distribution of statistics play an essential part.*

Meanwhile,† a solution of the exact distribution of χ^2 when applied to test the goodness of fit of regression formulæ had shown that a modification was required in this case, which, in fact, involved dropping the approximate assumption that n_2 was infinite, and reduced the general distribution to the same form as that already found in the study of intraclass correlation. At the same time, the distribution of the correlation ratio, η, derived from uncorrelated material, was shown to belong to the same class with n_1, equal to one less than the number of arrays; and the distribution of regression coefficients, whether total or partial, and whether employed in a linear or a non-linear formula, were shown to conform to "Student's" distribution. The solution of the distribution of the correlation ratio η really included also that of the multiple correlation coefficient for samples drawn from uncorrelated material, the distribution of which was given in its appropriate notation in 1924.‡

In the same year§ it was found possible to use the representation in hyperspace to demonstrate that the distribution of the partial correlation coefficients is exactly the same as that primarily found for the total correlation, provided that unity is deducted from the sample number for each variate eliminated.

Each distinct type of distribution found has thus occurred repeatedly in different investigations; whereas, however, nearly all cases are reducible to a common type capable of exact treatment by the same simple arithmetical procedure,|| and requiring the same fundamental table, the distribution of the
* (intraclass) correlation coefficient, total or partial, stood aside from the main system, and was capable of only an approximate treatment by using the distribution of z.

The distribution of the multiple correlation coefficient, apart from the practical necessity of applying to observed results sufficiently exact tests of significance, is thus of great theoretical interest owing to the close connection which must exist between it and the simple correlation coefficient, on the one hand, and, on the other, to the form already obtained from uncorrelated material.

* 'Phil. Trans.,' A, vol. 222, p. 309 (1922).

† 'J. R. Stat. Soc.,' vol. 85, p. 597 (1922).

‡ 'Phil. Trans.,' B, vol. 213, p. 89 (1924).

§ 'Metron.,' vol. 3, p. 329 (1924).

|| 'Statistical Methods for Research Workers,' 2nd ed., Oliver & Boyd, Edinburgh, 1928.

* For intraclass, read interclass.

The latter solution involves, besides the variate and frequency, the two parameters n_1 and n_2, and is therefore a functional relation between four variables. The new solution necessarily involves also the multiple correlation in the population sampled, making a fifth variable; complete tabulation of the results would thus require a table of fourfold entry; even confining attention to specified points of special importance, such as the 5 per cent. and 1 per cent. points, a procedure that has made tabulation practicable for the distribution of z, we should still have tables of triple entry. The problem of adequate tabulation is certainly not insurmountable, but to ascertain the proper method to adopt in its presentation will require further study of the nature of the function. The table of the 5 per cent. points of the distribution of B (Section 5) should in the author's opinion provide sufficient guidance for the greater number of practical applications.

2. *Method of Solution.*

The primary problem of the sampling distribution of the correlation coefficient between two variates, x and y, was originally solved by interpreting the n individual values of either variate appearing in the sample as the co-ordinates of a point in Euclidean space of n dimensions. It then easily appeared that the correlation coefficient between the variates was the cosine of the angle between the two *radii vectores* drawn from the origin to points, the co-ordinates of which represented the deviations from the mean of the sample of the two variates concerned..

The frequency with which r, the observed correlation coefficient, falls in any infinitesimal range dr may be usefully thought of as the product of two factors, one being the generalised volume in which the second sample point may lie so that the correlation may fall within the assigned range, this value being independent of the correlational properties of the population sampled, while the second is a factor by which the frequency density in any element of volume is modified by the correlation between x and y in the population. With zero correlation in the population, the frequency density at any point depends only on its distance from the origin, and since for any given distance the point is free to move over a sphere in $(n - 1)$ dimensions, one dimension having been eliminated by using the sample mean as origin, it is easy to see that for this case the frequency distribution of r is given by

$$df = \frac{[\frac{1}{2}(n-3)]!}{[\frac{1}{2}(n-4)]!\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(n-4)}\,dr.$$

The general solution of the primary problem*

$$df = \frac{n-2}{\pi}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}\int_0^\infty \frac{dz}{(\cosh z - \rho r)^{n-1}} \cdot dr,$$

may be written with advantage

$$df = \frac{[\frac{1}{2}(n-3)]!}{[\frac{1}{2}(n-4)]!\sqrt{\pi}}(1-r^2)^{\frac{1}{2}(n-4)}\,dr \times \frac{[\frac{1}{2}(n-2)]!}{[\frac{1}{2}(n-3)]!\sqrt{\pi}}(1-\rho^2)^{\frac{1}{2}(n-1)}\int_{-\infty}^{\infty}\frac{dz}{(\cosh z - \rho r)^{n-1}}.$$

The second factor then represents the effect upon the frequency density, in the region represented by dr, of a correlation ρ in the sampled population: the numerical part of this factor is merely such as to reduce it to unity when $\rho = 0$.

With multiple correlations we are concerned with the correlations between a dependent variate y, and a number of independent variates, $x_1, x_2, \ldots, x_{n_1}$, and, moreover, with the correlations of the latter among themselves. It was not at first obvious that the sampling distribution did not involve this whole matrix of correlations, in which case, even if it could be determined, it would be of no practical use. The argument, by which it can be seen to depend from only a single parameter of the population, is therefore of special interest, as by its general character it is applicable to a number of statistical problems, and leads in this case directly to the solution.

The multiple correlation of y with $x_1, x_2, \ldots, x_{n_1}$ is the correlation between y and that linear function of $x_1, x_2, \ldots, x_{n_1}$ with which its correlation is highest. If, therefore, for the dependent variates, x, we substitute an equal number of new variates, ξ, defined as linear functions of the n_1 variates, x, then the multiple correlation in the population, and in every sample from it, will remain unchanged. In particular we may choose as ξ_1, that linear function the correlation of which with y in the population sampled is highest, and for the remaining variates, ξ, we can choose linear functions of x, uncorrelated with ξ_1, or with each other. In choosing the last of these we have no more than $n_1 - 1$ conditions to be satisfied by the ratios of n_1 coefficients. If this is done it is easy to see, or to demonstrate, that all of the variates ξ, except ξ_1, have zero correlation with y. Using the variates ξ the sampling distribution of the multiple correlation R can only depend on the correlation in the population sampled between ξ_1 and y, namely, on the multiple correlation in the population sampled, which we may designate by ρ.

* Fisher, 'Biometrika,' vol. 10, p. 507 (1915).

The geometrical interpretation of the multiple correlation coefficient R is that it is the cosine of the angle between the *radius vector* of the dependent variate y and the planar region including the *radii vectores* of the n_1 independent variates; its distribution when $\rho = 0$, which depends only on the geometrical elements of volume, has been thus shown* to be

$$df = \frac{[\frac{1}{2}(n_1 + n_2 - 2)]!}{[\frac{1}{2}(n_1 - 2)]![\frac{1}{2}(n_2 - 2)]!} (R^2)^{\frac{1}{2}(n_1-2)} (1 - R^2)^{\frac{1}{2}(n_2-2)} d(R^2),$$

where n, the sample number, is replaced by $n_1 + n_2 + 1$; but in what way this distribution is modified when ρ is not zero has been hitherto entirely unknown.

It is evident, however, that since we have reduced the problem of the multiple correlation coefficient to one involving only a single correlation, the frequency density of any configuration will be affected merely by a factor

$$\frac{[\frac{1}{2}(n-2)]!}{[\frac{1}{2}(n-3)]!\sqrt{\pi}} (1 - \rho^2)^{\frac{1}{2}(n-1)} \int_{-\infty}^{\infty} \frac{dz}{(\cosh z - \rho r)^{n-1}},$$

in which r is the correlation in the sample between y and ξ_1; this factor will, however, vary, because r varies in the different configurations which give rise to the same value of R. Consider now a third variate, Y, representing the linear function of the independent variates which in the sample is most closely correlated with y, or, in other words, the prediction formula for y. Its correlation with ξ_1 we may represent by cos ψ, and since the partial correlation of y with ξ_1 (or any other linear function of the independent variates) when Y is eliminated, must be zero, it is evident that

$$r = R \cos \psi.$$

For a given value of ψ, therefore, the density factor is constant in the different configurations which give the same value of R, but, in the absence of correlation, the frequency with which ψ falls in the range $d\psi$ is evidently

$$\frac{[\frac{1}{2}(n_1 - 2)]!}{[\frac{1}{2}(n_1 - 3)]!\sqrt{\pi}} \sin^{n_1-2}\psi \, d\psi;$$

hence integrating over all values of ψ, the density factor becomes

$$\frac{(1 - \rho^2)^{\frac{1}{2}(n_1+n_2)}}{\pi} \cdot \frac{[\frac{1}{2}(n_1 + n_2 - 1)]!}{[\frac{1}{2}(n_1 + n_2 - 2)]!} \times \frac{[\frac{1}{2}(n_1 - 2)]!}{[\frac{1}{2}(n_1 - 3)]!} \int_0^{\pi} d\psi \int_{-\infty}^{\infty} \frac{\sin^{n_1-2}\psi \, . \, dz}{(\cosh z - \rho R \cos \psi)^{n_1+n_2}},$$

* Fisher, 'Phil. Trans.,' B, vol. 213, p. 89 (1924).

and the complete expression for the distribution of R is

$$df = \frac{[\frac{1}{2}(n_1+n_2-1)]!}{[\frac{1}{2}(n_2-2)]![\frac{1}{2}(n_1-3)]!}\cdot\frac{(1-\rho^2)^{\frac{1}{2}(n_1+n_2)}}{\pi}$$
$$\times (R^2)^{\frac{1}{2}(n_1-2)}(1-R^2)^{\frac{1}{2}(n_1-2)}\,d(R^2)\int_0^{\pi}d\psi\int_{-\infty}^{\infty}\frac{\sin^{n_1-2}\psi\,.\,dz}{(\cosh z-\rho R\cos\psi)^{n_1+n_2}}.$$

3. *The Hypergeometric Form.*

Apart from the factor,

$$(1-\rho^2)^{\frac{1}{2}(n_1+n_2)},$$

the density factor may be reduced to a hypergeometric function. For in the expression,

$$\frac{1}{\pi}\cdot\frac{[\frac{1}{2}(n_1+n_2-1)]!}{[\frac{1}{2}(n_1+n_2-2)]!}\cdot\frac{[\frac{1}{2}(n_1-2)]!}{[\frac{1}{2}(n_1-3)]!}\int_0^{\pi}d\psi\int_{-\infty}^{\infty}\frac{\sin^{n_1-2}\psi\,.\,dz}{(\cosh z-\rho R\cos\psi)^{n_1+n_2}},$$

the integrand may be expanded in the uniformly convergent series

$$\sum_{t=0}^{\infty}\frac{(n_1+n_2+2t-1)!}{(n_1+n_2-1)!\,(2t)!}\cdot\frac{\cos^{2t}\psi\sin^{n_1-2}\psi}{\cosh^{n_1+n_2+2t}z}(\rho^2R^2)^t,$$

in which the odd powers of $\cos\psi$, which evidently disappear on integration, have been omitted. Remembering now that

$$\int_0^{\pi}\cos^{2t}\psi\sin^{n_1-2}\psi\,d\psi=\frac{[\frac{1}{2}(2t-1)]![\frac{1}{2}(n_1-3)]!}{[\frac{1}{2}(n_1+2t-2)]!},$$

and

$$\int_{-\infty}^{\infty}\frac{dz}{\cosh^{n_1+n_2+2t}z}=\frac{[\frac{1}{2}(n_1+n_2+2t-2)]!\sqrt{\pi}}{[\frac{1}{2}(n_1+n_2+2t-1)]!},$$

we have a power series for the integral, which may be written

$$\frac{[\frac{1}{2}(n_1-2)]!}{[\frac{1}{2}(n_1+n_2-2)]!^2}\sum_{t=0}^{\infty}\frac{[\frac{1}{2}(n_1+n_2+2t-2)]!^2}{t!\,[\frac{1}{2}(n_1+2t-2)]!}(\rho^2R^2)^t,$$

or

$$F[\tfrac{1}{2}(n_1+n_2),\ \tfrac{1}{2}(n_1+n_2),\ \tfrac{1}{2}n_1,\ \rho^2R^2],$$

so that the distribution of R obtained in section 2 takes the form

$$df=\frac{[\frac{1}{2}(n_1+n_2-2)]!}{[\frac{1}{2}(n_1-2)]![\frac{1}{2}(n_2-2)]!}(1-\rho^2)^{\frac{1}{2}(n_1+n_2)}F[\tfrac{1}{2}(n_1+n_2),\tfrac{1}{2}(n_1+n_2),\tfrac{1}{2}n_1,\rho^2R^2]$$
$$\times(R^2)^{\frac{1}{2}(n_1-2)}(1-R^2)^{\frac{1}{2}(n_1-2)}\,d(R^2).\quad\text{(A)}$$

4. *Elementary Cases.*

4.1. *When n_2 is even.*—When n_2 is even the identity,

$$F[\tfrac{1}{2}(n_1+n_2),\ \tfrac{1}{2}(n_1+n_2),\ \tfrac{1}{2}n_1,\ \rho^2R^2] = (1-\rho^2R^2)^{-\frac{1}{2}(n_1+2n_2)}\ F(-\tfrac{1}{2}n_2,\ -\tfrac{1}{2}n_2,\ \tfrac{1}{2}n_1,\ \rho^2R^2),$$

gives a terminating series.

Thus, when $n_2 = 2$, we have the series of distributions

$$df = (1-\rho^2)^{\frac{1}{2}(n_1+2)}(1-\rho^2R^2)^{-\frac{1}{2}(n_1+4)}(n_1+2\rho^2R^2)\,R^{n_1-1}\,dR,$$

having the special forms

$$(2.2)\qquad df = (1-\rho^2)^2(2+2\rho^2R^2)/(1-\rho^2R^2)^3\,.\,R\,dR,$$

$$(3.2)\qquad df = (1-\rho^2)^{5/2}(3+2\rho^2R^2)/(1-\rho^2R^2)^{7/2}\,.\,R^2\,dR,$$

$$(4.2)\qquad df = (1-\rho^2)^3(4+2\rho^2R^2)/(1-\rho^2R^2)^4\,.\,R^3\,dR$$

when n_1 is 2, 3 and 4.

When $n_2 = 4$, we have a somewhat less simple series of distributions

$$df = \frac{2(1-\rho^2)^{\frac{1}{2}(n_1+4)}}{(1-\rho^2R^2)^{\frac{1}{2}(n_1+8)}}\left\{\frac{n_1(n_1+2)}{2\,.\,4}+2\,\frac{n_1+2}{2}\rho^2R^2+\rho^4R^4\right\}R^{n_1-2}(1-R^2)\,d(R^2),$$

and when $n_2 = 6$, a series which may be written

$$df = \frac{3(1-\rho^2)^{\frac{1}{2}(n_1+6)}}{(1-\rho^2R^2)^{\frac{1}{2}(n_1+12)}}\left\{\frac{n_1(n_1+2)(n_1+4)}{2\,.\,4\,.\,6}+3\,\frac{(n_1+2)(n_1+4)}{2\,.\,4}\rho^2R^2 + 3\,\frac{n_1+4}{2}\rho^4R^4+\rho^6R^6\right\}R^{n_1-2}(1-R^2)^2\,d(R^2),$$

an expression in which the general method of formation of the terms is readily seen.

4.2. *When n_1 and n_2 are both odd.*—A second group of cases in which the frequency element is expressible in finite terms in elementary functions occurs when both n_1 and n_2 are odd. If, for example, we put $n_1 = 3$ in the expression

$$\int_0^\pi d\psi\int_{-\infty}^{\infty}\frac{\sin^{n_1-2}\psi\,.\,dz}{(\cosh z-\rho R\cos\psi)^{n_1+n_2}},$$

and integrate with respect to $\cos\psi$, we obtain

$$\int_{-\infty}^{\infty}\frac{dz}{(n_2+2)\,\rho R}\{(\cosh z-\rho R)^{-(n_2+2)}-(\cosh z+\rho R)^{-n_2+2}\},$$

a form of integral which, as was shown in the case of the simple correlation coefficient*, is expressible in finite terms by the aid of the circular functions.

* Fisher, 'Biometrika,' vol. 10, p. 507 (1915).

For

$$\int_{-\infty}^{\infty} \frac{dz}{\cosh z - \rho R} = \frac{2\theta}{\sin\theta},$$

where $\cos\theta = -\rho R$, and θ does not exceed the bounds 0 to π; hence

$$\int_{-\infty}^{\infty} \frac{dz}{(\cosh z - \rho R)^{n_2+2}} = \frac{2}{(n_2+1)!}\left(\frac{d}{d\cos\theta}\right)^{n_2+1} \frac{\theta}{\sin\theta},$$

and therefore, if n_2 is odd,

$$\frac{1}{(n_2+2)\,\rho R}\int_{-\infty}^{\infty} \{(\cosh z - \rho R)^{-(n_2+2)} - (\cosh z + \rho R)^{-(n_2+2)}\}\,dz$$
$$= \frac{4}{(n_2+2)!\,\rho R}\left(\frac{d}{d(\rho R)}\right)^{n_2+1} \frac{\sin^{-1}\rho R}{\sqrt{1-\rho^2R^2}}.$$

Hence for the determination of the simpler distributions of this series we require

$$\left(\frac{d}{\cos\phi\, d\phi}\right)^2 \frac{\phi}{\cos\phi} = \frac{1}{\cos^3\phi}(\phi + 3\tan\phi + 3\phi\tan^2\phi)$$

$$\left(\frac{d}{\cos\phi\, d\phi}\right)^4 \frac{\phi}{\cos\phi} = \frac{1}{\cos^5\phi}(9\phi + 55t + 90\phi t^2 + 105t^3 + 105\phi t^4)$$

$$\left(\frac{d}{\cos\phi\, d\phi}\right)^6 \frac{\phi}{\cos\phi} = \frac{9}{\cos^7\phi}(25\phi + 231t + 525\phi t^2 + 1190t^3$$
$$+ 1575\phi t^4 + 1155t^5 + 1155\phi t^6,$$

which lead directly to the distributions,

$$(3.1)\quad df = \frac{1}{\pi}(1-\rho^2)^2\,(1-R^2)^{-\frac{1}{2}}\,(1-\rho^2R^2)^{-2}\,\{3+\alpha\,(1+2\rho^2R^2)\}\,R^2dR,$$

in which α stands for

$$\frac{\sin^{-1}(\rho R)}{\rho R\sqrt{1-p^2R^2}} = 1 + \tfrac{2}{3}\rho^2R^2 + \frac{2\cdot 4}{3\cdot 5}\rho^4R^4 + \quad \ldots \textit{ad inf.}$$

$$(3.3)\quad df = \frac{1}{4\pi}(1-\rho^2)^3\,(1-R^2)^{\frac{1}{2}}\,(1-\rho^2R^2)^{-4}\,R^2dR$$
$$\times\{5\,(11+10\rho^2R^2) + 3\alpha\,(3+24\rho^2R^2+8\rho^4R^4)\},$$

$$(3.5)\quad df = \frac{1}{8\pi}(1-\rho^2)^4\,(1-R^2)^{3/2}\,(1-\rho^2R^2)^{-6}\,R^2\,dR$$
$$\times\{7\,(33+104\rho^2R^2+28\rho^4R^4) + 5\alpha(5+90\rho^2R^2+120\rho^4R^4+16\rho^6R^6)\}.$$

A similar process of integration is available for other odd values of n_1; for $n_1 = 5$ we have the distributions

$$(5.1)\qquad df = \frac{1}{4\pi\rho^2}(1-\rho^2)^3\,(1-\mathrm{R}^2)^{-\frac{1}{2}}\,(1-\rho^2\mathrm{R}^2)^{-3}\,\mathrm{R}^2\,d\mathrm{R} \times \{1 + 14\rho^2\mathrm{R}^2 + \alpha\,(-1 + 8\rho^2\mathrm{R}^2 + 8\rho^4\mathrm{R}^4)\}.$$

$$(5.3)\qquad df = \frac{3}{8\pi\rho^2}(1-\rho^2)^4\,(1-\mathrm{R}^2)^{\frac{1}{2}}\,(1-\rho^2\mathrm{R}^2)^{-5}\,\mathrm{R}^2\,d\mathrm{R} \times \{1 + 68\rho^2\mathrm{R}^2 + 36\rho^4\mathrm{R}^4 + \alpha\,(-1 + 18\rho^2\mathrm{R}^2 + 72\rho^4\mathrm{R}^4 + 16\rho^6\mathrm{R}^6)\}.$$

$$(5.5)\qquad df = \frac{1}{64\pi\rho^2}(1-\rho^2)^5\,(1-\mathrm{R}^2)^{3/2}\,(1-\rho^2\mathrm{R}^2)^{-7}\,\mathrm{R}^2\,d\mathrm{R} \times \{25 + 4678\rho^2\mathrm{R}^2 + 8664\rho^4\mathrm{R}^4 + 1648\rho^6\mathrm{R}^6 + \alpha\,(-25 + 800\rho^2\mathrm{R}^2 + 7200\rho^4\mathrm{R}^4 + 6400\rho^6\mathrm{R}^6 + 640\rho^8\mathrm{R}^8)\}.$$

The polynomial coefficient of α in the hypergeometric function is itself easily expressed in terms of a function of this sort, in the forms

$$\frac{3\,.\,5\,\ldots\,(n_1-2)\,.\,3^2\,.\,5^2\,\ldots\,n_2^2}{2^2\,.\,4^2\,\ldots\,(n_1+n_2-2)^2}(-\rho^2\mathrm{R}^2)^{-\frac{1}{2}(n_1-3)}\,\mathrm{F}[-\tfrac{1}{2}(n_1+n_2-2),\ \tfrac{1}{2}(n_1+n_2-2),\ \tfrac{1}{2}(2-n_1),\ \rho^2\mathrm{R}^2],$$

or

$$\frac{n_1-2}{2\,.\,4\,\ldots\,(n_1+n_2-2)}(\rho^2\mathrm{R}^2)^{\frac{1}{2}(n_1+1)}\,\mathrm{F}\left[-\frac{n_2}{2},\ -\frac{n_1+n_2-2}{2},\ 1,\ \frac{1}{\rho^2\mathrm{R}^2}\right];$$

from this the remainder may in any particular case be found fairly easily by equating coefficients in the initial terms of the expansion of

$$\mathrm{F}\,(\tfrac{1}{2}(n_1+n_2), \tfrac{1}{2}(n_1+n_2), \tfrac{1}{2}n_1, \rho^2\mathrm{R}^2).$$

5. *The Problem of Large Samples.*

Some confusion has been caused by the fact that, while for any finite value of ρ, however small, the distribution of R will be normal for a sufficiently large sample, yet when $\rho = 0$ the distribution is far from normal. The approximate distribution appropriate to the theory of large samples, for different values of $\rho\sqrt{n_2}$, may be found as follows.

If we write $n_2\rho^2 = \beta^2$, $n_2\mathrm{R}^2 = \mathrm{B}^2$, and allow n_2 to increase indefinitely, the limiting form taken by the general distribution is

$$df = \frac{(\frac{1}{2}\mathrm{B}^2)^{\frac{1}{2}(n_1-2)}}{[\frac{1}{2}(n_1-2)]!}e^{-\frac{1}{2}\mathrm{B}^2-\frac{1}{2}\beta^2}\left\{1 + \frac{1}{n_1}\frac{\beta^2\mathrm{B}^2}{2} + \frac{1}{n_1(n_1+2)}\frac{\beta^4\mathrm{B}^4}{2\,.\,4} + \ldots\right\}d(\tfrac{1}{2}\mathrm{B}^2), \quad (\mathrm{B})$$

which may be written in terms of a Bessel function as

$$(\mathrm{B}/i\beta)^{\frac{1}{2}(n_1-2)}\,e^{-\frac{1}{2}(\mathrm{B}^2-\beta^2)}\,.\,\mathrm{J}_{\frac{1}{2}(n_1-2)}\,(i\beta\mathrm{B})\,.\,d\,(\tfrac{1}{2}\mathrm{B}^2).$$

When n_1 is odd, these may be reduced to elementary functions; thus for $n_1 = 3$, we have

$$df = \frac{1}{\sqrt{2\pi}} \frac{B}{\beta} \{e^{-\frac{1}{2}(B-\beta)^2} - e^{-\frac{1}{2}(B+\beta)^2}\} . dB,$$

an interesting distribution which connects the extreme forms found by making β zero for uncorrelated populations, and large for populations with a significant though still small correlation. When $\beta = 0$, we have

$$df = (2/\pi)^{\frac{1}{2}} B^2 \exp(-\tfrac{1}{2}B^2)\, dB,$$

the distribution of χ for 3 degrees of freedom, while when β is large, B is distributed normally about β, in the form

$$df = (2\pi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(B-\beta)^2\}\, dB,$$

and therefore R is distributed normally about ρ, with variance which may be equated to $1/n_2$.

When $n_1 = 5$, the system of distributions is

$$df = \frac{1}{\sqrt{2\pi}} \frac{B^2}{\beta^2} \left\{\left(1 - \frac{1}{\beta B}\right) \exp\{-\tfrac{1}{2}(B-\beta)^2\} + \left(1 + \frac{1}{\beta B}\right) \exp\{-\tfrac{1}{2}(B+\beta)^2\}\right\} dB,$$

and when $n_1 = 7$

$$df = \frac{1}{\sqrt{2\pi}} \frac{B^3}{\beta^3} \left\{\left(1 - \frac{3}{\beta B} + \frac{3}{\beta^2 B^2}\right) \exp\{-\tfrac{1}{2}(B-\beta)^2\} - \left(1 + \frac{3}{\beta B} + \frac{3}{\beta^2 B^2}\right) \exp\{-\tfrac{1}{2}(B+\beta)^2\}\right\} dB,$$

In the cases in which n_1 is even, the probability of exceeding a given value B may be written

$$\int_B^\infty 2^{-\frac{1}{2}(n_1-2)} e^{-\frac{1}{2}\beta^2} \sum_{t=0}^{\infty} \frac{\beta^{2t}}{2^{2t} . t ! [\frac{1}{2}(n_1 + 2t - 2)] !} x^{n_1+2t-1} e^{-\frac{1}{2}x^2}\, dx;$$

using the fact that when k is odd

$$\int_B^\infty \frac{x^k}{2^{\frac{1}{2}(k-1)} . [\frac{1}{2}(k-1)] !} e^{-\frac{1}{2}x^2}\, dx = e^{-\frac{1}{2}B^2} \left\{1 + \frac{B^2}{2} + \frac{B^4}{2.4} + \ldots + \frac{B^{k-1}}{2.4 \ldots (k-1)}\right\},$$

the integral becomes

$$\sum_{t=0}^{\infty} e^{-\frac{1}{2}\beta^2} \frac{(\frac{1}{2}\beta^2)^t}{t!} \sum_{u=0}^{t+\frac{1}{2}(n_1-2)} e^{-\frac{1}{2}B^2} \frac{(\frac{1}{2}B^2)^u}{u!},$$

involving only the terms of two Poisson Series with mean values $\frac{1}{2}\beta^2$ and $\frac{1}{2}B^2$. If t and u be regarded as variates distributed independently in two such series,

the probability may be identified with the probability that u should not exceed t by $\frac{1}{2}n_1$, or more.

The distributions developed in this section are limiting forms appropriate to large samples, in which exact account is taken of the positive bias of small observed multiple correlations; they will provide at least an approximate treatment of those cases of great practical importance in which n_2 does not exceed 100, and in which, therefore, the positive bias is prominent for observed values of R which are not small. The fact that sampling errors of the simple correlation coefficient have been successfully represented by a normal distribution by means of the transformation $z = \tanh^{-1} r$, suggests that pending fuller tests than are at present practicable, the transformation

$$\mathrm{B} = \sqrt{n_2}\,\tanh^{-1}\mathrm{R}, \quad \beta = \sqrt{n_2}\,\tanh^{-1}\rho,$$

will supply tests of significance of precision, sufficient for practical purposes, in the important region alluded to.

Table I (table of B) shows the 5 per cent. points of these distributions, for

Table of 5 per cent. points of the distribution of B.

Values of β	Value of n_1						
	1.	2.	3.	4.	5.	6.	7.
0	1·9600	2·4477	2·7955	3·0802	3·3272	3·5485	3·7506
0·2	1·9985	2·4720	2·8140	3·0955	3·3405	3·5602	3·7613
0·4	2·1070	2·5419	2·8680	3·1405	3·3796	3·5951	3·7930
0·6	2·2654	2·6497	2·9533	3·2125	3·4426	3·6517	3·8445
0·8	2·4505	2·7855	3·0640	3·3076	3·5268	3·7278	3·9144
1·0	2·6461	2·9398	3·1941	3·4216	3·6291	3·8210	4·0005
1·2	2·8451	3·1059	3·3386	3·5505	3·7462	3·9289	4·1008
1·4	3·0449	3·2796	3·4935	3·6911	3·8756	4·0491	4·2134
1·6	3·2449	3·4584	3·6561	3·8408	4·0148	4·1796	4·3363
1·8	3·4449	3·6410	3·8246	3·9978	4·1620	4·3184	4·4681
2·0	3·6449	3·8263	3·9976	4·1604	4·3158	4·4645	4·6074
2·2	3·8449	4·0137	4·1743	4·3278	4·4750	4·6166	4·7531
2·4	4·0449	4·2027	4·3539	4·4990	4·6388	4·7738	4·9043
2·6	4·2449	4·3932	4·5359	4·6735	4·8065	4·9353	5·0603
2·8	4·4449	4·5847	4·7199	4·8506	4·9774	5·1006	5·2204
3·0	4·6449	4·7772	4·9055	5·0301	5·1512	5·2691	5·3840
3·2	4·8449	4·9705	5·0926	5·2115	5·3273	5·4404	5·5508
3·4	5·0449	5·1644	5·2809	5·3946	5·5056	5·6142	5·7204
3·6	5·2449	5·3589	5·4703	5·5792	5·6857	5·7901	5·8924
3·8	5·4449	5·5539	5·6606	5·7650	5·8675	5·9679	6·0665
4·0	5·6449	5·7493	5·8516	5·9521	6·0506	6·1475	6·2426
4·2	5·8449	5·9451	6·0434	6·1401	6·2351	6·3285	6·4204
4·4	6·0449	6·1412	6·2359	6·3290	6·4206	6·5109	6·5998
4·6	6·2449	6·3376	6·4288	6·5187	6·6072	6·6945	6·7805
4·8	6·4449	6·5342	6·6223	6·7091	6·7947	6·8792	6·9625
5·0	6·6449	6·7311	6·8162	6·9002	6·9831	7·0649	7·1457

values of β from 0 to 5 and of n_1 from 1 to 7. The values tabulated are the values of B which will be exceeded by chance in 5 per cent. random trials, and which therefore give a presumption that β is really greater than the value postulated. Thus, when $n_1 = 3$, it may be seen at a glance that a value $B = 5{\cdot}7$ indicates that β probably exceeds $3{\cdot}8$.

For a great part of the labour of constructing this Table I am indebted to Mr. A. J. Page, I.C.S., whose assistance in my laboratory while on leave has thus enabled me to press forward with the theoretical investigation of the new distributions.

6. *The Probability Integral.*

For calculations involving finite probabilities of occurrence, including tests whether an observed R is or is not significantly discrepant from a hypothetical ρ, it is not the frequency element but its integral that is required. It is fortunate that the frequency distribution we have found when n_2 is even leads to a probability integral of a tolerably simple form.

The frequency element

$$(1-\rho^2)^{\frac{1}{2}(n_1+n_2)}\,\frac{n_1+n_2-2}{2}\,!\,\frac{(R^2)^{\frac{1}{2}(n_1-2)}}{[\frac{1}{2}(n_1-2)]\,!}$$
$$\times\,\frac{(1-R^2)^{\frac{1}{2}(n_2-2)}}{[\frac{1}{2}(n_2-2)]\,!}\,F\left[\frac{n_1+n_2}{2},\ \frac{n_1+n_2}{2},\ \frac{n_1}{2},\ \rho^2R^2\right]d(R^2);$$

may be written

$$(1-\rho^2)^{\frac{1}{2}(n_1+n_2)}\sum_{t=0}^{\infty}\frac{[\frac{1}{2}(n_1+n_2+2t-2)]\,!^2}{[\frac{1}{2}(n_1+n_2-2)]\,!\,t\,!}\,\rho^{2t}\,\frac{(R^2)^{\frac{1}{2}(n_1+2t-2)}}{[\frac{1}{2}(n_1+2t-2)]\,!}$$
$$\times\,\frac{(1-R^2)^{\frac{1}{2}(n_2-2)}}{[\frac{1}{2}(n_1-2)]\,!}\,d(R^2);$$

but if n_2 is even

$$\int_0^{R^2}\frac{n_1+n_2+2t-2}{2}\,!\,\frac{(R^2)^{\frac{1}{2}(n_1+2t-2)}}{[\frac{1}{2}(n_1+2t-2)]\,!}\cdot\frac{(1-R^2)^{\frac{1}{2}(n_2-2)}}{[\frac{1}{2}(n_2-2)]\,!}\,d(R^2)$$

is

$$R^{n_1+2t}\left\{1+\frac{n_1+2t}{2}(1-R^2)+\frac{(n_1+2t)(n_1+2t+2)}{2\,.\,4}(1-R^2)^2+\ldots\right.$$
$$\left.+\,\frac{(n_1+2t)\ldots(n_1+2t+n_2-4)}{2\,.\,4\,\ldots\,(n_2-2)}\,(1-R^2)^{\frac{1}{2}(n_2-2)}\right\}$$

or

$$\sum_{p=0}^{\frac{1}{2}(n_2-2)}\frac{(1-R^2)^p}{p\,!}\cdot\frac{[\frac{1}{2}(n_1+2t+2p-2)]\,!}{[\frac{1}{2}(n_1+2t-2)]\,!}\,(R^2)^{\frac{1}{2}(n_1+2t)}.$$

Again

$$\sum_{t=0}^{\infty} \frac{[\frac{1}{2}(n_1+n_2+2t-2)]!}{[\frac{1}{2}(n_1+n_2-2)]!\,t!}\rho^{2t}\frac{[\frac{1}{2}(n_1+2t+2p-2)]!}{[\frac{1}{2}(n_1+2t-2)]!}(\mathrm{R}^2)^{\frac{1}{2}(n_1+2t)}$$

is

$$\mathrm{R}^{n_1}\frac{[\frac{1}{2}(n_1+2p-2)]!}{[\frac{1}{2}(n_1-2)]!}\mathrm{F}\left(\frac{n_1+n_2}{2},\ \frac{n_1+2p}{2},\ \frac{n_1}{2},\ \rho^2\mathrm{R}^2\right)$$

or

$$\frac{[\frac{1}{2}(n_1+2p-2)]!}{[\frac{1}{2}(n_1-2)]!}\cdot\frac{\mathrm{R}^{n_1}}{(1-\rho^2\mathrm{R}^2)^{\frac{1}{2}(n_1+n_2+2p)}}\mathrm{F}\left(-p,\ -\frac{n_2}{2},\ \frac{n_1}{2},\ \rho^2\mathrm{R}^2\right),$$

which terminates in $p+1$ terms, and is equivalent to

$$\frac{\frac{1}{2}(n_2)!}{[\frac{1}{2}(n_2-2p)]!}\cdot\frac{\mathrm{R}^{n_1}(\rho^2\mathrm{R}^2)^p}{(1-\rho^2\mathrm{R}^2)^{\frac{1}{2}(n_1+n_2+2p)}}\mathrm{F}\left(-p,\ -\frac{n_1+2p-2}{2},\ \frac{n_1-2p+2}{2},\ \frac{1}{\rho^2\mathrm{R}^2}\right),$$

or

$$\frac{(\frac{1}{2}n_2)!}{[\frac{1}{2}(n_2-2p)]!}\cdot\frac{\mathrm{R}^{n_1}(-)^p}{(1-\rho^2\mathrm{R}^2)^{\frac{1}{2}(n_1+n_2)}}\mathrm{F}\left(-p,\ \frac{n_1+n_2}{2},\ \frac{n_2-2p+2}{2},\ \frac{1}{1-\rho^2\mathrm{R}^2}\right).$$

The probability integral, when n_2 is even, may therefore be written in the forms

$$(1-\rho^2)^{\frac{1}{2}(n_1+n_2)}\mathrm{R}^{n_1}\sum_{p=0}^{\frac{1}{2}(n_2-2)}\frac{[\frac{1}{2}(n_1+2p-2)]!}{[\frac{1}{2}(n_1-2)]!\,p!}\frac{(1-\mathrm{R}^2)^p}{(1-\rho^2\mathrm{R}^2)^{\frac{1}{2}(n_1+n_2+2p)}}\mathrm{F}\left(-p,\ -\frac{n_2}{2},\ \frac{n_1}{2},\ \rho^2\mathrm{R}^2\right),$$

or

$$\left(\frac{1-\rho^2}{1-\rho^2\mathrm{R}^2}\right)^{\frac{1}{2}(n_1+n_2)}\mathrm{R}^{n_1}\sum_{p=0}^{\frac{1}{2}(n_2-2)}(-)^p\frac{(\frac{1}{2}n_2)!}{[\frac{1}{2}(n_2-2p)]!\,p!}(1-\mathrm{R}^2)^p\mathrm{F}\left(-p,\ \frac{n_1+n_2}{2},\ \frac{n_1-2p+2}{2},\ \frac{1}{1-\rho^2\mathrm{R}^2}\right),$$

both of which terminate in $\frac{n_2}{8}(n_2+2)$ elementary terms.

When $n_2=2$, we have the simple probability integral

$$\{(1-\rho^2)/(1-\rho^2\mathrm{R}^2)\}^{\frac{1}{2}(n_1+2)}\mathrm{R}^{n_1};$$

when $n_2=4$, it becomes

$$\left(\frac{1-\rho^2}{1-\rho^2\mathrm{R}^2}\right)^{\frac{1}{2}(n_1+4)}\left\{\frac{n_1+4}{2}\frac{1-\mathrm{R}^2}{1-\rho^2\mathrm{R}^2}-(1-2\mathrm{R}^2)\right\}\mathrm{R}^{n_1},$$

and, when $n_2 = 6$,

$$\left(\frac{1-\rho^2}{1-\rho^2R^2}\right)^{\frac{1}{2}(n_1+6)}\left\{\frac{(n_1+6)(n_1+8)}{2\,.\,4}\left(\frac{1-R^2}{1-\rho^2R^2}\right)^2 - \frac{n_1+6}{2}(2-3R^2)\frac{1-R^2}{1-\rho^2R^2} + (1-3R^2+3R^4)\right\} R^{n_1}.$$

It should be observed that the coefficient of $\{(1-R^2)/(1-\rho^2R^2)\}^p$ is given by

$$\frac{[\frac{1}{2}(n_1+n_2+2p-2)]!}{[\frac{1}{2}(n_1+n_2-2)]!\;p!^2}\frac{d^p}{dx^p}\left\{\frac{(R^2+x)^{\frac{1}{2}n_2}-R^{n_2}}{x}\right\}$$

when $x = -1$.

7. *Extension of the Analysis of Variance.*

The distribution of the simple correlation coefficient, although one of the first sampling distributions to be determined with exactitude*, has hitherto occupied a somewhat isolated position. For all the exact distributions of statistics since discovered have grouped themselves in a single system; they are all amenable to the same technical procedure known as the analysis of variance; and all may be reduced to an equivalent problem of the distribution of the difference of the logarithms of two independent estimates of variance, based respectively upon n_1 and n_2 degrees of freedom.

The distribution of such an estimate s_1^2 derived from n_1 degrees of freedom is given by

$$df = \frac{1}{[\frac{1}{2}(n_1-2)]!}t_1^{\frac{1}{2}(n_1-2)}e^{-t_1}\,dt_1, \text{ where } t_1 = \frac{n_1 s_1^2}{2\sigma^2},$$

and σ is the parameter of which s_1 is the first estimate.

If, now, $t_2 = n_2 s_2^2/2\sigma^2$, and

$$z = \log s_1 - \log s_2,$$

it follows that

$$t_1 = (n_1/n_2)\,e^{2z}\,t_2,$$

and the simultaneous distribution

$$df = \frac{1}{[\frac{1}{2}(n_1-2)]!}t_1^{\frac{1}{2}(n_1-2)}e^{-t_1}\,dt_1 \cdot \frac{1}{[\frac{1}{2}(n_2-2)]!}t_2^{\frac{1}{2}(n_2-2)}e^{-t_2}\,dt_2$$

may be written

$$df = \frac{2}{[\frac{1}{2}(n_1-2)]!\;[\frac{1}{2}(n_2-2)]!}\left(\frac{n_1}{n_2}e^{2z}\right)^{\frac{1}{2}n_1} t_2^{\frac{1}{2}(n_1+n_2-2)}e^{t_2\left(1+\frac{n_1}{n_2}e^{2z}\right)}\,dt_2\,dz\,;$$

* Fisher, 'Biometrika,' vol. 10, p. 507 (1915).

this expression may be integrated with respect to t_2 to yield the distribution of z, in the form

$$df = 2\,\frac{[\frac{1}{2}(n_1+n_2-2)]!}{[\frac{1}{2}(n_1-2)]!\;[\frac{1}{2}(n_2-2)]!}\cdot\frac{n_2^{\frac{1}{2}n_2}\,n_1^{\frac{1}{2}n_1}\,e^{n_1 z}}{(n_2+n_1e^{2z})^{\frac{1}{2}(n_1+n_2)}}\cdot dz,$$

completely independent of the unknown variance.

By the insertion of the appropriate values of n_1 and n_2, including the important bounding values of unity and infinity, the appropriate distribution of z for the analysis of variance is obtained. In the case, for example, of the multiple correlation coefficient drawn from uncorrelated material, n_1 is equated to the number of independent variates, $n_1 + n_2 + 1$ to the sample number, and $2z$ to

$$\log(\mathrm{R}^2/n_1) - \log(\overline{1-\mathrm{R}^2}/n_2).$$

It was from the first obvious that this system was capable without formal modification of extension to the case in which s_1 and s_2 were estimates of two different parameters σ_1 and σ_2; for in such cases we have only to write $\zeta = \log\sigma_1 - \log\sigma_2$, and the distribution found above will be that appropriate to the variate $z - \zeta$.

The new system of distributions found for the multiple correlation coefficient derived from correlated material is not only a generalisation of that previously found* for the simple correlation coefficient, but provides an extension of a different kind from that mentioned above to the analysis of variance. For the limiting distribution found in section 5 (distribution of B) may be interpreted as the distribution of the sum of the squares of n_1 variates normally distributed with equal variance, but not with zero means as in all cases previously discussed.

To show this, let $\mathrm{T} = \frac{1}{2\sigma^2}\sum_{p=1}^{n_1}(x_p - a_p)^2$, in which $x_1, \ldots, x_{n_1}$ are variates distributed independently about zero with common variance σ^2. Let $\xi = \mathrm{S}(ax)/\sigma\mathrm{S}(a^2)$, then ξ will be normally distributed about zero with unit variance, and if we write $\frac{1}{2}\chi^2$ for $\mathrm{T} - \frac{1}{2}(\xi - \sqrt{\mathrm{S}(a^2)}/\sigma)^2$ or $\frac{1}{2}\sum_1^{n_1}(x^2/\sigma^2) - \frac{1}{2}\xi^2$, which is the sum of the squares of $(n_1 - 1)$ quantities independently distributed about zero with unit variance, it appears that the distribution of χ^2 is of the familiar form

$$\frac{1}{[\frac{1}{2}(n_1-3)]!}\,(\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n_1-3)}\,e^{-\frac{1}{2}\chi^2}\,d(\tfrac{1}{2}\chi^2),$$

and is independent of that of ξ, namely $(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}\xi^2}\,d\xi$.

* Fisher, 'Biometrika,' vol. 10, p. 507 (1915).

If, now, β is written for $\sqrt{S(a^2)}/\sigma$ and x for $\xi - \beta$, it follows that

$$\tfrac{1}{2}\chi^2 = T - \tfrac{1}{2}x^2,$$

and as the same value of x^2 is provided by the two values of ξ, $\beta \pm \sqrt{x^2}$, the frequency element required from the distribution of ξ is

$$\frac{1}{\sqrt{2\pi}}\{e^{-\frac{1}{2}(\beta+x)^2} + e^{-\frac{1}{2}(\beta-x)^2}\}\,dx,$$

only positive values of x being now considered. Substituting for χ^2 in terms of T and x, the frequency distribution of the latter two variates will be given by

$$df = \frac{1}{[\frac{1}{2}(n_1-3)]!}(T-\tfrac{1}{2}x^2)^{\frac{1}{2}(n_1-3)}\,dT\,.\,\frac{2}{\sqrt{2\pi}}e^{-\frac{1}{2}\beta^2}e^{-T}\cosh(\beta x)\,.\,dx.$$

For a given value of T, the variate x cannot exceed $\sqrt{2T}$, and the random sampling distribution of T is therefore found by integrating between 0 and $\sqrt{2T}$. Expanding the hyperbolic cosine in powers of x, and integrating term by term, since

$$\int_0^{\sqrt{2T}}(T-\tfrac{1}{2}x^2)^{\frac{1}{2}(n_1-3)}\frac{x^{2p}\beta^{2p}}{(2p)!}\,dx$$
$$= \frac{[\frac{1}{2}(2p-1)]!\,[\frac{1}{2}(n_1-3)]!}{[\frac{1}{2}(n_1+2p-2)]!}\,T^{\frac{1}{2}(n_1+2p-2)}\,2^{\frac{1}{2}(2p-1)}\frac{\beta^{2p}}{(2p)!},$$

we have the distribution of T in the form.

$$df = e^{-\frac{1}{2}\beta^2}e^{-T}\sum_{p=0}^{\infty}\frac{T^{\frac{1}{2}(n_1+2p-2)}\beta^{2p}}{[\frac{1}{2}(n_1+2p-2)]!\,2^p\,.\,p!}\,dT,$$

or

$$df = e^{-\frac{1}{2}\beta^2}\frac{T^{\frac{1}{2}(n_1-2)}}{[\frac{1}{2}(n_1-2)]!}e^{-T}\left\{1+\frac{1}{n_1}(T\beta^2)+\frac{1}{n_1(n_1+2)}\frac{(T\beta^2)^2}{2!}+\ldots\right\}dT,$$

which is the B-distribution of section 5 if T is equated to $\frac{1}{2}B^2$.

This interpretation of the distribution previously obtained adds greatly to its importance, for it is seen to replace the χ^2 distribution of the analysis of variance for cases in which the sum of squares corresponding to n_1 degrees of freedom is derived theoretically for non-central deviations with fixed central displacements. This will be similar to, but not identical with, the case of the n_1 degrees of freedom in multiple correlation in its proper form; for although these are non-central, the displacements will depend on the variation *in the sample* of the independent variates, and this will vary from sample to sample. In many cases, however, such as the dependence of weather upon the position and altitudes of a number of fixed meteorological stations, we are not interested

in the effects of possible variations in the positions of the stations, but solely in the possible variations of the weather at these spots. In fact, the problem of practical importance is often that in which the central displacements are constant, and although it may be urged, rightly enough, that to such cases the purely empirical concept of multiple correlation is not the most appropriate approach, yet it remains true that of the practical applications of multiple correlation methods many are of this kind.

The direct extension of the analysis of variance for non-central squares may be completed by writing

$$df = \frac{1}{[\frac{1}{2}(n_2-2)]!}\, t_2^{\frac{1}{2}(n_2-2)}\, e^{-t}\, dt_2 \quad \text{and} \quad \frac{T}{t_2} = \frac{R^2}{1-R^2} = \frac{n_1}{n_2} e^{2z},$$

then, if, in spite of the caution above, we choose to express our results in terms of R,

$$t_2 = \frac{1-R^2}{R^2} T, \quad dt_2 = -\frac{1-R^2}{R^2} T \frac{d(R^2)}{R^2(1-R^2)}, \quad t_2 + T = \frac{T}{R^2},$$

and the distribution of R is found by integrating with respect to T from 0 to ∞ the expression

$$\frac{1}{[\frac{1}{2}(n_2-2)]!}\, e^{-\frac{1}{2}\beta^2}\, e^{-T/R^2} \left(\frac{1-R^2}{R^2}\right)^{\frac{1}{2}n_2} \frac{dR^2}{R^2(1-R^2)} \sum_{p=0}^{\infty} \frac{T^{\frac{1}{2}(n_1+n_2+2p-2)}\, \beta^{2p}}{[\frac{1}{2}(n_1+2p-2)]!\, 2^p \,.\, p!} dT,$$

a process which yields

$$df = (R^2)^{\frac{1}{2}(n_1-2)} \frac{(1-R^2)^{\frac{1}{2}(n_2-2)}}{[\frac{1}{2}(n_1-2)]!}\, e^{-\frac{1}{2}\beta^2} \sum_{p=0}^{\infty} \frac{[\frac{1}{2}(n_1+n_2+2p-2)]!}{[\frac{1}{2}(n_1+2p-2)]!} \frac{(R^2\beta^2)^p}{2^p \,.\, p!} d(R^2),$$

or

$$df = \frac{[\frac{1}{2}(n_1+n_2-2)]!}{[\frac{1}{2}(n_1-2)]!\,[\frac{1}{2}(n_2-2)]!} (R^2)^{\frac{1}{2}(n_1-2)} (1-R^2)^{\frac{1}{2}(n_2-2)} e^{-\frac{1}{2}\beta^2} \left\{1 + \frac{n_1+n_2}{n_1 \,.\, 1!} \frac{R^2\beta^2}{2} + \frac{(n_1+n_2)(n_1+n_2+2)}{n_1(n_1+2) \,.\, 2!} \left(\frac{R^2\beta^2}{2}\right)^2 + \ldots\right\} d(R^2), \quad (C)$$

a third general distribution of this interesting group.

Although it will not be possible within the limits of this paper to give an account of the properties of the distribution of Type (C), beyond indicating their analogy with those of Type (A), it should not be overlooked that in the problems in which the multiple correlation coefficient is actually employed, distributions of Type (C) will be, owing to the absence or irrelevance of sampling variation in the variances of the independent variates, of at least as frequent occurrence as those of Type (A).

A typical example of the distinction here drawn is provided by the correlation ratio. If corresponding to any value x of the independent variate a number of values n_x of the dependent variate y is observed, then the correlation ratio η^2 of y on x is defined by the relation

$$\frac{\eta^2}{1-\eta^2} = \frac{\mathrm{S}\{n_x(\bar{y}_x - \bar{y})^2\}}{\mathrm{S}(y-\bar{y}_x)^2},$$

in which $\bar{y}_x$ is the mean of y in any array, and $\bar{y}$ is the general mean; the variance in all arrays is supposed equal, and the summation in the numerator is applied to the several arrays, while that in the denominator is applied to the whole of the individual observations. In most practical cases the idea of a sampling distribution of η^2 can only be given a definite meaning by supposing the number n_x in each array to be the same for all samples. In such a case the distribution of η^2 will be that of R^2 in distribution (C), with n_1 equal to one less than the number of arrays, and $n_1 + n_2 + 1$ equal to the total number of observations. If, however, the numbers n_x be regarded as subject to sampling variations, then the distribution (A) may be used, and will be exact, apart from grouping errors, if the expectations of y for the values of x in the sampled population are normally distributed.

Summary.

By an appropriate linear transformation of the independent variates it may be shown that the sampling distribution of the multiple correlation coefficient does not depend on the whole matrix of correlations between these variates, but solely upon the multiple correlation in the population sampled.

The actual distribution (A) may then easily be obtained by similar methods to those by which the distribution of the simple correlation coefficient has been obtained.

The frequency function involves a hypergeometric function of $\rho^2 R^2$ which is a rational function when n_1 and n_2 are both even, algebraic when n_2 only is even, and reducible to circular functions when n_1 and n_2 are both odd.

The case of large samples yields a series of distributions (B) of great interest, involving Bessel functions, which connect the χ^2 distributions with the Gaussian, and are intimately related to a double Poisson summation. Owing to the practical importance of this limiting form a table of its 5 per cent. points is given up to seven independent variates.

When n_2 is even, the probability integral of the general distribution is expressible in finite terms which are developed in section 6.

The (B) distribution of Section 5 replaces the χ^2 distribution in the analysis of variance if the squares summed are non-central. An analysis of variance so extended leads to a third group of distributions (C), closely related to (A), and tending like it to a common limit (B). The distinction between (A) and (C) arises from the fact that in cases proper to the multiple correlation the central displacements will vary from sample to sample owing to variations in the second order moment coefficients of the independent variates, and for such cases (A) is the correct distribution. The type (C), however, is of frequent occurrence owing to the absence or irrelevance of such variation.