FREQUENCY DISTRIBUTION OF THE VALUES OF THE CORRELATION COEFFICIENT IN SAMPLES FROM AN INDEFINITELY LARGE POPULATION.

By R. A. FISHER.

1. My attention was drawn to the problem of the frequency distribution of the correlation coefficient by an article published by Mr H. E. Soper* in 1913. Seeing that the problem might be attacked by means of geometrical ideas, which I had previously found helpful in the consideration of samples, I have examined the two articles by "Student†," upon which Mr Soper's more elaborate work was based, with a view to checking and verifying the conclusions there attained.

"Student," if I do not mistake his intention, desiring primarily to obtain a just estimate of the accuracy to be ascribed to the mean of a small sample, found it necessary to allow for the fact that the mean square error of such a sample is not generally equal to the standard deviation of the normal population from which it is drawn. He was led, in fact, to study the frequency distribution of the mean square error. He calculated algebraically the first four moments of this frequency curve, both about the zero point, and about its mean, observed a simple law to connect the successive moments, and discovered a frequency curve, which fitted his moments, and gave the required law.

Thus if $x_1, x_2, \ldots x_n$ are the members of a sample,

$$n\bar{x} = x_1 + x_2 + \ldots + x_n,$$

and

$$n\mu^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2,$$

the frequency with which the mean square error lies in the range $d\mu$ is proportional to

$$\mu^{n-2} e^{-\frac{n\mu^2}{2\sigma^2}} d\mu.$$

This result, although arrived at by empirical methods, was established almost beyond reasonable doubt in the first of "Student's" papers. It is, however, of interest to notice that the form establishes itself instantly, when the distribution of the sample is viewed geometrically.

* *Biometrika*, Vol. IX. p. 91. † *Ibid.* Vol. VI. pp. 1 and 302.

In the second of these two papers the more difficult problem of the frequency distribution of the correlation coefficient is attempted. For samples of 2 the frequency distribution between the only two possible values -1 and $+1$ was determined by Sheppard's theorem to be in the ratio $\frac{\pi}{2}+\sin^{-1}\rho : \frac{\pi}{2}-\sin^{-1}\rho$, where ρ is the correlation of the population. Besides this theoretical result, "Student" appeals only to experimental data. From these he derives an empirical form for the distribution when $\rho = 0$, and makes several valuable suggestions. It has been the greatest pleasure and interest to myself to observe with what accuracy "Student's" insight has led him to the right conclusions. The form when $\rho = 0$ is absolutely correct, and as a further instance I may quote the remark* "I have dealt with the cases of samples of 2 at some length, because it is possible that this limiting value of the distribution, with its mean of $\frac{2}{\pi}\sin^{-1}\rho$ and its second moment coefficient of $1-\left(\frac{2}{\pi}\sin^{-1}\rho\right)^2$, may furnish a clue to the distribution when n is greater than 2." As a matter of fact it is just these quantities with which we shall be concerned.

To Mr Soper's laborious and intricate paper I cannot hope to do justice. I have been able to establish the substantial accuracy and value of his approximations. It is one of the advantages of approaching a problem from opposite standpoints that Mr Soper's forms are most accurate for those larger values of n, where the exact formulae become most complicated.

2. The problem of the frequency distribution of the correlation coefficient r, derived from a sample of n pairs, taken at random from an infinite population, may be solved, when that population can be represented by a normal surface, with the aid of certain very general conceptions derived from the geometry of n dimensional space. In this paper the general form will first be demonstrated, and for a few important cases some of the successive moments will be derived. Incidentally it will be of interest to compare the exact form with Mr Soper's approximation, and with reference to the experimental data supplied by "Student."

If the frequency distribution of the population be specified by the form

$$df = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{1-\rho^2}\left\{\frac{(x-m_1)^2}{2\sigma_1^2}-\frac{2\rho\,(x-m_1)\,(y-m_2)}{2\sigma_1\sigma_2}+\frac{(y-m_2)^2}{2\sigma_2^2}\right\}} dx\,dy,$$

where df is the chance that any observation should fall into the range $dx\,dy$, then the chance that n pairs should fall within their specified elements is

$$\frac{1}{(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^n}\, e^{-\frac{1}{1-\rho^2}\sum_1^n\left\{\frac{(x-m_1)^2}{2\sigma_1^2}-\frac{2\rho\,(x-m_1)\,(y-m_2)}{2\sigma_1\sigma_2}+\frac{(y-m_2)^2}{2\sigma_2^2}\right\}} dx_1dy_1\ldots dx_ndy_n \ldots (\mathrm{I}),$$

and this we interpret as a simple density distribution in $2n$ dimensions.

* *Biometrika*, Vol. VI. p. 304.

For the variables x and y it is now necessary to substitute the statistical derivatives determined by the equations

$$n\bar{x} = \sum_1^n (x), \qquad n\bar{y} = \sum_1^n (y),$$

$$n\mu_1^2 = \sum_1^n (x-\bar{x})^2, \quad n\mu_2^2 = \sum_1^n (y-\bar{y})^2,$$

$$nr\mu_1\mu_2 = \sum_1^n (x-\bar{x})(y-\bar{y}),$$

and it is evident that the only difficulty lies in the expression of an element of volume in $2n$ dimensional space in terms of these derivatives.

The five quantities above defined have, in fact, an exceedingly beautiful interpretation in generalised space, which we may now examine.

3. Considering first the space of n dimensions in which the variations of x are represented, the mean and mean square error of n observations are determined by the relations of P, the point representing the n observations, to the line

$$x_1 = x_2 = x_3 = \ldots = x_n,$$

for the perpendicular PM drawn from P upon this line will lie in the region

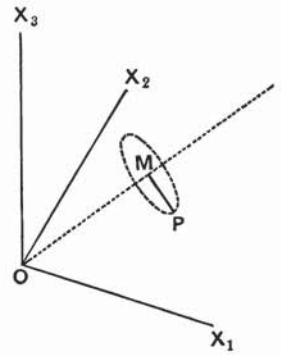
$$x_1 + x_2 + \ldots + x_n = n\bar{x},$$

and will meet it at the point M, where

$$x_1 = \bar{x}, \quad x_2 = \bar{x}, \quad \ldots x_n = \bar{x};$$

further, since, $$PM^2 = (x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2,$$

the length of PM is $\mu_1\sqrt{n}$.



An element of volume in this n dimensional space may now without difficulty be specified in terms of $\bar{x}$ and μ_1; for, given $\bar{x}$ and μ_1, P must lie on a sphere in $n-1$ dimensions, lying at right angles to the line OM, and the element of volume is

$$C\mu_1^{n-2} d\mu_1 d\bar{x},$$

where C is some constant, which need not be determined.

The point in $2n$ dimensional space which is represented by the n pairs of observations must be such that its projection on the n dimensional space, in which x is represented, lies upon a certain sphere of radius $\mu_1\sqrt{n}$, and on the space in which y is represented, upon another sphere of radius $\mu_2\sqrt{n}$, and now, when we come to the interpretation of r, we must observe that to each point on the first sphere there corresponds a certain point on the second sphere, to which it bears the relation

$$\frac{x_1-\bar{x}}{y_1-\bar{y}}=\frac{x_2-\bar{x}}{y_2-\bar{y}}=\ldots=\frac{x_n-\bar{x}}{y_n-\bar{y}}.$$

In general this relation does not hold for the n pairs of observations, and the two projections will not fall at corresponding points on the two spheres. If now one of the spheres be turned round so as to occupy the same space as the other, and so that the lines upon which x_1 and y_1, and the other pairs of coordinates, are measured, coincide, then corresponding points will lie on the same radii, and the correlation coefficient r measures the cosine of the angle between the radii to the two points specified by the observations.

Taking one of the projections as fixed at any point on the sphere of radius μ_2, the region for which r lies in the range dr, is a zone, on the other sphere in $n-1$ dimensions, of radius $\mu_1\sqrt{n}\sqrt{1-r^2}$, and of width $\mu_1\sqrt{n}\,dr/\sqrt{1-r^2}$, and therefore having a volume proportional to $\mu_1^{n-2}(1-r^2)^{\frac{n-4}{2}}\,dr$.

4. We may now turn to the direct simplification of the expression (I), at each stage discarding any factors which do not involve r.

$$e^{-\frac{1}{1-\rho^2}\sum_1^n\left\{\frac{(x-m_1)^2}{2\sigma_1^2}-\frac{2\rho\,(x-m_1)\,(y-m_2)}{2\sigma_1\sigma_2}+\frac{(y-m_2)^2}{2\sigma_2^2}\right\}}dx_1dy_1dx_2dy_2\ldots dx_ndy_n$$

may be reduced to

$$e^{-\frac{n}{1-\rho^2}\left\{\frac{(\bar{x}-m_1)^2+\mu_1^2}{2\sigma_1^2}-\frac{2\rho\{r\mu_1\mu_2+(\bar{x}-m_1)(\bar{y}-m_2)\}}{2\sigma_1\sigma_2}+\frac{(\bar{y}-m_2)^2+\mu_2^2}{2\sigma_2^2}\right\}}$$

$$d\bar{x}d\bar{y}\,\mu_1^{n-2}d\mu_1\mu_2^{n-2}d\mu_2(1-r^2)^{\frac{n-4}{2}}\,dr,$$

or to

$$e^{-\frac{n}{1-\rho^2}\left\{\frac{\mu_1^2}{2\sigma_1^2}-\frac{2\rho r\mu_1\mu_2}{2\sigma_1\sigma_2}+\frac{\mu_2^2}{2\sigma_2^2}\right\}}\mu_1^{n-2}\mu_2^{n-2}(1-r^2)^{\frac{n-4}{2}}\,d\mu_1d\mu_2dr.$$

In order to integrate this expression from 0 to ∞, with respect to μ_1 and μ_2, let

$$\zeta=\frac{\mu_1\mu_2}{\sigma_1\sigma_2},\qquad e^z=\frac{\mu_1\sigma_2}{\mu_2\sigma_1},$$

and we have

$$\int_{-\infty}^{\infty}dz\int_0^{\infty}\zeta^{n-2}d\zeta\,.\,e^{-\frac{n}{1-\rho^2}(\cosh z-\rho r)\zeta}\,.\,(1-r^2)^{\frac{n-4}{2}}\,dr,$$

or

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

which, on substituting $\cos\theta$ for $-\rho r$, may be expressed in terms of a Legendre function in the form

$$(i\operatorname{cosec}\theta)^{n-1}\,Q_{n-2}(i\cot\theta).(1-r^2)^{\frac{n-4}{2}}\,dr \quad \text{.....(II).}$$

Again
$$\int_0^\infty \frac{dz}{\cosh z+\cos\theta} = \frac{\theta}{\sin\theta},$$

so that
$$\int_0^\infty \frac{dz}{(\cosh z+\cos\theta)^{n-1}} = \frac{1}{\underline{n-2}}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-2}\frac{\theta}{\sin\theta},$$

and since this is a function of ρr only, we may express the frequency distribution by the convenient expression

$$(1-r^2)^{\frac{n-4}{2}}\frac{\partial^{n-2}}{\partial r^{n-2}}\left(\frac{\theta}{\sin\theta}\right)dr.$$

Professor Pearson has shown that this last result can be obtained directly from Sheppard's theorem* that

$$\frac{1}{2\pi\Sigma_1\Sigma_2\sqrt{1-R^2}}\int_0^\infty\int_0^\infty e^{-\frac{1}{2(1-R^2)}\left(\frac{\mu_1^2}{\Sigma_1^2}-\frac{2R\mu_1\mu_2}{\Sigma_1\Sigma_2}+\frac{\mu_2^2}{\Sigma_2^2}\right)}d\mu_1 d\mu_2 = \frac{\cos^{-1}(-R)}{2\pi};$$

making the substitutions

$$\frac{1}{(1-R^2)\Sigma_1^2} = \frac{n}{(1-\rho^2)\sigma_1^2},$$

$$\frac{1}{(1-R^2)\Sigma_2^2} = \frac{n}{(1-\rho^2)\sigma_2^2},$$

$$\frac{R}{(1-R^2)\Sigma_1\Sigma_2} = \frac{nr\rho}{(1-\rho^2)\sigma_1\sigma_2},$$

which give
$$R=\rho r$$

and
$$\cos^{-1}(-R)=\theta,$$

we obtain

$$\frac{n}{\sigma_1\sigma_2(1-\rho^2)}\int_0^\infty\int_0^\infty e^{-\frac{n}{2(1-\rho^2)}\left(\frac{\mu_1^2}{\sigma_1^2}-\frac{2\rho r\mu_1\mu_2}{\sigma_1\sigma_2}+\frac{\mu_2^2}{\sigma_2^2}\right)}d\mu_1 d\mu_2 = \frac{\theta}{\sin\theta}$$

and hence differentiating $(n-2)$ times with respect to r, the required expression is obtained.

5. The form which we have now obtained may be applied without difficulty to all small even values of n, and in such cases is peculiarly suitable for the calculation of moments.

When $n=2$ the ordinate of the curve, with abscissa r, is

$$\frac{\theta}{(1-r^2)\sin\theta},$$

which becomes hyperbolic in the neighbourhoods of -1 and $+1$. The value

* *Phil. Trans.* Vol. 192, A, p. 141.

of r is, therefore, as we know, either -1 or $+1$, and the proportion, in which these occur, depends upon ρ. The ratio of the infinite areas included with the asymptotes of the above curve is

$$\frac{\cos^{-1}\rho}{\cos^{-1}(-\rho)},$$

so that the mean value of a number of observations is $\dfrac{\sin^{-1}\rho}{\frac{\pi}{2}}$.

When $n=4$ there is still no approach to normality, the curve takes the form

$$\frac{1}{\sin^3\theta}(\theta - 3\cot\theta + 3\theta\cot^2\theta),$$

which, when r is positive, increases regularly from its value of $\frac{4}{15}$ when $\theta = 0$, to infinity, to which it approaches as θ approaches π. Unless ρ is actually equal to 1, in which case r is also 1 of necessity, the curve has finite ordinates at both extremes. For calculating the number of values which should fall within any given range, the integral, $\dfrac{1}{\sin^2\theta}(1-\theta\cot\theta)$, may be directly tabulated, as has been done in forming the accompanying table of "Student's" observations, and the corresponding expectations. The values given by Mr Soper's formula are apposed for comparison.

Table for comparison with p. 114, *Biometrika, Vol. IX.*

r	Calculated frequency m	Observed	Difference e	$\frac{e^2}{m}$	H.E. Soper's approximation	Difference e	$\frac{e^2}{m}$
·905—1	202·1	175·5	−15·0	·69	230·3	−17·2	·90
·805—·905	124·9	136·5			98·9		
·705—·805	88·7	84	− 3·8	·09	72·1	+20·3	3·18
·605—	65·1	66			57·6		
·505—	49·9	55	+12·3	1·73	48·0	+11·8	1·58
·405—	37·8	45			40·2		
·305—	30·6	24·5	− 6·4	·74	34·3	−15·0	3·52
·205—	24·8	24·5			29·7		
·105—	20·5	19	−11·6	3·58	25·6	−21·6	9·80
·005—	17·1	7			22·0		
$\bar{1}$·905—	14·5	22	+ 7·1	1·87	18·8	− ·8	0·02
$\bar{1}$·805—	12·4	12			16·0		
$\bar{1}$·705—	10·7	13	− 4·0	·80	13·5	− 8·7	3·06
$\bar{1}$·605—	9·3	3			11·2		
$\bar{1}$·505—	8·1	12	+12·7	10·54	9·0	+12·1	9·21
$\bar{1}$·405—	7·2	16			6·9		
$\bar{1}$·305—	6·3	7	+ 5·1	2·19	5·1	+ 8·6	8·80
$\bar{1}$·205—	5·6	10			3·3		
$\bar{1}$·105—	5·1	4	+ 3·6	1·38	1·9	+10·5	44·10
$\bar{1}$—$\bar{1}$·105	4·3	9			·6		
—	—	745	—	23·61	—	—	84·17

6. The direct process of integration by parts applied to such expressions as

$$\int_{-1}^{+1}(1-r^2)^{\frac{n-4}{2}}\frac{\partial^{n-1}}{\partial r^{n-1}}\frac{\theta^2}{2}\,dr \quad \text{and} \quad \int_{-1}^{+1}(1-r^2)^{\frac{n-4}{2}}\,r\,\frac{\partial^{n-1}}{\partial r^{n-1}}\frac{\theta^2}{2}\,dr,$$

when n is even, merely introduces the sums and differences of the terms $\frac{\partial^p}{\partial r^p}\frac{\theta^2}{2}$ at the extremes, where r is -1 or $+1$, with coefficients which are, in any particular case, easily calculable.

Thus, n being 6,

$$\int_{-1}^{+1}(1-r^2)\frac{\partial^5}{\partial r^5}\frac{\theta^2}{2}\,dr=\left[(1-r^2)\frac{\partial^4}{\partial r^4}\frac{\theta^2}{2}\right]_{-1}^{+1}+\left[2r\frac{\partial^3}{\partial r^3}\frac{\theta^2}{2}\right]_{-1}^{+1}-\left[2\frac{\partial^2}{\partial r^2}\frac{\theta^2}{2}\right]_{-1}^{+1}$$

$$=2\times\text{the sum of the extreme values of }\frac{\rho^3}{\sin^3\theta}(\theta-3\cot\theta+3\theta\cot^2\theta)$$

$$-2\times\text{the difference of the extreme values of }\frac{\rho^2}{\sin^2\theta}(1-\theta\cot\theta).$$

If $\rho=\sin\alpha$, so that the extreme values of θ are $\frac{\pi}{2}-\alpha$ and $\frac{\pi}{2}+\alpha$, the sums and differences may readily be expressed in terms of α, and the first few may here be tabulated: the table has been carried back as far as is necessary for the calculation of the fourth moment.

	sum	difference
$\frac{\sin^2\theta}{4\rho^2}\left\{\frac{7+2\theta^2}{4}-3\theta\cot\theta-\frac{7-6\theta^2}{4}\cot^2\theta\right\}$	—	$\frac{\pi\cot^2\theta}{4}(\alpha+3\tan\alpha+3\alpha\tan^2\alpha)$
$\frac{\sin\theta}{\rho}\left\{\theta+\left(1-\frac{\theta^2}{2}\right)\cot\theta\right\}$	$\pi\cot\alpha\,(1+\alpha\tan\alpha)$	$\cot\alpha\left\{2\alpha-2\tan\alpha+\left(\frac{\pi^2}{4}+\alpha^2\right)\tan\alpha\right\}$
$\frac{\theta^2}{2}$	$\frac{\pi^2}{4}+\alpha^2$	$\pi\alpha$
$\frac{\rho}{\sin\theta}\cdot\theta$	$\pi\tan\alpha$	$2\alpha\tan\alpha$
$\frac{\rho^2}{\sin^2\theta}(1-\theta\cot\theta)$	$2\tan^2\alpha\,(1+\alpha\tan\alpha)$	$\pi\tan^3\alpha$
$\frac{\rho^3}{\sin^3\alpha}(\theta-3\cot\theta+3\theta\cot^2\theta)$	$\pi\tan^2\alpha\,(1+3\tan^2\alpha)$	$2\tan^3\alpha\,(\alpha+3\tan\alpha+3\alpha\tan^2\alpha)$
$\frac{\rho^4}{\sin^4\alpha}(4-9\theta\cot\theta+15\cot^2\theta-15\theta\cot^3\theta)$	$2\tan^4\alpha(4+9\alpha\tan\alpha+15\tan^2\alpha+15\alpha\tan^3\alpha)$	$\pi\tan^4\alpha\,(9\tan\alpha+15\tan^3\alpha)$

There are here two natural series, which appear alternately as sums and differences; the simpler, which may be expressed in the form

$$\frac{\pi}{2}\sin^p\alpha\left(\frac{\partial}{\cos\alpha\partial\alpha}\right)^p\alpha,$$

is essentially a series of Legendre functions of the first kind; and may be expressed as

$$\frac{\pi}{2} \cdot \tan^p \alpha \, \frac{\lfloor p-1}{i^{p-1}} P_{p-1}(i \tan \alpha);$$

and it is these only which occur in the evaluation of the even moments.

7. It is, however, desirable to obtain general expressions for these integrals in terms of n and ρ, and to evaluate them when n is odd.

For this purpose let us introduce a quantity ϕ, such that

$$\cos \phi = \cos \theta - k,$$

then, when k is sufficiently small, we may expand ϕ^2 by Taylor's theorem, so that

$$\frac{\phi^2}{2} = \frac{\theta^2}{2} + k \frac{\partial}{\sin \theta \partial \theta} \frac{\theta^2}{2} + \frac{k^2}{\lfloor 2} \left(\frac{\partial}{\sin \theta \partial \theta} \right)^2 \frac{\theta^2}{2} + \ldots.$$

Now let $$k = \rho h \sqrt{1-r^2},$$

then $$\frac{\phi^2}{2} = \frac{\theta^2}{2} + \rho h \sqrt{1-r^2} \frac{\partial}{\sin \theta \partial \theta} \frac{\theta^2}{2} + \frac{\rho^2 h^2 (1-r^2)}{\lfloor 2} \left(\frac{\partial}{\sin \theta \partial \theta} \right)^2 \frac{\theta^2}{2} + \ldots,$$

and differentiating twice with respect to h

$$\rho^2 (1-r^2) \left(\frac{\partial}{\sin \phi \partial \phi} \right)^2 \frac{\phi^2}{2} = \rho^2 (1-r^2) \left(\frac{\partial}{\sin \theta \partial \theta} \right)^2 \frac{\theta^2}{2} + h \rho^3 (1-r^2)^{\frac{3}{2}} \left(\frac{\partial}{\sin \theta \partial \theta} \right)^3 \frac{\theta^2}{2} + \ldots,$$

whence, dividing by $(1-r^2)^{\frac{3}{2}}$, we obtain

$$\frac{\rho^2}{\sqrt{1-r^2}} \left(\frac{\partial}{\sin \phi \partial \phi} \right)^2 \frac{\phi^2}{2} = \frac{\rho^2}{(1-r^2)^{\frac{1}{2}}} \left(\frac{\partial}{\sin \theta \partial \theta} \right)^2 \frac{\theta^2}{2} + h \rho^3 \left(\frac{\partial}{\sin \theta \partial \theta} \right)^3 \frac{\theta^2}{2} + \frac{\rho^4 h^2}{\lfloor 2} (1-r^2)^{\frac{1}{2}} \left(\frac{\partial}{\sin \theta \partial \theta} \right)^4 \frac{\theta^2}{2} + \ldots,$$

so that $$\int_{-1}^{+1} r^p (1-r^2)^{\frac{n-4}{2}} \frac{\partial^{n-1}}{\partial r^{n-1}} \frac{\theta^2}{2} \, dr$$

may be obtained by multiplying by $\lfloor n-3$ the coefficient of h^{n-3} in

$$\rho^2 \int_{-1}^{+1} \frac{r^p \, dr}{\sqrt{1-r^2}} \cdot \frac{1 - \phi \cot \phi}{\sin^2 \phi},$$

when $\cos \phi = \cos \theta - \rho h \sqrt{1-r^2} = -\rho (r + h \sqrt{1-r^2})$.

Our object might equally be achieved by the evaluation of the integral

$$\rho \int_{-1}^{+1} \frac{r^p \, dr}{1-r^2} \left(\frac{\phi}{\sin \phi} - \frac{\theta}{\sin \theta} \right).$$

The quantity ϕ is determined by the equation

$$\cos \phi = \cos \theta - \rho h \sqrt{1-r^2},$$

that is $$\cos \phi = -\rho (r + h \sqrt{1-r^2}).$$

If now $r = \sin\beta,$

$h = \tan\epsilon,$

then $\cos\theta = -\rho\sin\beta,$

$$\cos\phi = -\rho\sqrt{1+h^2}\sin(\beta+\epsilon) = -\rho\sqrt{1+h^2}\sin\beta',$$

and as r passes from -1 to $+1$,

β passes from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$,

θ from $\frac{\pi}{2}-\alpha$ to $\frac{\pi}{2}+\alpha$,

β' from $-\frac{\pi}{2}+\epsilon$ to $\frac{\pi}{2}$ and thence to $\frac{\pi}{2}+\epsilon$,

and ϕ from $\frac{\pi}{2}-\alpha$ to $\frac{\pi}{2}+\alpha'$ and thence back to $\frac{\pi}{2}+\alpha$,

where $\sin\alpha' = \rho\sqrt{1+h^2}$, ϕ oscillates in the same manner as θ, with a somewhat greater amplitude, and slightly in advance in respect of phase.

The expression $$\rho^2\int_{-1}^{+1}\frac{1-\phi\cot\phi}{\sin^2\phi}\frac{dr}{\sqrt{1-r^2}}$$

may now be reduced to

$$\rho^2\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}\frac{1-\phi\cot\phi}{\sin^2\phi}d\beta = \rho^2\int_{-\frac{\pi}{2}+\epsilon}^{\frac{\pi}{2}+\epsilon}\left(\frac{1}{1-\sin^2\alpha'\sin^2\beta'}+\frac{\phi\sin\alpha'\sin\beta'}{(1-\sin^2\alpha'\sin^2\beta')^{\frac{3}{2}}}\right)d\beta'$$

$$= \rho^2\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}\frac{d\beta'}{1-\sin^2\alpha'\sin^2\beta'} + \pi\rho^2\int_{+\frac{\pi}{2}}^{+\frac{\pi}{2}+\epsilon}\frac{\sin\alpha'\sin\beta'\,d\beta'}{(1-\sin^2\alpha'\sin^2\beta')^{\frac{3}{2}}}$$

$$+\rho^2\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\frac{(\phi)\sin\alpha'\sin\beta'\,d\beta'}{(1-\sin^2\alpha'\sin^2\beta')^{\frac{3}{2}}}$$

$$= \frac{\rho^2\pi}{\cos\alpha'} + \frac{\pi\rho^2\sin\alpha'}{\cos^2\alpha'}\left(\frac{\sin\epsilon}{\cos\alpha}\right) + \frac{\pi\rho^2}{\cos^2\alpha'}(1-\cos\alpha')$$

$$= \frac{\rho^2\pi}{\cos^2\alpha'}\left(1+\frac{\sin\alpha\tan\epsilon}{\cos\alpha}\right),$$

but $$\cos^2\alpha' = 1-\rho^2(1+h^2) = \cos^2\alpha - \sin^2\alpha\tan^2\epsilon,$$

so that $$\rho^2\int_{-1}^{+1}\frac{1-\phi\cot\phi}{\sin^2\phi}\frac{dr}{\sqrt{1-r^2}} = \frac{\pi\tan^2\alpha}{1-h\tan\alpha}$$

From this evaluation we deduce the general form

$$\int_{-1}^{+1}(1-r^2)^{\frac{n-4}{2}}\frac{\partial^{n-1}}{\partial r^{n-1}}\frac{\theta^2}{2}\,dr = \underline{|n-3}\,\pi\tan^{n-1}\alpha \quad \text{.....(III)}$$

The absolute frequency df, with which r falls in the range dr, is therefore

$$\frac{(1-\rho^2)^{\frac{n-1}{2}}}{\pi \underline{|n-3}}(1-r^2)^{\frac{n-4}{2}}\left(\frac{\partial}{\sin\theta\partial\theta}\right)^{n-2}\frac{\theta}{\sin\theta}\,dr.$$

8. I do not see how to integrate the other expressions of the type

$$\rho^2\int_{-1}^{+1}\frac{1-\phi\cot\phi}{\sin^2\phi}\frac{r^p\,dr}{\sqrt{1-r^2}},$$

although a form could probably be obtained when p is even. The general expression for the second moment may, however, be deduced by means of a reduction formula.

By a process of integration by parts it appears that, if we write

$$\int_{-1}^{+1}(1-r^2)^{\frac{n-4}{2}}r^p\frac{\partial^{n-1}}{\partial r^{n-1}}\frac{\theta^2}{2}\,dr = I_{n.p},$$

then
$$I_{n+2.2} = I_{n+2.0} + nI_{n.0} - n(n-1)I_{n.2},$$

and since
$$I_{4.2} = 2\pi\left(\frac{\tan^3\alpha}{2} - \tan\alpha + \alpha\right),$$

we may obtain successively

$$I_{6.2} = 24\pi\left(\frac{\tan^5\alpha}{4} - \frac{\tan^3\alpha}{3} + \tan\alpha - \alpha\right),$$

$$I_{8.2} = 720\pi\left(\frac{\tan^7\alpha}{6} - \frac{\tan^5\alpha}{5} + \frac{\tan^3\alpha}{3} - \tan\alpha + \alpha\right),$$

and so on, yielding, when n is even, the expression

$$I_{n.2} = I_{n.0} - \pi\,\underline{|n-2}\int_0^\alpha \tan^{n-2}x\,dx,$$

a form which may well hold when n is odd.

The above expressions are useful in tabulating the numerical values of the second moment, $\bar{r}^2+\sigma^2$, of the unit curve, which may easily be calculated in succession for different values of n when $\tan^2\alpha$ is taken to have some simple value.

9. Before leaving this aspect of the subject it is worth while to give a more detailed examination of the mean of the frequency curves of r when $n=4$.

Two formulae are arrived at by Mr Soper, which are equivalent approximations of the second degree

I. $$\bar{r} = \rho\left[1 - \frac{1-\rho^2}{2n}\left\{1 + \frac{3}{4n}(1+3\rho^2)\right\}\right] = \rho\left[1 - \frac{1-\rho^2}{8}\left\{1 + \frac{3}{16}(1+3\rho^2)\right\}\right],$$

II. $$\bar{r} = \rho\left[1 - \frac{1-\rho^2}{2(n-1)}\left\{1 - \frac{1}{4(n-1)}(1-9\rho^2)\right\}\right] = \rho\left[1 - \frac{1-\rho^2}{6}\left\{1 - \frac{1}{12}(1-9\rho^2)\right\}\right],$$

and these we shall compare with the form

III. $$\bar{r} = \frac{2}{\pi}(\alpha + \cot\alpha - \alpha\cot^2\alpha),$$

ρ	·1000	·2000	·3000	·4000	·5000	·6000	·7000	·8000	·9000	·9500
I	·0853	·1710	·2578	·3463	·4377	·5333	·6347	·7443	·8649	·9304
II	·0847	·1697	·2555	·3419	·4310	·5241	·6236	·7330	·8566	·9254
III	·0850	·1704	·2570	·3451	·4360	·5301	·6290	·7357	·8540	·9209

It will be observed that the approximations lie on either side of the exact value over the greater part of the range, and that the error of the first approximation increases up to the value when $\rho = ·9$. The second formula gives the correct value somewhere between ·8 and ·9, and is thereafter too large.

For the particular case $\rho = ·6608$,

I find (formula III) $\bar{r} = ·5897$, nearly the maximum difference from ρ,

Mr Soper gives (p. 109) the value ·5933

and the experimental data ·5609.

The two theoretical values are much nearer to each other than either is to the experimental value. On the whole, it is obvious that even in this unfavourable case Mr Soper's formulae possess remarkable accuracy.

10. The use of the correlation coefficient r as independent variable of these frequency curves is in some respects highly unsatisfactory. For high values of r the curve becomes extremely distorted and cramped, and although this very cramping forces the mean $\bar{r}$ to approach ρ, the difference compared with $1-\rho$ becomes inordinately great. Even for high values of n, the distortion in this region becomes extreme, and since at the same time the curve rapidly changes its shape, the values of the mean and standard deviation cease to have any very useful meaning. It would appear essential in order to draw just conclusions from an observed high value of the correlation coefficient, say ·99, that the frequency curves should be reasonably constant in form.

The previous paragraphs suggest that more natural variables for the treatment of our formulae are afforded by the transformations

$$t = \tan\beta = \frac{r}{\sqrt{1-r^2}},$$

$$\tau = \tan\alpha = \frac{\rho}{\sqrt{1-\rho^2}}.$$

The expression for the frequency curve (II)

$$(1-r^2)^{\frac{n-4}{2}}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\,dr$$

now becomes

$$\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\,\frac{dt}{(1+t^2)^{\frac{n-1}{2}}}$$

and the range of the curve is extended from $-\infty$ to $+\infty$.

It is interesting that in the important case, $r=0$, the frequency reduces to $\dfrac{dt}{(1+t^2)^{\frac{n-1}{2}}}$ and the curves are identical with those found by "Student" for z, the probability integral of which he has tabulated in his first paper.

11. The moments of these curves are obtained by the evaluation of the expressions

$$\int_{-\infty}^{\infty}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\,\frac{dt}{(1+t^2)^{\frac{n-1}{2}}},\quad \int_{-\infty}^{\infty}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\,\frac{t\,dt}{(1+t^2)^{\frac{n-1}{2}}},$$

and so on; of these the first is known already (III) to have the value

$$\frac{\pi\,\lfloor n-3}{(1-\rho^2)^{\frac{n-1}{2}}}$$

and the others may be obtained in succession, for

$$I_{n.p}=\int_{-\infty}^{\infty}\frac{\partial^{n-1}}{(\sin\theta\,\partial\theta)^{n-1}}\frac{\theta^2}{2}\,\frac{t^p\,dt}{(1+t^2)^{\frac{n-1}{2}}}=\frac{\partial^{n-1}}{\partial\rho^{n-1}}\int_{-\infty}^{\infty}\frac{1}{r^{n-1}}\,\frac{\theta^2}{2}\,\frac{t^p\,dt}{(1+t^2)^{\frac{n-1}{2}}}$$

$$=\frac{\partial^{n-1}}{\partial\rho^{n-1}}\int_{-\infty}^{\infty}\frac{\theta^2}{2}\cdot\frac{dt}{t^{n-1-p}}=\frac{\partial^p}{\partial\rho^p}\,I_{n-p.0},$$

so that the first moment

$$\int_{-\infty}^{\infty}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\cdot\frac{t\,dt}{(1+t^2)^{\frac{n-1}{2}}}=\frac{\partial}{\partial\rho}\cdot\frac{\pi\,\lfloor n-4}{(1-\rho^2)^{\frac{n-2}{2}}}=\frac{\pi\,\lfloor n-4\,(n-2)\,\rho}{(1-\rho^2)^{\frac{n-4}{2}}};$$

hence

$$\bar{t}=\frac{n-2}{n-3}\,\frac{\rho}{\sqrt{1-\rho^2}}=\frac{n-2}{n-3}\,\tau.$$

The mean, therefore, is greater than the true value τ by a constant fraction of its value. And this fraction decreases in the simplest possible manner as n increases.

In the same way, we may evaluate the second moment,

$$\bar{t}^2+\sigma^2=\frac{1}{n-4}\,\{1+(n-1)\,\tau^2\}$$

and

$$\sigma^2=\frac{1}{n-4}\left\{1+\tau^2+\frac{(n-2)}{(n-3)^2}\,\tau^2\right\};$$

the third moment

$$\sqrt{\beta_1}\sigma^3=\frac{(n-2)\,\tau}{(n-3)\,(n-4)\,(n-5)}\left\{3\,(1+\tau^2)+\frac{2\tau^2\,(n-1)}{(n-3)^2}\right\},$$

and the fourth moment

$$\beta_2\sigma^4=\frac{3}{(n-4)(n-6)}\left\{(1+\tau^2)^2+\frac{6(n-2)\tau^2}{(n-3)(n-5)}(1+\tau^2)+\frac{6(n-2)(3n^2-11n+12)\tau^4}{(n-3)^4(n-5)}\right\}.$$

For high values of n, all but the first terms tend to vanish; β_1 tends to vary as ρ^2, and β_2 tends to become independent of ρ. In effect for high values of τ, where ρ^2 is nearly equal to unity, the form of the curve is nearly constant, but the skewness measured by β_1 decreases to zero at the origin, and changes its sense, when τ and ρ change their sign.

Tables are appended for inspection rather than for reference which show the nature and extent of these changes in the form of the curves.

Table of σ^2.

$\tau^2=$	·01	·03	·10	·30	1·00	3·00	10·00	30·00	100·00
$n=$									
8	·2531	·2593	·2810	·3430	·5600	1·140	3·350	9·550	31·250
13	·1123	·1148	·1234	·1481	·2344	·4811	1·344	3·811	12·444
18	·07219	·07372	·07908	·09438	·1479	·3010	·8365	2·367	7·722
23	·05319	·05429	·05817	·06925	·1080	·2188	·6066	1·714	5·592
33	·03484	·03555	·03805	·04518	·7015	·1415	·3912	1·105	3·602
43	·02590	·02643	·02827	·03353	·05194	·1045	·2886	·8146	2·655
53	·02062	·02103	·02249	·02666	·04123	·08288	·2287	·6451	2·103

Table of β_1.

$\tau^2=$	·01	·03	·10	·30	1·00	3·00	10·00	30·00	100·00	∞
$n=$										
8	·05685	·1662	·5076	1·230	2·450	3·788	3·965	4·153	4·184	4·252
13	·01517	·04776	·1376	·3400	·7058	1·018	1·205	1·271	1·296	1·3065
18	·008399	·02463	·07645	·1914	·4016	·5857	·6990	·7395	·7546	·7619
23	·005757	·01691	·05247	·1317	·3016	·4093	·4910	·5208	·5314	·5361
33	·003518	·01035	·03214	·08100	·1731	·2559	·3031	·3260	·3334	·3366
43	·002530	·007435	·02315	·05841	·1251	·1858	·2237	·2376	·2429	·2452
53	·001973	·005798	·01807	·04562	·09800	·1458	·1757	·1868	·1910	·1928

Table of β_2.

$\tau^2=$	00	·01	·03	·10	·30	1·00	3·00	10·00	30·00	100·00	∞
$n=$											
8	6·0000	6·1137	6·3179	7·0179	8·4767	10·9668	12·9652	14·1116	14·5024	14·6508	14·7159
13	3·8571	3·8802	3·9248	4·0663	4·3770	4·9397	5·4240	5·7147	5·8186	5·8578	5·8750
18	3·5000	3·5121	3·5356	3·6104	3·7937	4·0828	4·3532	4·5186	4·5783	4·6009	4·6109
23	3·3529	3·3612	3·3768	3·4271	3·5556	3·7486	3·9356	4·0511	4·0930	4·1089	4·1159
33	3·2222	3·2271	3·2365	3·2667	3·3343	3·4619	3·5773	3·6493	3·6756	3·6856	3·6899
43	3·1622	3·1656	3·1723	3·1938	3·2422	3·3261	3·4172	3·4692	3·4886	3·4958	3·4991
53	3·1277	3·1303	3·1356	3·1522	3·1898	3·2640	3·3281	3·3676	3·3826	3·3883	3·3909

12. The fact that the mean value $\bar{r}$ of the observed correlation coefficient is numerically less than ρ might have been interpreted as meaning that given a single observed value r, the true value of the correlation coefficient of the population from which the sample is drawn is likely to be greater than r. This reasoning is altogether fallacious. The mean $\bar{r}$ is not an intrinsic feature of the frequency distribution. It depends upon the choice of the particular variable r in terms of which the frequency distribution is represented. When we use t as variable, the situation is reversed. Whereas in using r we cramp all the high values of the correlation into the small space in the neighbourhood of $r = 1$, producing a frequency curve which trails out in the negative direction and so tending to reduce the value of the mean, by using t, we spread out the region of high values, producing asymmetry in the opposite sense, and obtain a value $\bar{t}$ which is greater than τ. The mean might, in fact, be brought to any chosen point, by stretching and compressing different parts of the scale in the required manner. For the interpretation of a single observation the relation between $\bar{t}$ and τ is in no way superior to that between $\bar{r}$ and ρ. The variable t has been chosen primarily in order to give stability of form to the frequency curves in different parts of the scale. It is in addition a variable to which the analysis naturally leads us, and which enables the mean and moments to be readily calculated, and so a comparison to be made with the standard Pearson curves, but it is not, with these advantages, in a unique position. In some respects the function, $\log \tan \frac{1}{2}\left(\alpha + \frac{\pi}{2}\right)$, is its superior as independent variable.

I have given elsewhere* a criterion, independent of scaling, suitable for obtaining the relation between an observed correlation of a sample and the most probable value of the correlation of the whole population. Since the chance of any observation falling in the range dr is proportional to

$$(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\,dr$$

for variations of ρ, we must find that value of ρ for which this quantity is a maximum, and thereby obtain the equation

$$\frac{\partial}{\partial\rho}\left\{(1-\rho^2)^{\frac{n-1}{2}}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}\right\} = 0.$$

Since
$$\int_0^\infty \frac{dx}{(\cosh x + \cos\theta)^{n-1}} = \frac{1}{\lfloor n-1}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-1}\frac{\theta^2}{2}$$

we have
$$\int_0^\infty \frac{\partial}{\partial\rho}\left\{(1-\rho^2)^{\frac{n-1}{2}}\frac{dx}{(\cosh x+\cos\theta)^{n-1}}\right\} = 0,$$

* R. A. Fisher, "On an absolute criterion for fitting frequency curves," *Messenger of Mathematics*, February, 1912.

which leads by a process of simplification to the equation

$$\int_0^\infty \frac{dx}{(\cosh x - \rho r)^n}(r - \rho \cosh x) = 0.$$

Since $\cosh x$ is always greater than ρr, the factor in the numerator, $r - \rho \cosh x$, must change sign in the range of integration. We therefore see that r is greater than ρ. Further an approximate solution may be obtained for large values of n. The integrand is negligible save when x is very small, and we may write

$$1 + \frac{x^2}{2} \text{ for } \cosh x$$

and

$$(1 - \rho r)^n e^{\frac{nx^2}{2(1-\rho r)}} \text{ for } (\cosh x - \rho r)^n.$$

Then

$$r\int_0^\infty e^{-\frac{nx^2}{2(1-\rho r)}} dx = \rho \int_0^\infty \left(1 + \frac{x^2}{2}\right) e^{-\frac{nx^2}{2(1-\rho r)}} dx,$$

and in consequence, as a first approximation,

$$r = \rho\left(1 + \frac{1 - r^2}{2n}\right).$$

The corresponding relation between t and τ is evidently

$$t = \tau\left(1 + \frac{1}{2n}\right).$$

It is now apparent that the most likely value of the correlation will in general be less than that observed, but the difference will be only half of that suggested by the mean, t.

It might plausibly be urged that in the choice of an independent variable we should aim at making the relation between the mean and the true value approach the above equation, or rather that to which the above is an approximation, or that we should aim at reducing the asymmetry of the curves, or at approximate constancy of the standard deviation. In these respects the function

$$\log \tan \tfrac{1}{2}\left(\alpha + \frac{\pi}{2}\right) \text{ that is, } \tanh^{-1}\rho$$

is not a little attractive, but so far as I have examined it, it does not tend to simplify the analysis, and approaches relative constancy at the expense of the constancy proportionate to the variable, which the expressions in τ exhibit*.

* [It may be worth noting that Mr Fisher's t is the ϕ-square root mean square contingency—of the more usual notation, and is the expression used in determining the probability that correlated material has been obtained by random sampling from uncorrelated material. ED.]