August, 1965

SOME NEW RESULTS ON THE DISTRIBUTION OF THE SAMPLE CORRELATION COEFFICIENT

Harold Ruben

Technical Report No. 59

SOME NEW RESULTS ON THE DISTRIBUTION OF THE SAMPLE CORRELATION COEFFICIENT*

BY HAROLD RUBEN
University of Minnesota

1. INTRODUCTION AND SUMMARY

The correlation coefficient r in samples of size $n > 2$ from a non-singular bivariate normal population with correlation coefficient ρ can be represented in the form (using an obvious generic notation)

$$\tilde{r} = \frac{\xi + \tilde{\rho}\,\chi_{n-1}}{\chi_{n-2}} ,$$

where $\tilde{r} = r/(1-r^2)^{\frac{1}{2}}$, $\tilde{\rho} = \rho/(1-\rho^2)^{\frac{1}{2}}$, ξ is a standardized normal variate and $\xi, \chi_{n-1}, \chi_{n-2}$ are independent. (An equality sign between two variates, here and subsequently, is to be interpreted as meaning that the variates have a common distribution.) Approximation of the χ-variates in the representation, using Fisher's normalization of χ^2, results in the approximation that

$$\frac{(n - \frac{5}{2})^{\frac{1}{2}}\,\tilde{r} - (n - \frac{3}{2})^{\frac{1}{2}}\,\tilde{\rho}}{\{1 + \frac{1}{2}(\tilde{r}^2 + \tilde{\rho}^2)\}^{\frac{1}{2}}}$$

is distributed as a standardized normal variate. High accuracy of the latter approximation is indicated by theoretical considerations and corroborated by some computations. These suggest that the approximation is generally more precise than that based on the Fisherian z (the inverse hyperbolic tangent) transform, corrected for bias or otherwise, and is about as precise as Hotelling's refinement of z. Further, when n is sufficiently large for r to be approximated effectively by a normal variate, it is far superior to that approximation.

The above representation of r has independent value and interest in that it throws some light on various earlier results, and can also be used to yield additional

*This research was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research under Contract AF-AFOSR-885-65.

approximations and expansion for the probability integral of r, such as a high-accuracy Cornish-Fisher expansion and an approximation, valid for high n and ρ, based on F with $n - 1$ and $n - 2$ degrees of freedom. Finally, we may note that the representation provides an exceedingly simple derivation† of the exact distribution of r in normal samples.

2. A SIMPLE REPRESENTATION OF r

Denote the means, standard deviations and correlation coefficient of an arbitrary (not necessarily normal) non-singular, bivariate population by $\mu_x, \mu_y, \sigma_x, \sigma_y, \rho$. Let

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})\}^{\frac{1}{2}}} \tag{1}$$

be the sample correlation coefficient based on a random sample $((x_1,y_1),\ldots,(x_n,y_n))$ of size $n > 2$ $(\Sigma \equiv \sum_1^n, \ \bar{x} = \Sigma\, x_i/n, \ \bar{y} = \Sigma\, y_i/n)$. Defining the transformed correlations $\tilde{\rho}, \tilde{r}$ by

$$\tilde{\rho} = \rho/(1-\rho^2)^{\frac{1}{2}}, \quad \tilde{r} = r/(1-r^2)^{\frac{1}{2}} \tag{2}$$

and the variates u, v, w by

$$u = \frac{b - \beta}{\sigma_y(1-\rho^2)^{\frac{1}{2}}/\{\Sigma(x_i - \bar{x})^2\}^{\frac{1}{2}}}, \tag{3}$$

$$v = \left(\frac{\Sigma(x_i - \bar{x})^2}{\sigma_x^{\,2}}\right)^{\frac{1}{2}}, \tag{4}$$

$$w = \frac{[\Sigma\{y_i - \bar{y} - b(x_i - \bar{x})\}^2]^{\frac{1}{2}}}{\sigma_y(1 - \rho^2)^{\frac{1}{2}}}, \tag{5}$$

---

† I have used essentially this derivation (which appears to be much simpler than previous derivations) in class over the past few years. This experience suggests that the derivation should not present any difficulty to mathematics students in their first year of statistics.

where $\beta = \rho\sigma_y/\sigma_x$, $b = \Sigma(x_i - \bar{x})(y_i - \bar{y})/\Sigma(x_i - \bar{x})^2$ (β, b are the slopes of the mean-square linear regression of y on x in the population and sample, respectively), we have the purely algebraic result

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2 - \{\Sigma(x_i - \bar{x})(y_i - \bar{y})\}^2]^{\frac{1}{2}}} = \frac{u + \tilde{\rho}\, v}{w} . \quad (6)$$

Defining further

$$\epsilon_i^* = \frac{y_i - \mu_y - \beta(x_i - \mu_x)}{\sigma_y(1-\rho^2)^{\frac{1}{2}}} \quad (7)$$

(ϵ_i^* is the residual of y_i from the population mean-square regression, standardized by the factor $\sigma_y(1-\rho^2)^{\frac{1}{2}}$), we also have the decomposition

$$\Sigma\epsilon_i^{*2} = (\Sigma\, \epsilon_i^*/\sqrt{n})^2 + u^2 + w^2 , \quad (8)$$

where it will be noted that

$$u = \sum \frac{(x_i - \bar{x})}{\{\Sigma(x_i - \bar{x})^2\}^{\frac{1}{2}}}\, \epsilon_i^* . \quad (9)$$

(8) and (9) are again purely algebraic in character.

For a normal population, u is a standardized normal variate (call this ξ), v is a χ_{n-1} , w is a χ_{n-2} , and u, v, w are independent; that is, (6) becomes

$$\tilde{r} = \frac{\xi + \tilde{\rho}\, \chi_{n-1}}{\chi_{n-2}} , \quad (10)$$

in which the three variates on the right of (10) are independent. (Conditionally on fixed $x_1,\ldots,x_n$, the ϵ_i^* are independent standardized normal variates, and so therefore are also an arbitrary set of η_i defined by an orthogonal transformation of the ϵ_i^* with $\eta_1 = \Sigma\epsilon_i^*/\sqrt{n}$, $\eta_2 = \Sigma(x_i - \bar{x})\epsilon_i^*/\{\Sigma(x_i - \bar{x})^2\}^{\frac{1}{2}}$. Consequently, conditionally on fixed $x_1,\ldots,x_n$, $u(=\eta_2)$ and $w(=\sqrt{\sum_3^n \xi_i^{*2}} = \sqrt{\sum_3^n \eta_i^2})$ are independent and distributed as a $\mathcal{N}(0,1)$ and a χ_{n-2} . The latter conditional distribution does not involve the x_i , and furthermore v is distributed as a χ_{n-1} , thus justifying the assertion prior to (10).)

It is of interest to relate (10) to previous results.

(a) Distribution of r in normal samples. For $\rho = 0$, (10) reduces to

$$\frac{(n-2)^{\frac{1}{2}}r}{(1-r^2)^{\frac{1}{2}}} = \frac{\xi}{\chi_{n-2}/(n-2)^{\frac{1}{2}}} ,$$

or r is distributed as t with n-2 degrees of freedom. For general ρ , the distribution follows almost as simply. Thus according to (6) and (10), $\hat{r}$, conditionally on fixed v and w , is normal with mean $\tilde{\rho}\, v/w$ and standard deviation $1/w$, the conditional frequency function being, then,

$$w\varphi(\hat{r}w - \hat{\rho}v) ,$$

where $$\varphi(x) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2}$$

is the standardized normal density function, so that the (unconditional) frequency function of $\hat{r}$ is

$$\int_0^\infty \int_0^\infty w\varphi(\hat{r}w - \hat{\rho}v)f_{n-2}(w)f_{n-1}(v)dvdw ,$$

where $f_\nu(x) = \{2^{\frac{1}{2}(\nu-2)}\Gamma(\frac{\nu}{2})\}^{-1}e^{-\frac{1}{2}x^2}x^{\nu-1} \qquad (x > 0)$

is the frequention of a χ_ν . Correspondingly, the frequency function of r $(d\hat{r}/dr = (1-r^2)^{-3/2})$ is

$$(1-r^2)^{-3/2}\int_0^\infty \int_0^\infty w\varphi(\hat{r}w - \hat{\rho}v)f_{n-2}(w)f_{n-1}(v)dvdw = \frac{(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}}{(n-3)!\pi}$$

$$\times \int_0^\infty \int_0^\infty (v'w')^{n-2}e^{-\frac{1}{2}[v'^2-2\rho rv'w'+w'^2]}dv'dw' ,$$

on using the substitutions $v' = v/(1-\rho^2)^{\frac{1}{2}}$, $w' = w/(1-r^2)^{\frac{1}{2}}$ and simplifying the normalizing constant by the duplication formula

$$2^{n-3}\Gamma(\tfrac{n-1}{2})\Gamma(\tfrac{n-2}{2}) = \Gamma(n-2)\sqrt{\pi} .$$

Term by term integration after expansion of $\exp(\rho r v'w')$ as a power series in $(\rho r v'w')$ yields

$$\frac{2^{n-3}(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}}{(n-3)!\pi} \sum_{j=0}^{\infty} \Gamma^2\left(\frac{n-1+j}{2}\right) \frac{(2\rho r)^j}{j!} \tag{11}$$

for the frequency function of r (see, e.g., Cramér, 1946, p. 398), a result useful in the evaluations of the moments of r. On the other hand, from the easily proved result†

$$\int_0^{\infty}\int_0^{\infty} e^{-\frac{1}{2}(v'^2 - 2\rho r v'w' + w'^2)}\, dv'dw' = \frac{\cos^{-1}(-\rho r)}{(1-\rho^2 r^2)^{\frac{1}{2}}} ,$$

the frequency function can be expressed alternatively as

$$\frac{(1-\rho^2)^{\frac{1}{2}(n-1)}(1-r^2)^{\frac{1}{2}(n-4)}}{(n-3)!\pi} \frac{\partial^{n-2}}{\partial(\rho r)^{n-2}} \left(\frac{\cos^{-1}(-\rho r)}{(1-\rho^2 r^2)^{\frac{1}{2}}}\right) , \tag{12}$$

which is Fisher's original form (1915).

(b) Probability of a positive correlation in normal samples. From (10),

$$\begin{aligned} P(r > 0) &= P(\xi + \hat{\rho}\chi_{n-1} > 0) \\ &= P\left(\frac{\xi}{\chi_{n-1}/(n-1)^{\frac{1}{2}}} > -(n-1)^{\frac{1}{2}}\hat{\rho}\right) \\ &= P(t_{n-1} > -(n-1)^{\frac{1}{2}}\rho) . \end{aligned} \tag{13}$$

---

† E.g., by transforming v', w' to polar coordinates, or by a linear transformation which converts $v'^2 - 2\rho r v'w' + w'^2$ to a sum of squares. The result is essentially Sheppard's celebrated formula (see, e.g., Kendall and Stuart, Vol. 1, 1961, p. 351) for the probability of the positive quadrant in a bivariate normal distribution.

Formula (13) has been given by Armsen[†] (1955) and Ruben (1963).

(c) Moments of $\tilde{r}$ in normal samples. A general formula for the moments of $\hat{r}$ has been given by Harley (1957), who also evaluated explicitly the first four moments. (These four moments had already been determined by Fisher in his 1915 paper.) Formula (10) makes the determination, in explicit form, of all moments of $\tilde{r}$ a trivial matter. Thus, for $m < n - 2$,

$$E(\tilde{r}^m) = E(\chi_{n-2}^{-m})E(\xi + \tilde{\rho}\,\chi_{n-1})^m$$

$$= E(\chi_{n-2}^{-m}) \sum_{j=0}^{m} \binom{m}{j} E(\xi^j) E(\chi_{n-1}^{m-j})\tilde{\rho}^{m-j} .$$

After simplification

$$E(\tilde{r}^m) = \frac{1}{\langle n-m-2\rangle_{\frac{1}{2}(m+1)}} \sum_{k=0}^{\frac{1}{2}(m-1)} \langle 1\rangle_k \; \langle n-2\rangle_{\frac{1}{2}(m+1)-k} \binom{m}{2k} \tilde{\rho}^{m-2k} \quad (m=1,3,\ldots), \tag{14a}$$

$$E(\tilde{r}^m) = \frac{1}{\langle n-m-2\rangle_{\frac{1}{2}m}} \sum_{k=0}^{\frac{1}{2}m} \langle 1\rangle_k \langle n-1\rangle_{\frac{1}{2}m-k} \binom{m}{2k} \tilde{\rho}^{m-2k} \quad (m = 2,4,\ldots) \tag{14b}$$

mth degree polynomials in $\tilde{\rho}$. Here $\langle a\rangle_0 = 1$, $\langle a\rangle_j = a(a+2)(a+4)\ldots(a+2j-2)$ for $j = 1,2,\ldots$.

(d) Approximate identification of $\tilde{r}$ with non-central t. With the aim of supplementing tables of non-central t by means of David's tables (1958) of the probability integral of r, Harley (1957) approximated $(n-2)r/(1-r^2)^{\frac{1}{2}}$ to a multiple of a non-central t, the multiple and non-centrality parameter being determined in an ad hoc manner by moment considerations. Insight into the nature of the approximation can be obtained from (10), as follows. Approximate χ_{n-1} by a normal variate with mean $(n-3/2)^{\frac{1}{2}}$

[†] Dr. Linhart kindly informed me of Dr. Armsen's paper after publication of my own note.

and variance 1/2 (Fisher's normalization of χ^2). The numerator of (10) is then distributed approximately as a normal variate with mean

$$\mu = (n - \tfrac{3}{2})^{\frac{1}{2}}\tilde{\rho}$$

and variance

$$\sigma^2 = 1 + \tfrac{1}{2}\tilde{\rho}^2 \quad .$$

Accordingly, $(n-2)^{\frac{1}{2}}\tilde{r}$ is approximated by

$$\frac{\mu + \sigma\eta}{\chi_{n-2}/(n-2)^{\frac{1}{2}}} = \sigma \frac{\eta + \mu/\sigma}{\chi_{n-2}/(n-2)^{\frac{1}{2}}} \quad ,$$

where η is a $N(0,1)$ variate and η, χ_{n-2} are independent; that is, $(n-2)^{\frac{1}{2}}\tilde{r}/\sigma$ is approximated by a non-central t with n-2 degrees of freedom and non-centrality parameter μ/σ . This is Harley's result.

Referring again to (6) and (10), it is clear that $(n-2)^{\frac{1}{2}}\tilde{r}$, the square root of the regression mean-square ratio in an analysis of variance of the y_i (x_i fixed) used in testing $\beta = 0$, can be regarded as a weighted non-central t (rather than, as above, a weighted normal), thereby providing an alternative, though less direct, derivation of the distribution of r. Conditionally on given v , $(n-2)^{\frac{1}{2}}\tilde{r}$ is a non-central t with n-2 degrees of freedom and non-centrality parameter $\tilde{\rho}v$, the corresponding conditional frequency function of $\tilde{r}$ being then (see, e.g., Wilks, 1962, p. 247)

$$\frac{e^{-\frac{1}{2}\tilde{\rho}^2 v^2}}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi}} \sum_{j=0}^{\infty} \Gamma\left(\frac{n-1+j}{2}\right) \frac{(\tilde{\rho}v\sqrt{2})^j}{j!} (\tilde{r}^2)^{\frac{1}{2}j} (1 + \tilde{r}^2)^{-\frac{1}{2}(n-1+j)} \quad .$$

Averaging with respect to the distribution of v and multiplying by $(1-r^2)^{-3/2}$ yields again the series (11) for the (unconditional) frequency function of r. It will be noted that this derivation is not essentially distinct from the familiar derivation of the distribution of the multiple correlation coefficient R in normal samples (e.g.,

Anderson, 1958, pp. 92-94) consisting in weighting the conditional distribution of the regression mean-square ratio $\frac{R^2/(p-1)}{(1-R^2)/(n-p)}$ for given values of the p-1 'independent' variables with respect to the distribution of those variables.†

†The failure to carry this out for p = 2 (except in the case $\rho = 0$, as in Anderson, 1958, pp. 61-64) provides an interesting historical reflection. At the time of appearance of Fisher's 1915 paper, when (in Hotelling's phrase) statisticians were enthusiastically exploring the universe with the correlation coefficient as their chief measuring instrument and the shift in emphasis from correlational to regression and analysis of variance concepts still lay well ahead in the future, 'regression' and conditional derivations (simple as these are) for the distribution of r could not be expected. However, there appears to be no reason (other than inertia) why such derivations should not be presented in textbooks and in class. On the other hand, the derivations do have an element of asymmetry which is absent in the corresponding derivations for $p > 2$.

3. A NEW APPROXIMATION FOR THE DISTRIBUTION OF r

From (10), for an arbitrary fixed value r_0 ($|r_0| < 1$) with $\tilde{r}_0 = r_0/(1 - r_0^2)^{\frac{1}{2}}$,

$$P(r \leq r_0) = P(\tilde{r} \leq \tilde{r}_0)$$

$$= P(\xi + \tilde{\rho}\,\chi_{n-1} - \tilde{r}_0\,\chi_{n-2} \leq 0), \tag{15}$$

the evaluation of the probability integral of r being then equivalent to the probability that the variate L, defined by

$$L = \xi + \tilde{\rho}\,\chi_{n-1} - \tilde{r}_0\,\chi_{n-2} \tag{16}$$

is non-positive. (Cf. Kendall and Stuart, Vol. 1, 1961; p.299.) Since L is a linear combination of three independent variates, one of which (ξ) is exactly normal and two (χ_{n-1} and χ_{n-2}) are approximately normal, it is evident that L itself will be approximately normal, and one would further expect L to be more nearly normal than either of the two χ-variates. This gains urgency from a consideration of the skewness and kurtosis of L.

The mth cumulant of L, $\kappa_m(L)$, is related to the corresponding cumulants of ξ, χ_{n-1} and χ_{n-2} by

$$\kappa_m(L) = \kappa_m(\xi) + \tilde{\rho}^m\,\kappa_m(\chi_{n-1}) + (-\tilde{r}_0)^m\,\kappa_m(\chi_{n-2}). \tag{17}$$

On using

$$\kappa_3(\chi_\nu) \sim \frac{1}{4\nu^{\frac{1}{2}}}, \quad \kappa_4(\chi_\nu) \sim \frac{3}{16\nu^2},$$

one obtains for the standardized third and fourth cumulants of L,

$$\gamma_1(L) \equiv \frac{\kappa_3(L)}{\{\kappa_2(L)\}^{\frac{3}{2}}} = \frac{\tilde{\rho}^3 - \tilde{r}_0^3}{\{1 + \frac{1}{2}(\tilde{\rho}^2 + \tilde{r}_0^2)\}^{\frac{3}{2}}}\;\frac{1}{4n^{\frac{1}{2}}} + o(\tfrac{1}{n}),$$

$$\gamma_2(L) \equiv \frac{\kappa_4(L)}{\{\kappa_2(L)\}^2} = \frac{\tilde{\rho}^4 + \tilde{r}_0^4}{\{1 + \frac{1}{2}(\tilde{\rho}^2 + \tilde{r}_0^2)\}^2}\;\frac{3}{16n^2} + o(\tfrac{1}{n^2}).$$

Asymptotically, then

$$|\gamma_1(L)| \leq 2^{\frac{3}{2}} \frac{|\tilde{\rho}|^3 + |\tilde{r}_0|^3}{\{|\tilde{\rho}|^2 + |\tilde{r}_0|^2\}^{\frac{3}{2}}} \frac{1}{4n^{\frac{1}{2}}} \leq \frac{1}{(2n)^{\frac{1}{2}}} = \gamma_1(\chi_n) , \qquad (18)$$

$$\gamma_2(L) \leq 2^2 \frac{|\tilde{\rho}|^4 + |\tilde{r}_0|^4}{\{|\tilde{\rho}|^2 + |\tilde{r}_0|^2\}^2} \frac{3}{16n^2} \leq \frac{3}{4n^2} = \gamma_2(\chi_n) . \qquad (19)$$

In view of (18) and (19), we shall approximate L by a normal variate. By Fisher's normalization of χ^2, χ_{n-1} and χ_{n-2} are approximately normal with means $(n - \frac{3}{2})^{\frac{1}{2}}$, $(n - \frac{5}{2})^{\frac{1}{2}}$ and variances $\frac{1}{2}$. Hence L is approximately normal with mean

$$(n - \tfrac{3}{2})\tilde{\rho} - (n - \tfrac{5}{2})\tilde{r}_0$$

and variance

$$1 + \tfrac{1}{2}(\tilde{\rho}^2 + \tilde{r}_0^2).$$

Accordingly,

$$P(r \leq r_0) \equiv P(L \leq 0)$$

$$\doteq \Phi\left[\frac{(n - \frac{5}{2})^{\frac{1}{2}}\tilde{r}_0 - (n - \frac{3}{2})^{\frac{1}{2}}\tilde{\rho}}{\{1 + \frac{1}{2}(\tilde{r}_0^2 + \tilde{\rho}^2)\}^{\frac{1}{2}}}\right] , \qquad (20)$$

where Φ denotes the standardized cumulative normal distribution function; that is, replacing $\tilde{r}_0$ by $\tilde{r}$,

$$g_n(r,\rho) \equiv \frac{(n - \frac{5}{2})^{\frac{1}{2}}\tilde{r} - (n - \frac{3}{2})^{\frac{1}{2}}\tilde{\rho}}{\{1 + \frac{1}{2}(\tilde{r}^2 + \tilde{\rho}^2)\}^{\frac{1}{2}}} \qquad (21)$$

is approximately distributed as a standardized normal variate†. This result has the following interesting interpretation. r is asymptotically normal with mean ρ and variance $(1 - \rho^2)^2/n$. Hence $\tilde{r}$ $(= r/(1-r^2)^{\frac{1}{2}})$ is likewise asymptotically normal with mean $\tilde{\rho}$ $(= \rho/(1 - \rho^2)^{\frac{1}{2}})$ and variance

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

The corresponding estimate of variances is $(1 + \tilde{r}^2)/n$. An approximate standard error of $\tilde{r}$, based on <u>averaging</u> the last two variances, is

$$\left[\frac{1}{n} \{1 + \tfrac{1}{2}(\tilde{r}^2 + \tilde{\rho}^2)\}\right]^{\frac{1}{2}} ,$$

so that

$$\frac{n^{\frac{1}{2}}(\tilde{r} - \tilde{\rho})}{\{1 + \frac{1}{2}(\tilde{r}^2 + \tilde{\rho}^2)\}^{\frac{1}{2}}} ,$$

a variate which in large samples is not essentially distinct from $g_n(r,\rho)$, is asymptotically normal with zero mean and unit standard derivation.

The efficacy of the proposed approximation based on $g_n(r,\rho)$ (formula (20)) can be judged from the accompanying tables, in which C, F_1, F_2, H, N have the following meaning:

C = 10^5 × error when $g_n(r,\rho)$ is approximated by a standardized normal variate,

F_1 = 10^5 × error when $z = \frac{1}{2} \log \frac{1 + r}{1 - r}$ is approximated by a normal variate with mean $\zeta = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}$ and variance $\frac{1}{n - 3}$ (Fisher's normalization of r),

---

†This result can also be obtained directly from (10) with the aid of the following result due to Geary (1930). If q denotes the ratio of two independent normal variates with means μ_1, μ_2 and standard derivations σ_1, σ_2, then $(\mu_1 - \mu_2 q)/(\sigma_1^2 + \sigma_2^2 q^2)^{\frac{1}{2}}$ is approximately normal with zero mean and unit standard derivation, provided the range of the denominator variate is effectively positive. The required result follows from (10) on approximating χ_{n-1} and χ_{n-2} by normal variates as above. (Here $q = \tilde{r}$, $\mu_1 = (n - \frac{3}{2})^{\frac{1}{2}} \tilde{\rho}$, $\mu_2 = (n - \frac{5}{2})\tilde{r}$, $\sigma_1^2 = 1 + \frac{1}{2}\tilde{\rho}^2$, $\sigma_2^2 = \frac{1}{2}$.)

[illegible]

[illegible]

$$[illegible]$$

[illegible]

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

[illegible]

F_2 = $10^5 \times$ error when z is approximated by a normal variate with mean $\zeta + \frac{\rho}{2(n-1)}$ and variance $\frac{1}{n-3}$ (Fisher's normalization of r, corrected for bias),

H = $10^5 \times$ error when $z - \frac{3z + r}{n}$ is approximated by a normal variate with mean $\zeta - \frac{3\zeta + \rho}{n}$ and variance $\frac{1}{n-1}$ (Hotelling's† approximation, 1953),

N = $10^5 \times$ error when r is approximated by a normal variate with mean ρ and variance $(1 - \rho^2)^2/n$,

P(r) $\equiv$ $10^5 \times$ actual value of probability integral.

The values of n, r, ρ in the tables are those selected by Miss David (1938; pp. xxxii - xxxiii) in her own investigation of the accuracy of the bias-corrected z-transformation and the approximation which consists in replacing r by a normal variate.

The tables suggest the following conclusions.

(i) The proposed approximation is much superior to that based on z (good as this is);

(ii) it is still superior to the approximation based on the bias-corrected z over the entire range of distribution, except possibly at the tails;

(iii) it is of about the same order of accuracy as that based on Hotelling's refinement of z;

(iv) it is much superior to the approximation which consists in regarding r itself as normally distributed (when n is sufficiently large for the latter approximation to be at all effective).

---

† The transformation $z^* = z - (3z + r)/n$ was obtained by Hotelling for the purpose of improving on the stabilization of variance induced by the inverse hyperbolic tangent transformation, and not primarily for improving on the z-normalization. However (as Hotelling conjectured and as the present calculations indicate), improved normalization does in fact generally appear to result from the use of z^*.

4. CONCLUDING REMARKS

Equations (10) and (15) give rise to various additional approximations for the distribution of r. Although an investigation into the relative merits and demerits of these approximations hardly seems worthwhile, two of the approximations are of sufficient interest and importance to warrant some mention.

(i) The variate $|\xi|$ is stochastically negligible relative to the variate $|\tilde{\rho}|\chi_{n-1}$ if $n|\tilde{\rho}| \gg 1$. This suggests that under the latter condition (10) can be adequately replaced by the relation

$$\tilde{r} = \tilde{\rho}\frac{\chi_{n-1}}{\chi_{n-2}} ,$$

$$\text{i.e.,} \quad \frac{\tilde{r}}{\tilde{\rho}} = \left(\frac{n-1}{n-2}\right)^{\frac{1}{2}} F^{\frac{1}{2}}_{n-1,n-2} \qquad (|n\tilde{\rho}| \gg 1) , \tag{22}$$

or

$$\log \frac{\tilde{r}}{\tilde{\rho}} = \log \left\{\left(\frac{n-1}{n-2}\right)^{\frac{1}{2}}\right\} + z_{n-1,n-2} \qquad (|n\tilde{\rho}| \gg 1) , \tag{23}$$

where $z_{n-1,n-2}$ is a variate distributed as Fisher's z (half the naturel logarithm of a variance ratio) with n - 1 and n - 2 degrees of freedom (r and ρ are to be of the sign in (23)). An equivalent statement is that

$$\frac{1}{1 + \left(\frac{\tilde{r}}{\tilde{\rho}^2}\right)^2} = \beta_{\frac{1}{2}(n-2),\frac{1}{2}(n-1)} \qquad (|n\tilde{\rho}| \gg 1) , \tag{24}$$

where the variate of the right of (24) is a beta-variate of the first kind with parameters $\frac{1}{2}(n-2)$ and $\frac{1}{2}(n-1)$. Spot calculations indicate that this approximation gives remarkable accuracy. It should be remarked that the calculation of the probability integral of r for high n and ρ has proved in the past to be most difficult (see E. S. Pearson's preface to Miss David's tables), and that the approximation mentioned here performs best precisely under such conditions.

For $\rho = 0$, $(n-2)^{\frac{1}{2}}\tilde{r}$ is distributed exactly as t with $n - 2$ degrees of freedom. From general continuity considerations, $(n-2)^{\frac{1}{2}}\tilde{r}$ is then distributed approximately as t with $n - 2$ degrees of freedom if $|\rho|$ is small. The result of the previous paragraph provides an interesting contrast in that it asserts that the distribution of $\tilde{r}$ is to be referred to an F(with approximately equal degrees of freedom) if $|\rho|$ is large.

(ii) We have seen in Section 3 that the rather crude Fisherian normalization of χ^2 already produces high accuracy with the aid of (15), involving the computation of $P(L \leq 0)$. It is clear that extremely high accuracy can be obtained if the latter probability is evaluated by means of a Cornish-Fisher asymptotic expansion and that such accuracy can be obtained even if only the first correcting term in the expansion is used. The main term in the expansion is $\Phi(-\mu_L/\sigma_L)$, where μ_L, σ_L are the mean and standard deviation of L. This does not differ materially from (19). [(19) is obtained from $\Phi(-\mu_L/\sigma_L)$ by approximating μ_L and σ_L, in that the mean and variance of a χ_ν are taken as $(\nu - \frac{1}{2})^{\frac{1}{2}}$ and $\frac{1}{2}$.] The correcting terms, which are in powers of $n^{-\frac{1}{2}}$, involve the cumulants of L, and these are easily computed from (17). (Bounds to the error induced by termination after a given number of correcting terms can also be computed.)

The robustness of r (previously studied by Gayen, 1951) can also be similarly investigated through the relationship

$$\rho(r \leq r_0) \quad = \quad P(u + \tilde{\rho}v - \tilde{r}_0 w \leq 0), \tag{25}$$

where u, v, w are defined by (3), (4) and (5). The variate

$$L \quad = \quad u + \tilde{\rho}v - \tilde{r}_0 w \tag{26}$$

is a symmetric function of the observations whose asymptotic moments can be determined from Fisher's theory of k-statistics, and the Cornish-Fisher expansion for $P(L \leq 0)$ thereby determined (cf. David and Johnson, 1951). The variates u, v, w, being functions of central moments, are asymptotically jointly normal, whence L is itself asymptotically normal. The leading term in the expansion is again $\Phi(-\mu_L/\sigma_L)$.

We remark in conclusion that in comparing $g_n(r,\rho)$ (defined in (21)) with Fisher's z-transform of r from the point of view of statistical applications, rather than that (as in Section 3) of accuracy, it is clear that $g_n(r,\rho)$ lacks the flexibility of z arising from its variance-stabilization property: $g_n(r,\rho)$ cannot, for example, be used to test for the homogeneity of correlations in independent sets of data (unless, of course, one approximates further by substituting an estimate for the common unknown correlation). On the other hand, it can profitably be used (just as can the z-transform, but more accurately) to test the more restricted hypothesis that the population correlations have a specified common value (see pp. xxiv - xxv of Miss David's tables for examples when the latter is zero) by referring

$$\{g_{n_1}(r_1,\rho_0)\}^2 + \dots + \{g_{n_k}(r_k,\rho_0)\}^2 \tag{27}$$

to χ^2 with k degrees of freedom (ρ_0 is the postulated common value of the population correlations, r_i is the correlation computed from the ith set of data, n_i = size of ith set). Another and very obvious advantage is that only tables of the normal distribution are required for probability integrals and percentage points of r. In particular, tests and confidence intervals relating to ρ and based on $g_n(r,\rho)$ are easily contructed in the usual manner. Thus a two-sided confidence interval for $\tilde{\rho}$ with coverage probability $1 - \alpha$, arising from the inversion of the inequality

$|g_n(r,\rho)| < d_{\alpha/2}$ ($d_{\alpha/2}$ is the two sided $100\alpha\%$ point of a standard normal deviate) has limits

$$(2n-3-d^2_{\alpha/2})^{-1}[\{(2n-5)(2n-3)\}^{\frac{1}{2}}\tilde{r} \pm \{(4n-6-d^2_{\alpha/2}) + (4n-8-2d^2_{\alpha/2})\tilde{r}^2\}^{\frac{1}{2}}d_{\alpha/2}], \tag{28}$$

a rough approximation being

$$\frac{r}{(1 - r^2)^{\frac{1}{2}}} \pm d_{\alpha/2}\,\frac{1}{n^{\frac{1}{2}}}\,\frac{1}{(1 - r^2)^{\frac{1}{2}}} . \tag{29}$$

Denoting the confidence limits for $\tilde{\rho}$ by c_1, c_2, the corresponding limits for ρ are $c_1/(1 + c_1^2)^{\frac{1}{2}}$, $c_2/(1 + c_2^2)^{\frac{1}{2}}$.

REFERENCES

ANDERSON, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

ARMSEN, P. (1956). Note on a simple identity concerned with particular values of the distribution functions of correlation coefficients in bivariate normal populations. Jour. Nat. Inst. Personnel Res., 6, 175-6.

CRAMÉR, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

DAVID, F. N. (1938). Tables of the Correlation Coefficient. Cambridge University Press.

DAVID, F. N. and JOHNSON, N.L. (1951). The effect of non-normality on the power function of the F-test. Biometrika, 38, 43-57.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10, 507-21.

GAYEN, A.K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size down from non-normal universes. Biometrika, 38, 219-47.

GEARY, R. C. (1930). The frequency distribution of the quotient of two normal variables. Jour. Roy. Statist. Soc., 93, 442-46.

HARLEY, B. I. (1957). Relation between the distributions of non-central t and of a transformed correlation coefficient. Biometrika, 44, 219-24.

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms. Jour. Roy. Statist. Soc. B, 15, 193-225.

KENDALL, M. G. & STUART, A. (1961). The Advanced Theory of Statistics, Vol. 2. London: Charles Griffin.

RUBEN, H. (1963). Probability of a positive sample correlation. Ann. Math. Statist., 34, 694 (abstract).

WILKS, S. S. (1962). Mathematical Statistics. New York: Wiley.

Table 1. Accuracy of approximations to probability integral of r for n = 11, $\rho = 0.5$

r	P(r)	C	F_1	F_2	H
-.25	995	45	147	-48	-180
-.20	1414	52	257	-16	-194
-.15	1974	59	405	35	-197
-.10	2710	67	597	108	-188
-.05	3666	74	836	204	-163
.00	4893	80	1120	322	-124
.05	6449	83	1447	457	- 71
.80	93270	-346	717	-174	204
.825	95517	-387	580	- 56	278
.85	97267	-400	454	44	307
.875	98526	-377	332	103	281
.90	99335	-313	213	110	207

Table 2. Accuracy of approximations to probability integral of r for n = 11, ρ = 0.9

r	P(r)	C	F_1	F_2	H
.60	881	306	497	107	-120
.625	1165	340	694	164	-123
.65	1553	372	882	240	-122
.675	2086	402	1165	342	-112
.70	2826	427	1528	475	- 90
.725	3864	441	1989	644	- 54
.75	5335	438	2561	820	- 2
.80	10461	377	4071	1360	165
.81	12047	333	4398	1441	181
.82	13873	300	4744	1528	210
.83	15998	263	5086	1600	232
.84	18470	223	5413	1650	246
.85	21342	186	5714	1672	250
.86	24679	150	5966	1651	235
.87	28545	119	6151	1579	201
.88	33006	97	6244	1449	145
.89	38127	83	6217	1249	65
.90	43957	79	6043	979	- 36
.91	50517	77	5698	646	-147
.92	57776	64	5168	271	-250
.93	65614	19	4461	98	-313
.94	73783	- 86	3611	-395	-301
.95	81851	-266	2691	-532	-179
.97	94987	-669	1040	-179	259
.975	97083	-677	723	- 34	305
.98	98571	-606	450	64	279
.985	99470	-455	227	87	184

Table 3. Accuracy of approximations to probability integral of r for a = 160, ρ = 0.8

r	P(r)	C	F_1	F_2	H
.70	172	7	16	- 2	- 4
.75	5388	22	381	26	4
.80	48729	2	1271	14	0
.85	97402	- 26	179	- 6	5

Table 4. Accuracy of approximations to probability integral of r for n = 400, ρ = 0

r	P(r)	C	F_1	H	N
- .16	66	0	- 1	- 1	- 4
- .14	251	0	- 2	- 2	- 7
- .12	817	0	- 3	- 3	-10
- .10	2282	0	- 2	- 3	- 7
- .08	5507	0	2	2	5
- .06	11559	0	8	7	23
- .04	21249	0	1	11	34
- .02	34503	0	8	8	27
.00	50000	0	0	0	0

Table 5. Accuracy of approximations to probability integral of r for n = 400, ρ = 0.4

r†	P(r)	C	F_1	H	N
.22	4	- 1	- 1	- 1	- 4
.26	83	0	2	- 1	- 37
.30	1121	1	27	- 3	-220
.34	8132	0	157	3	-281
.38	31550	- 1	367	8	544
.42	60850	- 9	355	- 1	569
.46	92781	-24	110	-25	-342
.50	99341	23	44	27	-194
.54	99999	-16	- 15	-16	- 42

†In Miss David's tables each number in this column has a negative sign inadvertently attached to it.

Table 6. Accuracy of approximations to probability integral of r for n = 400, ρ = 0.9

r	P(r)	C	F_1	H	N
.85	1	0	0	0	- 1
.86	17	1	1	0	- 15
.87	261	4	17	- 3	-169
.88	2588	11	144	- 1	-658
.89	15254	7	561	10	63
.90	49100	0	900	0	1797
.91	86000	- 7	476	- 8	-278
.92	98941	- 9	62	4	-1670
.93	99988	0	2	1	- 67