PROBABLE ERROR OF A CORRELATION COEFFICIENT.

By STUDENT.

At the discussion of Mr R. H. Hooker's recent paper "The correlation of the weather and crops" (*Journ. Royal Stat. Soc.* 1907) Dr Shaw made an enquiry as to the significance of correlation coefficients derived from small numbers of cases.

His question was answered by Messrs Yule and Hooker and Professor Edgeworth, all of whom considered that Mr Hooker was probably safe in taking ·50 as his limit of significance for a sample of 21. They did not, however, answer Dr Shaw's question in any more general way. Now Mr Hooker is not the only statistician who is forced to work with very small samples, and until Dr Shaw's question has been properly answered the results of such investigations lack the criterion which would enable us to make full use of them. The present paper, which is an account of some sampling experiments, has two objects: (1) to throw some light by empirical methods on the problem itself, (2) to endeavour to interest mathematicians who have both time and ability to solve it.

Before proceeding further, it may be as well to state the problem which occurs in practice, for it is often confused with other allied questions.

A random sample has been obtained from an indefinitely large* population and r† calculated between two variable characters of the individuals composing the sample. We require the probability that R for the population from which the sample is drawn shall lie between any given limits.

It is clear that in order to solve this problem we must know two things: (1) the distribution of values of r derived from samples of a population which has a given

* Note that the indefinitely large population need not actually exist. In Mr Hooker's case his sample was 21 years of farming under modern conditions in England, and included all the years about which information was obtainable. Probably it could not actually have been made much larger without loss of homogeneity, due to the mixing with farming under conditions not modern; but one can imagine the population indefinitely increased and the 21 years to be a sample from this.

† Throughout the rest of this paper "r" is written for the correlation coefficient of a sample and R for correlation coefficient of a population.

R, and (2) the *à priori* probability that R for the population lies between any given limits. Now (2) can hardly ever be known, so that some arbitrary assumption must in general be made; when we know (1) it will be time enough to discuss what will be the best assumption to make, but meanwhile I may suggest two more or less obvious distributions. The first is that any value is equally likely between $+1$ and -1, and the second that the probability that x is the value is proportional to $1-x^2$: this I think is more in accordance with ordinary experience: the distribution of *à priori* distribution would then be expressed by the equation $y = \frac{2}{\pi}(1-x^2)$.

But whatever assumption be made, it will be necessary to know (1), so that the solution really turns on the distribution of r for samples drawn from the same population. Now this has been determined for *large* samples with as much accuracy as is required, for Pearson and Filon (*Phil. Trans.* Vol. 191 A, p. 229 *et seq.*) showed that the standard deviation is $\frac{1-r^2}{\sqrt{n}}$ and of course for large samples the distribution is sure to be practically normal unless r is very close to unity. But their method involves approximations which are not legitimate when the sample is small. Besides this the distribution is not then normal, so that even if we had the standard deviation a great deal would still remain unknown.

In order to throw some light on this question I took a correlation table* containing 3000 cases of stature and length of left middle finger of criminals, and proceeded to draw samples of four from this population†. This gave me 750 values of r for a population whose real correlation was ·66. By taking the statures of one sample with the middle finger lengths of the next sample I was enabled to get 750 values of r for a population whose real correlation was zero. Next I combined each of the samples of four with the tenth sample before it and with the tenth sample after it, thus obtaining two sets of 750‡ values from samples of 8, with real correlation ·66 and zero.

Besides this empirical work it is possible to calculate *à priori* the distribution for samples of two as follows.

For clearly the only values possible are $+1$ and -1, since two points must always lie on the regression line which joins them§.

Next consider the correlation between the difference between the values of one character in two successive individuals, and the difference between the values of the other character in the same individuals. It is well known to be the same as that between the values themselves, if the individuals be in random order.

* *Biometrika*, Vol. I. p. 219. W. R. Macdonnell. † *Biometrika*, Vol. VI. p. 13. Student.

‡ Not strictly independent, but practically sufficiently nearly so. This method was adopted in order to save arithmetic.

§ There are of course indeterminate cases when the values are the same for one character, but they become rarer as we decrease the unit of grouping until with an infinitesimal unit of grouping the statement in the text is true.

Also, if an indefinitely large number of such differences be taken, it is clear that the means of the distributions will have the value zero. Hence, if the correlation be determined from a fourfold division through zero we can apply Mr Sheppard's* result that if A and B be the numbers in the large and the small divisions of the table respectively $\cos \frac{\pi B}{A+B} = R$, where R is the correlation of the original system.

But if a pair of individuals whose difference falls in either of the small divisions be considered to be a random sample of 2, their r will be found to be -1, while that of a pair whose difference falls in one of the large divisions is $+1$. Hence the distribution of r for samples of 2 is AN at $+1$, and BN at -1, where $A+B=1$, and $B = \frac{\cos^{-1} R}{\pi}$.

When $R = 0$, there is of course even division, half the values being $+1$, and half -1; when $R = \cdot 66$, $B = \frac{\cos^{-1} \cdot 66}{\pi} = \cdot 271$, therefore $A = \cdot 729$, and the mean is at $\cdot 729 - \cdot 271 = \cdot 458$. The S.D. $= \sqrt{1-(\cdot 458)^2} = \cdot 889$. It is noteworthy that the mean value is considerably less than R.

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 $\frac{2}{\pi} \sin^{-1} R$ and its second moment coefficient of $1 - \left(\frac{2}{\pi} \sin^{-1} R\right)^2$ may furnish a clue to the distribution when n is greater than 2.

Besides these series, I have another shorter one of 100 values of r from samples of 30, when the real value is $\cdot 66$. The distributions of the various trials are given in the table.

Several peculiarities will be noticed which are due to the effects of grouping, particularly in the samples of 4. Firstly, there is a lump at zero; with such small numbers zero is not an uncommon value of the product moment and then, whatever the values of the standard deviations, $r = 0$.

Next there are five indeterminate cases in each of the distributions for samples of 4. These are due to the whole sample falling in the same group for one variable. In such a case, both the Standard Deviation and the product moment vanish and r is indeterminate.

Lastly, with such small samples one cannot use Sheppard's corrections for the Standard Deviations, as r often becomes greater than unity. So I did not use the corrections except in the case of the samples of 30, yet on the whole the values of the Standard Deviations are no doubt too large. This does not much affect the values of r in the neighbourhood of zero, but there is a tendency for larger values

* *Phil. Trans.* A. Vol. CXCII. p. 141.

Distribution of Correlation Coefficients.

Scale	−1·00−·98	−·97−·93	−·92−·88	−·87−·83	−·82−·78	−·77−·73	−·72−·68	−·67−·63	−·62−·58	−·57−·53	−·52−·48	−·47−·43	−·42−·38	−·37−·33	−·32−·28	−·27−·23	−·22−·18	−·17−·13	−·12−·08	−·07−·03	−·02+·02
No correlation, Samples of 4*	8	17	16	23	13	23	9½	19½	19	17	23	18½	25½	15	22	22	19	19½	23½	15	30
No correlation, Samples of 8	1	—	—	1	7	10	10	15	14	15	18	24	18	27	36	33	43	45	26½	26½	34
Correlation of ·66 Samples of 4†	3	3½	2½	2	2	8	3½	1½	5	9	8	7	3	1	4	10	5	5	3	6	16
Correlation of ·66 Samples of 8	—	—	—	—	—	—	—	—	—	2	—	2	—	3	—	3	6	4	3	4	5

Scale	+·03+·07	+·08+·12	+·13+·17	+·18+·22	+·23+·27	+·28+·32	+·33+·37	+·38+·42	+·43+·47	+·48+·52	+·53+·57	+·58+·62	+·63+·67	+·68+·72	+·73+·77	+·78+·82	+·83+·87	+·88+·92	+·93+·97	+·98+1·00
No correlation, Samples of 4*	18	14	23	22	22½	24	13½	17	23	24	24½	9½	17½	17½	11½	12½	9	22½	13½	9
No correlation, Samples of 8	42½	27½	34	23½	36½	33½	28½	19	24	22	15	13½	9	2½	7	5	3	—	—	—
Correlation of ·66 Samples of 4†	4	7	9	11	8	14	11	20½	18	30	23½	27	33½	45½	39½	41	80½	64	91	59
Correlation of ·66 Samples of 8	7	7	5	11	17	20	17	22	34	46½	44½	61	59	66½	56½	80½	60	66	33½	4

Scale	·26 to ·28	·29 to ·31	·32 to ·34	·35 to ·37	·38 to ·40	·41 to ·43	·44 to ·46	·47 to ·49	·50 to ·52	·53 to ·55	·56 to ·58	·59 to ·61	·62 to ·64	·65 to ·67	·68 to ·70	·71 to ·73	·74 to ·76	·77 to ·79	·80 to ·82	·83 to ·85
Distribution with samples of 30	1	—	—	1	—	1	—	1	6	4	5	9	12	10	15	12	7	10	5	1

* There are five indeterminate cases so that the total is 745, while there are 750 in the other two distributions.

† The moment coefficients of this distribution were actually calculated from a different grouping as below:

−1·00−·95	−·94−·90	−·89−·85	−·84−·80	−·79−·75	−·74−·70	−·69−·65	−·64−·60	−·59−·55	−·54−·50	−·49−·45	−·44−·40	−·39−·35	−·34−·30	−·29−·25	−·24−·20	−·19−·15	−·14−·10	−·09−·05	−·04−·00
5	4	2	2	6	4	3	4	7	9	7	5	1	2	6	7	7	5	4	18

+·01+·05	+·06+·10	+·11+·15	+·16+·20	+·21+·25	+·26+·30	+·31+·35	+·36+·40	+·41+·45	+·46+·50	+·51+·55	+·56+·60	+·61+·65	+·66+·70	+·71+·75	+·76+·80	+·81+·85	+·86+·90	+·91+·95	+·96+1·00
2	5	11	8	12	12½	7½	17	21	24	20	35	31	35	50	34	58½	76	77	98½

to come too low, so that there is a deficiency of cases towards 1 and -1. This introduces an error into the Standard Deviation of all the series to some extent, but of course the mean is unaltered when there is no correlation. The series for samples of 4 are affected more than those from samples of 8, as the mean Standard Deviation of samples of 4 is the smaller, so that the unit of grouping is comparatively larger.

The moment coefficients of the five distributions were determined, and the following values found*:—

	Mean	S.D.	μ_2	μ_3	μ_4	β_1	β_2
Samples of 4 ($r = 0$)	—	·5512	·3038	—	·1768	—	1·918
Samples of 8 ($r = 0$)	—	·3731	·1392	—	·0454	—	2·336
Samples of 4 ($r = ·66$)	·5609	·4680	·2190	− ·1570	·2152	2·245	4·489
Samples of 8 ($r = ·66$)	·6139	·2684	·07202	− ·02634	·02714	1·857	5·232
Samples of 30 ($r = ·66$)	·661	·1001	·01003	− ·000882	·000461	·7713	4·580

Considering first the "no correlation" distributions I attempted to fit a Pearson curve to the first of them. As might be expected, the range proved limited and as symmetry had been assumed in calculating the moments, a Type II curve resulted. The equation was $y = y_0\left(1 - \frac{x^2}{1·076}\right)^{·573}$, the range of which is 2·074.

Now the real range is clearly 2, and only a very small alteration in β_2 is required to make the value of the index zero. Consequently the equation $y = y_0(1 - x^2)^9$ was suggested. This means an even distribution of r between 1 and -1, with S.D. = ·5774 ± ·010 vice ·5512 actual, $\mu_2 = ·3333 \pm ·0116$ vice 3038, $\mu_4 = ·2000 \pm ·016$ vice ·1768 and $\beta_2 = 1·800 \pm ·12$ vice 1·918, all values as close as could perhaps be expected considering that the grouping must make both μ_2 and μ_4 too low.

Working from $y = y_0(1 - x^2)^9$ for samples of 4 I guessed the formula $y = y_0(1 - x^2)^{\frac{n-4}{2}}$ and proceeded to calculate the moments.

By using the transformation $x = \sin\theta$ we get $y = y_0\cos^{n-4}\theta$,

$$dx = \cos\theta\, d\theta,$$

$$2\int_0^1 y\,dx = 2y_0\int_0^{\frac{\pi}{2}} \cos^{n-1}\theta\, d\theta,$$

$$2\int_0^1 x^2 y\,dx = 2y_0\int_0^{\frac{\pi}{2}} \cos^{n-3}\theta\, d\theta - 2y_0\int_0^{\frac{\pi}{2}} \cos^{n-1}\theta\, d\theta,$$

and so on.

Whence

$$\mu_2 = \frac{1}{n-1}, \qquad \mu_4 = \frac{3}{(n-1)(n+1)}, \qquad \beta_2 = \frac{3(n-1)}{n+1} = 3 - \frac{6}{n+1}$$

In the cases of no correlation the moments were taken about zero, the known centroid of the distribution.

Putting $n = 8$ we get the equation $y = y_0(1 - x^2)^3$ and

$\mu_2 = \frac{1}{7} = \cdot1429 \pm \cdot0050$ instead of actual $\cdot1392$,

$\mu_4 = \frac{1}{21} = \cdot0476 \pm \cdot0038$ „ „ $\cdot0454$,

$\sigma = \cdot3780 \pm \cdot0066$ „ „ $\cdot3731$,

$\beta_2 = 3 - \frac{6}{7} = 2\cdot333 \pm \cdot012$ „ „ $2\cdot336$.

The equation calculated from the actual moments is $y = y_0\left(1 - \frac{x^2}{\cdot9802}\right)^{2\cdot021}$ whence the calculated range is 1·98, whereas it is known to be 2.

The following tables compare the actual distributions with those calculated from the equations.

Distribution of r from samples of 4 compared with the equation $y = 1\frac{1}{2}^{11}(1 - x^2)^3$.

	−1 to −·825	−·825 to −·675	−·675 to −·525	−·525 to −·375	−·375 to −·225	−·225 to −·075	−·075 to +·075	+·075 to +·225	+·225 to +·375	+·375 to +·525	+·525 to +·675	+·675 to +·825	+·825 to +1
Actual ...	64	45½	55½	67	59	62	63	58	60	64	51½	41½	54
Calculated	65	56	56	56	56	56	56	56	56	56	56	56	65
Difference	−1	−10½	−½	+11	+3	+6	+7	+2	+4	+8	−4½	−14½	−11

From this we get $\chi^2 = 13\cdot30$, $P = \cdot34$. It will however be noticed that the grouping has caused all the middle compartments to contain more than the calculated, as pointed out above.

Distribution of r from samples of 8 *compared with the equation*

$$y = \frac{750 \times 15}{16}(1 - x^2)^3.$$

	−1 to −·825	−·825 to −·675	−·675 to −·525	−·525 to −·375	−·375 to −·225	−·225 to −·075	−·075 to +·075	+·075 to +·225	+·225 to +·375	+·375 to +·525	+·525 to +·675	+·675 to +·825	+·825 to +1
Actual ...	2	27	44	60	96	114½	103	85	98½	65	37½	14½	3
Calculated	4½	20½	43	67	87	100½	105	100½	87	67	43	20½	4½
Difference	−2½	+6½	+1	−7	+9	+14	−2	−15½	+11½	−2	−5½	−6	−1½

whence $\chi^2 = 13\cdot94$, $P = \cdot30$.

In this case the grouping has had less influence and the largest contributions to χ^2 (in the second, sixth, eighth, and twelfth compartments) are due to differences of opposite sign on opposite sides, and may therefore be supposed to be entirely due to random sampling.

My equation then fits the two series of empirical results about as well as could be expected. I will now show that it is in accordance with the two theoretical cases n "large" and $n=2$, for $\sigma=\frac{1}{\sqrt{n-1}}$ which approximates sufficiently closely to Pearson and Filon's $\frac{1-r^2}{\sqrt{n}}$ when $r=0$ and n is large. Also when n is large β_2 becomes 3 and the distribution is normal.

And if $n=2$, the equation becomes $y=y_0(1-x^2)^{-1}$* where

$$y_0=\frac{N}{2\int_0^1(1-x^2)^{-1}dx}.$$

Put $x=\sin\theta$. Then $dx=\cos\theta\,d\theta$,

$$y_0=\frac{N}{2}\Big/\int_0^{\frac{\pi}{2}}\sec\theta\,d\theta=\frac{N}{2}\Big/\infty=0,$$

i.e. there is no frequency except where $(1-x^2)^{-1}$ is infinite, all the frequency is equally divided between $x=1$ and $x=-1$ which we know to be actually the case. Consequently I believe that the equation $y=y_0(1-x^2)^{\frac{n-4}{2}}$ probably represents the theoretical distribution of r when samples of n are drawn from a normally distributed population with no correlation. Even if it does not do so, I am sure that it will give a close approximation to it.

Let us consider Mr Hooker's limit of ·50 in the light of this equation. For 21 cases the equation becomes $\left.\begin{matrix}x=\sin\theta\\ y=y_0\cos^{17}\theta\end{matrix}\right\}$ and the proportion of the area lying beyond $x=\pm\,{\cdot}50$ will be

$$\frac{\int_{\theta=\sin^{-1}\cdot50}^{\theta=\frac{\pi}{2}}\cos^{18}\theta\,d\theta}{\int_0^{\frac{\pi}{2}}\cos^{18}\theta\,d\theta}.$$

I find this to be ·02099, or we may expect to find one case in 50 occurring outside the limits $\pm\,{\cdot}50$ when there is no correlation and the sample numbers 21.

* If a Pearson curve be fitted to the distribution whose moment coefficients are $\mu_2=1=\mu_4$ and $\mu_3=0$ we have $\beta_2=1$, $\beta_1=0$, hence the curve must be of Type II. and the equation is given by

$$y=y_0\left(1-\frac{x^2}{a^2}\right)^m \text{ where } a^2=\frac{2\mu_2\beta_2}{3-\beta_2}=1 \text{ and } m=\frac{5\beta_2-9}{2(3-\beta_2)} \text{ or } y=y_0(1-x^2)^{-1},$$

agreeing with the general formula.

When however there is correlation, I cannot suggest an equation which will accord with the facts, but as I have spent a good deal of time over the problem I will point out some of the necessities of the case.

(1) With small samples the value certainly lies nearer to zero than the real value of R, *e.g.*

samples of 2 : mean at $\frac{2}{\pi}\sin^{-1}R$,

samples of 4 (real value ·66) ·561* ± ·011,

samples of 8 (real value ·66) ·614† ± ·065.

But with samples of 30 (real value ·66) mean at ·6609 ± ·0067 shows that the mean value approaches the real value comparatively rapidly.

(2) The standard deviation is larger than accords with the formula $\frac{1-r^2}{\sqrt{n-1}}$ even if we give the mean value of r for samples of the size taken, *e.g.* for samples of 2,

$$\text{S.D.} = \sqrt{1-\left(\frac{2}{\pi}\sin^{-1}R\right)^2}.$$

For samples of 4, calculated‡ ·3957 ± ·0069; actual ·4680,

" " 8 " ·2355 ± ·0041; actual ·2684.

But samples of 30 calculated ·1046 ± ·0018, actual ·1001, again show that with samples as large as 30 the ordinary formula is justified.

(3) When there was no correlation the range found by fitting a Pearson curve to the distribution was accurately 2 in the theoretical case of samples of two, and well within the probable error for empirical distributions of samples of 4 and 8. But when we have correlation this process does not give the range closely for the empirical distribution (samples of 4 give 2·137, samples of 8 2·699, samples of 30 infinity) and the range calculated from samples of 2, which is

$$\frac{2\sqrt{4+3\mu_2+18\mu_2^2-9\mu_2^3}}{3+\mu_2},$$

$\left(\text{where } \mu_2 = 1-\left(\frac{2}{\pi}\sin^{-1}R\right)^2\right)$ is always less than 2 except in the case where μ_2 is 1, *i.e.* when there is no correlation.

Hence the distribution probably cannot be represented by any of Prof. Pearson's types of frequency curve unless $R = 0$.

(4) The distribution is skew with a tail towards zero.

* The value must be slightly larger than this (perhaps even by ·03) as Sheppard's corrections were not used.

† Again higher, but not by more than ·02.

‡ $\frac{1-r^2}{\sqrt{n-1}}$ where r is taken as the mean value for the size of the sample. If we took the real value R, the difference would be even greater.

(5) To sum up:—If $y = \phi(x, R, n)$ be the equation, it must satisfy the following requirements. If $R = 1$, 1 is the only value of x which gives the value of y other than zero. If $n = 2$, ± 1 are the only values of x to do so. If $R = 0$ the equation probably reduces to $y = y_0(1 - x^2)^{\frac{n-4}{2}}$.

Conclusions.

It has been shown that when there is no correlation between two normally distributed variables $y = y_0(1 - x^2)^{\frac{n-4}{2}}$ gives fairly closely the distribution of r found from samples of n.

Next, the general problem has been stated and three distributions of r have been given which show the sort of variation which occurs. I hope they may serve as illustrations for the successful solver of the problem.