RANK CORRELATION AND TESTS OF SIGNIFICANCE INVOLVING NO ASSUMPTION OF NORMALITY*†

By Harold Hotelling and Margaret Richards Pabst

1. Dependence of Tests of Significance on Normality

The powerful tests of significance, largely the work of R. A. Fisher, which have been revolutionizing statistical theory and practice, are in the main based on the assumption of a normal distribution in a hypothetical population from which the observations are a random sample. The nature and extent of the errors likely to result from the application of a test of significance assuming normality, where normality does not really exist, have been the subject of investigations both experimental and mathematical,[1] which however have not produced satisfactory substitutes for Fisher's methods. A false assumption of normality does not usually give rise to serious errors in the interpretation of simple means, since the distribution of a mean of any considerable number of cases is very nearly normal, no matter what the nature of the parent population, so long as it does not fall within a certain class having infinite range, and including the Cauchy distribution. The sampling distributions of second-order statistics are however more seriously disturbed by lack of normality, as is evident from their standard errors. For example the variance $(\mu_4 - \mu_2^2)/n$ of sample variances is much affected if μ_4/μ_2^2 differs considerably, as it often does, from the value 3 which it takes for a normal distribution. Likewise the approximate variance of the correlation coefficient,

$$\sigma_r^2 = \frac{1}{n\mu_{20}\mu_{02}}\left\{\mu_{22} + \frac{\mu_{40}\mu_{11}^2}{4\mu_{20}^2} + \frac{\mu_{04}\mu_{11}^2}{4\mu_{02}^2} - \frac{\mu_{31}\mu_{11}}{\mu_{20}} - \frac{\mu_{13}\mu_{11}}{\mu_{02}} + \frac{\mu_{22}\mu_{11}^2}{2\mu_{20}\mu_{02}}\right\},$$

* Research under a grant-in-aid from the Carnegie Corporation of New York.

† Presented to the American Mathematical Society at New York, Oct. 26, 1935.

[1] J. L. Carlson, *A Study of the Distribution of Means Estimated from Small Samples by the Method of Maximum Likelihood for Pearson's Type II Curve*, Unpublished M. A. Thesis, Leland Stanford Junior University, 1931.

Leone Chesire, Elena Oldis and Egon S. Pearson, *Further Experiments on the Sampling Distribution of the Correlation Coefficient*, Journal of the American Statistical Association, June, 1932, pp. 121–128.

Victor Perlo, *On the Distribution of Student's Ratio for Samples of Three Drawn from a Rectangular Distribution*, Biometrika, Vol. XXV, Parts I and II, May, 1933, pp. 203–204.

Paul R. Rider, *On the Distribution of the Ratio of Mean to Standard Deviation in Small Samples from Non-Normal Universes*, Biometrika, Vol. XXI, Parts I to IV, December, 1929, pp. 124–143.

H. L. Rietz, *Note on the Distribution of the Standard Deviation, etc.*, Biometrika, Vol. XXIII, 1931, pp. 424–426.

W. A. Shewhart and F. W. Winters, *Small Samples—New Experimental Results*, Bell Telephone Laboratories, Reprint B-327, July, 1928.

where μ_{ij} is the mean value of $x^i y^j$, and $\mu_{10} = \mu_{01} = 0$, may be substantially different from the value $(1 - \rho^2)^2/n$ commonly used, to which it reduces if the population has the bivariate normal distribution. It is however remarkable that if the variates are really independent, so that $\mu_{11} = 0$ and $\mu_{22} = \mu_{20}\mu_{02}$, this formula reduces to

$$\sigma_r^2 = \frac{1}{n}, \tag{1}$$

regardless of the form of the distribution. It should of course be remembered that these formulae give only the first term of an expansion in inverse powers of n, and also that the standard error fails for small samples to characterize the distribution adequately. But the sensitiveness of the standard error formula to deviations from normality in the population is a symptom of the grave dangers in using even those distributions which for normal populations are accurate, in the absence of definite evidence of normality.

To substitute in standard error formulae values of the higher moments estimated from the data does not meet the difficulty satisfactorily, since these higher moments are themselves subject to sampling errors which are often large, and since no exact distributions can ever be obtained in this way. The use of an arbitrary system of distributions such as the Pearson curves is subject to the same criticisms as that of the normal distribution. These and other special distributions may indeed be justified in special cases by general reasoning; an example of this in introducing a measure of relationship other than the correlation coefficient is to be found in the genetic discussion of Chapter 9 of Fisher's "Statistical Methods for Research Workers." But for a great deal of statistical work no such a priori reasoning is available and sufficient to specify a distribution in sufficient detail. If a specific form of distribution other than the normal can be relied on in a particular case, the mathematical problem of finding the exact distribution of the appropriate statistic will still commonly be found difficult or impossible.

2. Tests Independent of Normality Assumptions

A set of problems is thus encountered regarding the nature and methods of statistical inference possible without assuming any particular distribution of the variates in the population from which we have a sample. Tests of significance underlying such inferences must clearly be invariant under all transformations of each variate. We are thus forced to rely for our information on relations of *order*, or of qualitative classification, rather than upon magnitudes, excepting insofar as we can use inequalities such as that of Tchebycheff. Classification leads to the use of contingency tables, from which accurate probabilities are calculable for testing whether or not the two or more principles of cross-classification used are independent. If the probability obtained is so small as to render it incredible that independence exists, the further problem arises of measuring the degree of relationship; but in the absence of special assumptions, such as that

of the bivariate normal distribution, or those in Fisher's genetic example mentioned above, the problem of measuring degree of relationship is insoluble. Any measure of degree of relationship will change its value, unless this value corresponds to independence, when transformations other than those of a restricted class are applied to one of the variates. The problem of measuring *degree* of relationship, or correlation, is thus of quite a different character from that of testing the *existence* of a relationship, which is equivalent to absence of independence. The existence of correlation may be detected by methods of rank order or of classification; these can never, by themselves, be sufficient for its measurement.

To test the deviation of the center of a symmetrical population from some definite hypothetical value, Student's distribution, which is appropriate when the population is normal, may be replaced by the binomial distribution, which will sometimes show that the preponderance of cases on one side of the hypothetical value is too great to admit the hypothesis. Fisher applied this principle to Student's original example, showing at the same time that it can in certain cases be used to test the significance of the difference between the means of two samples.[2] Both this type of test and the use of contingency tables with grouped values of variates bring out clearly the fact that abandonment of the assumption of normality is equivalent to a certain loss of information, larger samples being required to make up for the lack of knowledge of the form of the population. The loss of information is greater for contingency tables arranged according to the values of the variates than when an appropriate method of rank correlation is used, for the contingency table may be regarded as derived from the ranks by grouping them, thus discarding some of the information.

We shall in §8 illustrate a combination of rank and contingency methods suitable for utilizing simultaneously two kinds of information contained in grouped data.

For large samples a method of treatment for which a great deal is to be said in many cases consists of replacing the observed variate by a new variate x to which a value is assigned for each individual or frequency class by interpolation in a table of the normal probability integral, in such a way that the distribution of x in the sample approximates normality. If this is done for each of two variates which do not have the bivariate normal distribution, the transformed values x and y may also lack the bivariate normal distribution, even approximately, though each is normally distributed, so far as we can speak of a sample as being normally distributed. Even if the bivariate distribution is normal, the correlation coefficient of x and y will not have the same distribution as the correlation coefficient in samples drawn from a bivariate normal distribution, since in the latter case the distributions of x and y separately would in most samples be less nearly normal than when the transformation to approximate normality is applied. From these considerations it follows that for the detection of correlation the normalizing transformation cannot be said in general to be the best

[2] R. A. Fisher, *Statistical Methods for Research Workers*, Art. 24, end.

method, even for large samples, though it may be a useful preliminary to the application of the method of least squares or to the use of correlation coefficients significantly different from zero in certain cases.

3. The Rank Correlation Coefficient

Suppose that n individuals are arranged in two orders with respect to two different attributes. Thus we might arrange a freshman class in order according to their grades in a language examination, and also according to their mathematical grades. As another example, we might be able to obtain ratings of various states with respect to penal law or practice, and also with respect to amount of crime. Continuous variates expressing these qualities are likely not to be normally distributed, so that the product-moment correlation coefficient r cannot be expected to have the exact distribution known for it in the case of samples from a normal population. We may therefore resort to the ranks, ignoring any exact values that have been assigned.

Calling X_i the rank of the ith individual with respect to one attribute, and Y_i his rank with respect to the other, so that $(X_1, X_2, \cdots, X_n)$ and $(Y_1, Y_2, \cdots, Y_n)$ are two permutations of the numbers $(1, 2, \cdots, n)$, let us put $x_i = X_i - \bar{x}$, $y_i = Y_i - \bar{y}$, where

$$\bar{x} = \bar{y} = \frac{n+1}{2}$$

The rank correlation coefficient is defined as

$$r' = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}, \tag{2}$$

the sums being over the n values in the sample. Now since the sum of the first n integers is $n(n+1)/2$, and the sum of their squares is $n(n+1)(2n+1)/6$, we have

$$\begin{aligned} \Sigma x^2 = \Sigma (X - \bar{x})^2 &= \Sigma X^2 - (\Sigma X)^2/n \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n^3 - n}{12}, \end{aligned} \tag{3}$$

and Σy^2 has the same value. Also, if we put d_i for the difference between the two ranks for the ith individual, so that

$$d_i = X_i - Y_i = x_i - y_i,$$

we have

$$\Sigma d^2 = \Sigma x^2 - 2\Sigma xy + \Sigma y^2 = \frac{n^3 - n}{6} - 2\Sigma xy.$$

Substituting in (2) the value of Σxy found from this equation, and also the values just obtained for Σx^2 and Σy^2, we have:

$$r' = 1 - \frac{6\Sigma d^2}{n^3 - n}. \tag{4}$$

This is the most convenient formula for computing r'.

Compared with certain other tests of correlation based on order, such as $\Sigma |d|$, or the number of inversions required to pass from one permutation of the n numbers to the other, r' appears to be a sensitive index of relationship, since for a given value of n it possesses a greater number of distinct values. But to assert without qualification that r' or any other statistic is the best possible test of correlation based on order relations alone would be meaningless. Indeed, a particular type of bivariate distribution might well have a parameter representing correlation whose significance could best be detected by a test adapted only to this particular bivariate distribution. However the rank correlation coefficient has properties that point to its value in more general use than it has heretofore received. It has been regarded chiefly as a more easily calculable substitute for the product-moment coefficient r. Karl Pearson has remarked that the rank correlation coefficient is the easier to compute for samples smaller than approximately forty, while r involves less labor for larger samples.

The great value of the rank correlation coefficient appears to us to consist in its use as a test of the existence of correlation, a test capable of exact interpretation in terms of probability, without any assumption of a normal or other special bivariate distribution. If a bivariate distribution is specified by $f(x, y)\,dx\,dy$, the condition of independence is that $f(x, y)$ shall be the product of a function of x by a function of y. If we put

$$\xi = \int_{-\infty}^{\infty}\int_{0}^{x} f(x', y')\,dx'\,dy', \qquad \eta = \int_{0}^{y}\int_{-\infty}^{\infty} f(x', y')\,dx'\,dy', \tag{5}$$

using the inner integral sign in each case to correspond to the inner differential, then each of the quantities ξ and η is distributed with uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$; and if x and y are independent, then ξ and η are also independent. The correlation ρ' of ξ with η may be called the rank correlation of x and y in the population. It will vanish in case of independence. It is for this case that we shall obtain in §§5, 6 and 7 the exact probability test for r' in small samples, the exact standard error and fourth moment, and asymptotic values for the higher moments, with a demonstration that, for sufficiently large samples, r' can be treated as normally distributed. In §9 we shall present, in a revised and simplified form, certain work of Karl Pearson relative to the estimation of the correlation ρ in a bivariate normal distribution, and apply the results to discuss the question of the importance of the lost information when measurements are replaced by ranks.

4. History of Rank Correlation Theory

Rank correlation seems to have had its origin in the method of representing the distribution of a variate by grades or percentiles introduced by Francis

Galton.[3] Later Spearman[4] proposed that rank be considered in place of the variate, and suggested that the correlation of ranks be used as a measure of the degree of dependence of the variates. Spearman also introduced the "footrule of correlation" based on $\Sigma \mid d \mid$.

The principal memoir on rank correlation is by Karl Pearson.[5] Assuming an underlying normal distribution, Pearson obtains a relation equivalent to

$$\rho = 2 \sin \frac{\pi}{6} \rho' , \tag{6}$$

where ρ is the correlation of x and y in the population, and ρ' is the correlation of uniformized variates ξ and η defined by (4). An estimate r'' of ρ may be based on the rank correlation r', in accordance with (6), by writing

$$r'' = 2 \sin \frac{\pi}{6} r' . \tag{7}$$

Pearson finds the first few terms of infinite series giving the standard errors of r' and r''. He deals similarly with the estimation of correlation by means of $\Sigma \mid d \mid$. The paper contains a neat proof, attributed to Student, of the probable error of r' under conditions of independence. It was this proof that suggested the analysis of §§6 and 7 below. This long memoir is very difficult to read and interpret accurately, owing chiefly to the failure to distinguish clearly between sample and population.

The use of the probable error formulae is valid only if the distributions of r' and r'' are sensibly normal. The question of approximate normality thus raised is investigated for the first time in the present paper. In order to use these formulae it is necessary to assume not only (1) that the underlying population has the bivariate normal distribution (an assumption which requires more than that each variate be normally distributed), (2) that the first few terms of the infinite series are enough, and (3) that the distributions of r' and r'' are practically normal, but also (4) that sample values can be put for population values in the formulae, or that population values are known independently or can be assumed. It is probably this last condition that has been least understood and has led to the greatest number of false conclusions regarding the significance of data.

A note by W. C. Eells[6] presents a compilation of numerous textbook versions of the probable errors of r' and r'', all differing from each other and from Pear-

[3] Francis Galton, *Natural Inheritance*, Macmillan, 1889, Chaps. 4 and 5.

[4] C. Spearman, *The Proof and Measurement of Association Between Two Things*, American Journal of Psychology, Vol. 15, 1904.

[5] Karl Pearson, *On Further Methods of Determining Correlation*, Drapers' Company Research Memoirs, Biometric Series IV, Mathematical Contributions to the Theory of Evolution, XVI, London, Dulau, 1907.

[6] W. C. Eells, *Formulas for Probable Errors of Coefficients of Correlation*, Journal of the American Statistical Association, Vol. 24, 1929, p. 170.

son's. Taking Pearson's formulae as correct, without discussing the assumptions implicit in their use, Eells presents a table for calculating the probable errors of r, r' and r''.

5. Significance of Rank Correlation in Small Samples

If the variates are independent we may without loss of generality assign the values 1, 2, $\cdots$, n in order to $X_1, X_2, \cdots, X_n$, and regard the Y's as made up by any one of the $n!$ permutations of these numbers, all permutations being equally probable. The probability of any particular value of r' is thus proportional to the number of permutations giving rise to this value. These may be enumerated with the help of (4). Thus for $n = 2$, each of the values ± 1 has the probability $\frac{1}{2}$. For $n = 3$, the possible values of r' are $-1, -\frac{1}{2}, \frac{1}{2}, 1$, with respective probabilities 1/6, 1/3, 1/3, 1/6. For $n = 4$ the values 1, 4/5, 3/5, 2/5, 1/5, 0 have the respective probabilities 1/24, 1/8, 1/24, 1/6, 1/12, 1/12.

From (2) it is evident that the distribution of r' in case of independence is symmetrical, since each permutation is exactly as probable as that of directly opposite order, and since a change of sign of all the x's or y's changes the sign of r' without affecting its absolute value. It is clear also that the values $r' = \pm 1$, corresponding to the two variates being in the same or opposite orders, are the extreme ones, and have each a probability $1/n!$. The next greatest value of $|r'|$ corresponds to the interchange of two consecutive individuals, who may be selected in $n - 1$ ways and makes $\Sigma d^2 = 2$. Thus the values $\pm(1 - 12/[n^3 - n])$ occur with probability $(n - 1)/n!$ each. Next to these, corresponding to $\Sigma d^2 = 4$, are the values $\pm(1 - 24/[n^3 - n])$, whose probabilities are each $(n - 2)(n - 3)/2(n!)$, since the numbers of pairs of mutually exclusive consecutive pairs in a sequence of n is $(n - 2)(n - 3)/2$. In like manner, but with greater complexity, it appears that the probability of the value $1 - 36/[n^3 - n]$ is $\dfrac{(n - 3)(n - 4)(n - 5) + 12(n - 2)}{6(n!)}$. Easy calculation from these results shows that, if we require for significance a probability $P = .01$ of a value of $|r'|$ as great as or greater than the value observed, then for samples of 5 it is impossible to obtain a significant value; for $n = 6$, significance requires that $r' = \pm 1$; and for $n = 7$ the significant values of $|r'|$ are 25/28 and more. For the less stringent standard $P = .05$, a unit correlation only is significant in a sample of 5; while 29/35 is not, but 31/35 is, significant in a sample of 6.

6. The Standard Error and Fourth Moment

For large samples the exact calculation of probabilities becomes very laborious, and we are forced to resort to approximations. The first step in the available approximations is the determination of the standard deviation of the distribution. The square of this quantity, the second moment or variance of r', may, since the mean value of r' in case of independence is zero, be written

$$\sigma_{r'}^2 = \mu_2 = Er'^2,$$

the symbol E denoting the expectation or mean value of the quantity following. The operation E has the properties that the expectation of a sum is the sum of the expectations of the terms, the expectation of the product of *independent* variates is the product of their expectations, and the expectation of the product of a constant by a variate is the product of the constant by the expectation of the variate. It is particularly to be noted that the first of these properties holds whether the terms of the sum are mutually independent or not.

From (2) and (3) we have

$$r' = \frac{12 \Sigma xy}{n^3 - n}. \tag{8}$$

Now we may regard $x_1, x_2, \cdots, x_n$ as taking the same values in all samples, these values being centered at zero and differing consecutively by unity. The y's are then variates, not independent of each other, taking this same set of values, but in a manner varying from sample to sample by chance. For any particular y, for example that associated with x_1, the chance distribution has moments of the form

$$Ey^p = \frac{\Sigma x^p}{n} = \frac{\Sigma y^p}{n} = \frac{s_p}{n}, \tag{9}$$

if we denote by s_p the sum of the pth powers of the n numbers differing consecutively by unity and centered at zero. It is clear that, for every odd value of p, $s_p = 0$. Also, from (3),

$$s_2 = \frac{n^3 - n}{12}.$$

In view of these facts, we have from (8),

$$\sigma_{r'}^2 = Er'^2 = \frac{E(\Sigma xy)^2}{s_2^2} = \frac{\Sigma x^2 Ey^2 + 2 \Sigma x_1 x_2 E y_1 y_2}{s_2^2},$$

where $\Sigma x_1 x_2$ stands for the sum of all the $n(n-1)/2$ *different* terms obtained by permuting the subscripts. We have

$$E y_1 y_2 = \frac{2 \Sigma x_1 x_2}{n(n-1)};$$

also

$$2 \Sigma x_1 x_2 = s_1^2 - s_2 = -s_2.$$

Combining these results we have:

$$\sigma_{r'}^2 = \frac{1}{s_2^2}\left\{\frac{s_2^2}{n} + \frac{s_2^2}{n(n-1)}\right\} = \frac{1}{n-1}. \tag{10}$$

This is the formula obtained by Student and incorporated in Pearson's memoir.

Any desired moment of r' may be obtained in this manner. However the complexity of the calculation increases rapidly with the order of the moment, and the derivation of even the fourth moment is too long to be included in this paper. The value obtained for the fourth moment is

$$\mu_4 = \frac{3(25n^4 - 13n^3 - 73n^2 + 37n + 72)}{25n(n+1)^2(n-1)^3}.$$

It will be observed immediately that the kurtosis, $\beta_2 = \mu_4/\mu_2^2$, approaches the normal value 3 as n increases.

For values of n which are not small enough for the exact probabilities to be computed easily, the Tchebycheff inequality,

$$P \leqq \frac{1}{(n-1)r'^2}, \tag{11}$$

where P is the probability of a deviation exceeding r', will often be of service. Thus, if $n = 25$ and $r' = .9$, (11) shows that P is less than .05, so that the evidence for existence of a relationship should by an ordinary standard be regarded as significant. However this does not in general give an accurate approximation to P, nor do the similar inequalities involving the higher moments.

7. The Higher Moments and the Approach to Normality

A general moment of r' of even order is defined by

$$\mu_{2\alpha} = E\, r'^{2\alpha} = \frac{1}{s_2^{2\alpha}} E\,(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}. \tag{12}$$

When the parenthesis is expanded we may take the expectation term by term, regarding the x's as constants. Now

$$E\, y_1^{2\alpha} = \frac{\Sigma\, x_1^{2\alpha}}{n}, \qquad E\, y_1^{2\alpha-1} y_2 = \frac{\Sigma\, x_1^{2\alpha-1} x_2}{n(n-1)},$$

and so forth, the sums on the right in the numerators being symmetric functions of the constants x, taken over all different terms obtained from that written by permuting subscripts, and the denominator being in each case the number of terms in the numerator. Thus

$$\mu_{2\alpha} = \frac{1}{s_2^{2\alpha}}\left\{\frac{(\Sigma\, x_1^{2\alpha})^2}{n} + A\,\frac{(\Sigma\, x_1^{2\alpha-1} x_2)^2}{n(n-1)} + B\,\frac{(\Sigma\, x_1^{2\alpha-2} x_2x_3)^2}{n(n-1)(n-2)} + \cdots\right\}, \tag{13}$$

where the coefficients A, B, $\cdots$ depend on α but not on n. With a view to determining the leading term in the expansion of $\mu_{2\alpha}$ in powers of n^{-1}, we shall select the term in the curly brackets in (13) of highest degree, meaning by the degree of one of these rational fractions the excess of the degree of the numerator over that of the denominator.

The symmetric functions are well known to be expressible as polynomials in

the power-sums s_p. In each term of such a polynomial corresponding to one of our symmetric function of degree 2α, the sum of the subscripts of the s_p's must be 2α, since if all the x's are multiplied by a constant such a polynomial must be multiplied by the 2αth power of the constant. Now s_p is a polynomial of degree $p + 1$ in n, if n is even, but vanishes identically if n is odd. Consequently the degree in n of any of the terms of the polynomial in the power-sums must exceed 2α by the number of power-sums appearing in this term. Therefore, the term of highest degree in n obtained, when one of the symmetric functions is expressed in terms of the s_p's and thence in terms of n, must contain the greatest possible number of the s_p's. If p is the number of distinct x's in a term of one of our symmetric functions, this function may be written in the form

$$\begin{aligned}\Sigma x_1^{a_1} x_2^{a_2} \cdots x_p^{a_p} &= c_0 s_{a_1} s_{a_2} \cdots s_{a_{p-1}} s_{a_p} - c_1 s_{a_1+a_p} s_{a_2} \cdots s_{a_{p-1}} \\ &\quad - c_2 s_{a_1} s_{a_2+a_p} \cdots s_{a_{p-1}} - \cdots - c_{p-1} s_{a_1} s_{a_2} \cdots s_{a_{p-1}+a_p} \\ &\quad - c' s_{a_1+a_2+a_p} s_{a_3} \cdots s_{a_{p-1}} - \cdots ,\end{aligned} \tag{14}$$

where $a_1 + a_2 + \cdots + a_p = 2\alpha$, and the c's do not involve n. In the right-hand member of the equation above, the first term involves p of the power-sums, while the remaining terms involve fewer of them. Hence, if all the indices $a_1, a_2, \cdots, a_p$ are even, the first term is a polynomial of degree $2\alpha + p$ in n, while the remaining terms are polynomials of lower degree, and are therefore negligible in comparison with the first term when n is sufficiently large. But if any of the indices a_i are odd, the first term vanishes identically, and the degree of (14), regarded as a polynomial in n, is then less than $2\alpha + p$. Since the sum of the indices is 2α, the number of odd ones among them must be even; let this number be denoted by $2q$, and let the number of even indices be m. Then $p = m + 2q$. The terms of highest degree in the right-hand member of (14) must be obtained by grouping the odd indices in pairs to form the subscripts of the s's. The degree is therefore $2\alpha + m + q$.

In (13), the degree of the denominator of each term in the curly brackets is the number of distinct x's appearing in a term of the symmetric function in the numerator, namely p, or $m + 2q$. Hence the excess of the degree of the numerator over that of the denominator is

$$2(2\alpha + m + q) - (m + 2q) = 4\alpha + m .$$

This will be a maximum when m is a maximum, and is independent of q. The maximum value of m is α, and occurs only for the symmetric function

$$\Sigma x_1^2 x_2^2 \cdots x_\alpha^2 . \tag{15}$$

The term involving this function is therefore the only one in the right-hand member of (13) we need consider. Since this symmetric function contains $n(n-1)(n-2)\cdots(n-\alpha+1)/(\alpha!)$ terms, and since in the expansion of

$$(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^{2\alpha}$$

the coefficient of $x_1^2x_2^2 \cdots x_\alpha^2y_1^2y_2^2 \cdots y_\alpha^2$ is, by the multinomial theorem $(2\alpha)!/2^\alpha$, we have from (13),

$$\mu_{2\alpha} \sim \frac{1}{s_2^{2\alpha}} \frac{(2\alpha)!}{2^\alpha} \frac{\alpha!\,(\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2)^2}{n^\alpha}.$$

To evaluate the symmetric function (15), so far as the term of highest order in, n is concerned, we of course need only the first term of (14), which reduces in this case to

$$\Sigma\, x_1^2x_2^2 \cdots x_\alpha^2 = c_0 s_2^\alpha - \cdots.$$

In the expansion of $s_2^\alpha = (x_1^2 + x_2^2 + \cdots + x_n^2)^\alpha$, the coefficient of (15) is $\alpha!$, which is therefore the reciprocal of c_0. Thus we obtain

$$\mu_{2\alpha} = \frac{(2\alpha)!}{\alpha!\,2^\alpha}\left[\frac{1}{n^\alpha} + \cdots\right],$$

the terms dropped being of higher order in n^{-1}.

The 2αth moment of the quotient of r' by its standard error, that is, of $r'\sqrt{n-1}$, is $(n-1)^\alpha$ times that of r', and therefore approaches, as n increases, the value

$$\frac{(2\alpha)!}{\alpha!\,2^\alpha}. \tag{16}$$

The odd moments are all zero because of the symmetry of the distribution of r'. But (16) is the moment of order 2α of a normal distribution of unit variance and zero mean. It follows therefore from the Second Limit Theorem of Probability[7] that the distribution tends to normality as n increases; that is, for any real number λ, the limit as n tends to infinity of the probability that $r'\sqrt{n-1} < \lambda$ is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{1}{2}x^2}\, dx.$$

The normality of the limiting distribution of the rank correlation coefficient is rather remarkable, since r', unlike the product-moment correlation coefficient r and other statistics in common use, is neither a mean of independent quantities nor a function of such means, so that the ultimate normality just established is not a corollary of known general theorems. It is unexpected also because the exact distribution of r' for samples smaller than six might lead one to anticipate a bimodal distribution.

An outstanding problem is to determine whether the distribution of r' in samples from a bivariate normal distribution for which $\rho \neq 0$ converges to normality. Without such an approach to normality, the probable error formulae

[7] First proved by Markoff. Cf. Fréchet and Shohat, *A Proof of the Generalized Second Limit Theorem in the Theory of Probability*, Transactions of the American Mathematical Society, Vol. 33, 1932, pp. 533-543.

discovered by Pearson are useless. Another problem is to find convenient and accurate approximations to the distribution of r', for moderate values of n, with close limits of error. A table calculated along the lines suggested in §5 would be very useful.

8. Combination of Rank and Contingency Methods

Suppose that a thousand school children are examined at the end of a course of instruction, and rated with the grades A, B, C and D. Five hundred of these children are of each sex. The results are:

	A	B	C	D	Totals
Boys	190	200	80	30	500
Girls	220	200	60	20	500
Totals	410	400	140	50	1000
Proportion of Girls	537	500	429	400	500

Regarding this as a 2 x 4 contingency table with three degrees of freedom, we calculate $\chi^2 = 7.52$, the probability of which value being exceeded by chance is .0570. The indications of a significant difference in distribution of grades between sexes may thus, if one holds to the .05 standard and uses only the χ^2 test, be regarded as not quite significant. There is, however, additional evidence in the fact that the proportion of girls diminishes steadily as we pass down the scale of grades. If we treat excellence in the subject as one variate and the proportion of girls in a group as another, we have a rank correlation of unity, with a sample of four. The probability of a correlation of ± 1 is .083, which also, by itself, would not be considered significant. But we may combine the two pieces of evidence by the method given by Fisher.[8] The process consists of adding the natural logarithms of the two probabilities, doubling, and treating the result as having the χ^2 distribution with four degrees of freedom. This gives a probability in the neighborhood of .03, which would be judged significant.

Similar cases are very common. The value of χ^2 is unchanged if the columns are permuted in any way, whereas r' depends solely on which of the possible permutations actually exists. Thus the two tests are *independent*, a property needed for the combination by the above method.

9. Efficiency of Replacement of Measures by Ranks, and the Estimation of ρ from Rank Correlation, for a Normal Population

Consider a population with a normal distribution in two variates x and y, each of which we shall without loss of generality assume to be of unit variance and zero mean. The density distribution is then specified by $z\,dx\,dy$, where

[8] R. A. Fisher, *Statistical Methods for Research Workers*, 4th and 5th editions, Art. 21.1.

$$z = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)}, \tag{17}$$

where ρ is the correlation of x and y, or the variate correlation. By ξ and η, as in §3, we denote the uniformized variates defined by (5), i.e., functions respectively of x and y having distributions of uniform density from $-\frac{1}{2}$ to $+\frac{1}{2}$. Then ξ and η will each have the variance 1/12. The rank correlation ρ' in the population is the correlation of ξ and η; consequently

$$\rho' = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\,\eta\,z\,dx\,dy\,. \tag{18}$$

Thus ρ' is a function of ρ, which obviously vanishes when $\rho = 0$.

From (17) the identity

$$\frac{\partial z}{\partial \rho} = \frac{\partial^2 z}{\partial x\,\partial y} \tag{19}$$

is readily calculated. With its help we have from (18) and integrations by parts,

$$\begin{aligned}\frac{d\rho'}{d\rho} &= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta\frac{\partial z}{\partial\rho}\,dx\,dy = 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\xi\eta\frac{\partial^2 z}{\partial x\,\partial y}\,dx\,dy \\ &= 12\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{d\xi}{dx}\frac{d\eta}{dy}\,z\,dx\,dy\,.\end{aligned} \tag{20}$$

Now since x and y are normally distributed with unit variance and zero means, the uniformized variates (5) take the form

$$\xi = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}\,dt\,, \qquad \eta = \frac{1}{\sqrt{2\pi}}\int_0^y e^{-\frac{t^2}{2}}\,dt\,.$$

Therefore

$$\frac{d\xi}{dx} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \qquad \frac{d\eta}{dy} = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}.$$

Substituting these values and (17) in the last integral in (20) we have,

$$\frac{d\rho'}{d\rho} = \frac{12}{4\pi^2\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{(2-\rho^2)x^2-2\rho xy+(2-\rho^2)y^2}{2(1-\rho^2)}}\,dx\,dy\,.$$

The double integral, as is well known, equals π divided by the square root of the discriminant of the quadratic form in the exponent. This gives

$$\frac{d\rho'}{d\rho} = \frac{6}{\pi\sqrt{4-\rho^2}}.$$

Therefore, since ρ' vanishes with ρ,

$$\rho' = \frac{6}{\pi} \sin^{-1} \frac{\rho}{2},$$

or

$$\rho = 2 \sin \frac{\pi \rho'}{6}.$$

This is essentially the process used by Pearson.

The last equation suggests that an estimate r'' of ρ be based on the rank correlation r' by means of the relation

$$r'' = 2 \sin \frac{\pi r'}{6}.$$

Prefixing a δ to denote a deviation of sample from population value we have by a Taylor expansion,

$$\delta r'' = \frac{\pi}{3} \cos \frac{\pi \rho'}{6} \delta r' + \cdots,$$

the terms dropped being of higher order in $\delta r'$ than those written, and consequently of higher order in n^{-1}. Squaring, taking the expectation, and ignoring the terms of higher order, we have for the case $\rho = \rho' = 0$, by (10),

$$\sigma_{r''}^2 = E(\delta r'')^2 = \frac{\pi^2}{9} \sigma_{r'}^2 = \frac{\pi^2}{9(n-1)},$$

approximately.

The last result enables us to measure the loss of information, at least for large samples, that results from neglecting the exact values of the variates and using only ranks. The product-moment correlation coefficient r has, if $\rho = 0$, the exact variance

$$\frac{1}{n-1},$$

the ratio of which to $\sigma_{r'}^2$ tends as n increases to $9/\pi^2$. Thus the efficiency of the rank correlation method in estimating ρ, if ρ is really zero, is $9/\pi^2 = .9119$. This means that the product-moment correlation is approximately as sensitive a test of the existence of a relationship in a normally distributed population with 91 cases as the rank correlation with 100 cases.

The efficiency of r' will of course be different for non-normal populations, and also for normal populations with $\rho \neq 0$. But if the form of the population is known, this knowledge may always be used to supplement the ranks to obtain a more accurate estimate of correlation, or test of relationship. This fact deserves some attention, since a superficial observation of the coincidence of the formula (1)

for the leading term of the variance of an arbitrary uncorrelated population, and the leading term of the formula (10) for the variance of the rank correlation, might suggest that r' is as accurate as r. But it may be surmised that the 9 % loss of information found for the bivariate normal distribution is the greatest loss of information in using r' in place of r to test for independence, since for non-normal populations the most efficient estimate of the correlation will not usually be r, but a more complicated function of the observations. Certainly where there is complete absence of knowledge of the form of the bivariate distribution, and especially if it is believed not to be normal, the rank correlation coefficient is to be strongly recommended as a means of testing the existence of relationship.

COLUMBIA UNIVERSITY.