DISTRIBUTIONS OF SUMS OF SQUARES OF RANK DIFFERENCES FOR SMALL NUMBERS OF INDIVIDUALS[1]

By E. G. Olds

I. INTRODUCTION

In a recent article,[2] reporting the results of research under a grant-in-aid from the Carnegie Corporation of New York, Hotelling and Pabst have given a comprehensive treatment of the theory and application of rank correlation and have contributed significantly to existing knowledge on the subject. It is not the purpose of this note to evaluate their contribution but to attempt the solution of a problem they suggest.

In §3[3] they have given the well-known formula for rank correlation, $r' = 1 - \frac{6\Sigma d^2}{n^3 - n}$, where n is the number of individuals ranked and $\Sigma d^2 = \sum_{i=1}^{n} d_i^2$ (d_i being the rank difference for the ith individual). In §5 the question of the significance of r' in small samples has been considered from the following point of view; if the value of r', obtained from a comparison of the ranks of n individuals as a possible measure of the relation between two attributes, is such that there exists a high probability that it could have occurred by virtue of a chance rearrangement of the n individuals, then the value of r' does not furnish a significant indication of relationship. Then one test of the significance of a particular value of r' is to note whether it has a probability less than P (P equal to .01 or, less stringently, equal to .05) of occurring because of a chance re-ranking.

To apply this test it is necessary to have some information regarding the distribution of r' for the chance rearrangements of the numbers from 1 to n. Hotelling and Pabst have given the distribution of r' for the cases, $n = 2, 3, 4$. They have noted that the distribution is symmetrical for each value of n and that it has a range from -1 to 1. From a consideration of the probabilities corresponding to $\Sigma d^2 = 0, 2, 4, 6$, they have discussed the significance of values of r' for $n = 5, 6, 7$. In §8 they have stated, "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." This statement, along with the interest manifested by others in private communications, has led to the investigation reported in this paper.

[1] Presented to the American Mathematical Society, December 29, 1936.

[2] Harold Hotelling and Margaret Pabst, *Rank Correlation and Tests of Significance Involving No Assumption of Normality*, Annals of Mathematical Statistics, Vol. VII, 1936, pp. 29-43.

[3] Loc. cit.

II. EXACT DISTRIBUTION OF SUMS OF SQUARED DIFFERENCES

In the paper mentioned above, the authors have given the exact probabilities for all possible values of r' for $n = 2, 3$, and 4. Since r' is a linear function of Σd^2 for any particular value of n, there is a one-to-one correspondence between values of Σd^2 and values of r'. For example, for the case of $n = 3$, we have the following:

$$\begin{array}{rcccc} \Sigma d^2 = & 0 & 2 & 6 & 8 \\ r' = & 1 & \frac{1}{2} & -\frac{1}{2} & -1 \\ p = & \frac{1}{3!} & \frac{2}{3!} & \frac{2}{3!} & \frac{1}{3!} \end{array}$$

where p represents the relative frequency of r' or of Σd^2. Therefore it seems pertinent to investigate the distribution of Σd^2 for various values of n.

If n individuals are ranked $1, 2, 3, \cdots n$, by one criterion and then are reranked at random there are $n!$ possibilities for the new ranking. Let us consider the differences between the numbers in the new and in the original rankings. Suppose these differences are represented by $d_1, d_2, \cdots d_n$. Then it is apparent that $\sum_{i=1}^{n} d_i^2 = 0$. If we let $a_1, a_2, \cdots a_k$ represent an arrangement for $n = k$, insert $k + 1$ after a_k and advance the cycle one position at a time, we have the following arrangements for the case, $n = k + 1$:

$$\begin{array}{cccccccc} a_1, & a_2, & a_3, & \cdot & \cdot & \cdot & a_k, & k+1 \\ a_2, & a_3, & a_4, & \cdot & \cdot & \cdot & k+1, & a_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_k, & k+1, & a_1, & \cdot & \cdot & \cdot & a_{k-2}, & a_{k-1} \\ k+1, & a_1, & a_2, & \cdot & \cdot & \cdot & a_{k-1}, & a_k \end{array} \tag{1}$$

Now, for $n = k$, $d_1 = a_1 - 1$, $d_2 = a_2 - 2$, $\cdots d_k = a_k - k$. If we list the differences for the $k + 1$ derived arrangements, we have

$$\begin{array}{cccccccc} d_1, & d_2, & d_3, & \cdot & \cdot & \cdot & d_k, & 0 \\ d_2+1, & d_3+1, & d_4+1, & \cdot & \cdot & \cdot & 1, & d_1-k \\ d_3+2, & d_4+2, & d_5+2, & \cdot & \cdot & 2, & d_1+1-k, & d_2+1-k \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ d_k+k-1, & k-1, & d_1-2, & \cdot & \cdot & \cdot & \cdot & d_{k-1}-2 \\ k, & d_1-1, & d_2-1, & \cdot & \cdot & \cdot & \cdot & d_k-1 \end{array} \tag{2}$$

It is apparent that each row of differences is formed as follows: the entry in the first column is formed by adding 1 to the entry in column two in the row above, the entry in the second column is obtained by adding 1 to the entry in the third column in the row above, and so on until we come to the entry in the last column which is obtained by subtracting k from the entry in the first column of the preceding row.

If we form the sum of squares of the entries in each row we observe an interesting property of the set; the sums are all congruent, modulus $(k + 1)$. Let us write the sums, denoting them by $S_1, S_2, \cdots S_{k+1}$. Also let d_{ij} represent the entry in the ith row and jth column. Then

$$S_i = \sum_{j=1}^{k+1} d^2_{i,j}$$

$$S_{i+1} = \sum_{j=1}^{k+1} d^2_{i+1,j} = \sum_{j=2}^{k+1} (d_{i,j} + 1)^2 + (d_{i,1} - k)^2$$

$$= \sum_{j=1}^{k+1} (d_{i,j} + 1)^2 + (d_{i,1} - k)^2 - (d_{i,1} + 1)^2 \tag{3}$$

$$= \sum_{j=1}^{k+1} (d^2_{i,j} + 2d_{i,j} + 1) - (2d_{i,1} - k + 1)(k + 1)$$

$$= S_i + 0 + (k + 1) - (2d_{i,1} - k + 1)(k + 1)$$

$$= S_i + (k - 2d_{i,1})(k + 1)$$

Noticing that $d_{i,1} = d_1 + i - 1$, for $i = 1, 2, \cdots k$, and $d_{k+1,1} = k$, we have

$$\begin{aligned} S_2 &= S_1 + (k - 2d_1)(k + 1) \\ S_3 &= S_2 + (k - 2d_2 - 2)(k + 1) \\ S_4 &= S_3 + (k - 2d_3 - 4)(k + 1) \\ &\vdots \\ S_{k+1} &= S_k + (k - 2d_k - 2k + 2)(k + 1) \\ S_{k+2} &= S_{k+1} + (k - 2k)(k + 1) = S_{k+1} - k(k + 1) \end{aligned} \tag{4}$$

Of course, $S_{k+2} = S_1$, as the $(k + 2)$nd row is identical with the first and the set is closed. So we may write

$$S_{k+1} = S_1 + k(k + 1) \tag{5}$$

The analysis given above not only establishes the congruence of the sums, modulus $(k + 1)$, but also indicates a method of deriving the sums for $n = k + 1$ from the sums for $n = k$, since $S_1 = \sum_{i=1}^{k} d_i^2$. It is also worth noticing that S_{i+1} depends not only on S_i (and therefore on S_1) but also on $d_{i,1}$ (and therefore on d_{i-1}).

Another matter needs attention. It is the relation between the sums of squares of deviations for a particular order and for the reverse order. Let a_1, a_2, $\cdots a_n$ be a particular arrangement. Then the reverse order is a_n, a_{n-1}, $\cdots a_1$. The sums of the squares of the deviates are, respectively,

$$S = (a_1 - 1)^2 + (a_2 \quad - 2)^2 + \cdots (a_k - k)^2$$

and

$$\bar{S} = (a_k - 1)^2 + (a_{k-1} - 2)^2 + \cdots (a_1 - k)^2 \tag{6}$$

Then

$$\begin{aligned} S + \bar{S} &= [(a_1 - 1)^2 + (a_1 - k)^2] + [(a_2 - 2)^2 + (a_2 - k + 1)^2] \\ &\qquad + \cdots [(a_k - k)^2 + (a_k - 1)^2] \\ &= \sum_{r=1}^{k} [(a_r - r)^2 + (a_r - k + r - 1)^2] \\ &= \sum_{r=1}^{k} [(a_r - r) + (a_r - k + r - 1)]^2 - 2 \sum_{r=1}^{k} (a_r - r)(a_r - k + r - 1) \\ &= 4 \sum_{r=1}^{k} a_r^2 - 4(k+1) \sum_{r=1}^{k} a_r + (k+1) \sum_{r=1}^{k} 1 - 2 \sum_{r=1}^{k} a_r^2 \\ &\qquad + 2(k+1) \sum_{r=1}^{k} a_r - 2(k+1) \sum_{r=1}^{k} r + 2 \sum_{r=1}^{k} r^2. \end{aligned}$$

Noting that $\Sigma a_r^2 = \Sigma r^2$ and $\Sigma a_r = \Sigma r$, we readily obtain the result[4]

$$S + \bar{S} = \frac{k^3 - k}{3} \tag{7}$$

It is now apparent that the sums range from 0 to $\dfrac{k^3 - k}{3}$ with a mean of $\dfrac{k^3 - k}{6}$.

As the exact frequencies for sums of squares do not seem to be available, it seems useful to compute them for certain small values of n and, at the same time

[4] The geometric representation of the problem may be of some interest. Let the coordinates of point R, in Euclidean n-space be $(1, 2, 3, \cdots n)$, the coordinates of $\bar{R}$ be $(n, n - 1, \cdots 2, 1)$, and the coordinates of P be $(x_1, x_2, \cdots x_n)$. Let us restrict the x's to be the numbers $(1, 2, 3, \cdots n)$, but not necessarily in the order given, i.e. the locus of P is a set of $n!$ points, corresponding to the permutations of the numbers $1, 2, 3, \cdots n$. Then it is easy to see that $\sum_{i=1}^{n} x_i = \dfrac{n^2 + n}{2}$ and that points P lie on an n-flat or hyperplane. Also $\sum_{i=1}^{n} x_i^2 = \dfrac{n(n+1)(2n+1)}{6}$ so points P lie on a hypersphere with center at the origin. Let us consider the joins PR and $P\bar{R}$. It is readily established that they are orthogonal. Then $(PR)^2 + (P\bar{R})^2 = (R\bar{R})^2 = \dfrac{n^3 - n}{3}$ or, since $S = \sum_{i=1}^{n} (x_i - i)^2$ and $\bar{S} = \sum_{i=1}^{n} (x_i - n + i - 1)^2$, $S + \bar{S} = \dfrac{n^3 - n}{3}$ a result previously established otherwise.

to devise a method which can be used successfully to extend the computation to larger values of n if desired. The details of the method follow.

Let D_n represent any series of n differences, $d_1, d_2, \cdots d_n$, and let O_n be an operator such that O_n operating on D_n (written $O_n(D_n)$) means that $D_n = (d_1, d_2, \cdots d_n)$ is changed to $(d_2 + 1, d_3 + 1 \cdots d_n + 1, d_1 - (n - 1))$. Let $\underline{m}$, written following $d_1, d_2, \cdots d_n$, indicate that $\Sigma d^2 = m$. For $n = 3$

$$\begin{aligned} D_{3,1} &= \phantom{D_{3,2} =} (0, \ \ 0, \ \ 0):\underline{0} \\ O_3(D_{3,1}) &= D_{3,2} = (1, \ \ 1, -2):\underline{6} \\ O_3(D_{3,2}) &= D_{3,3} = (2, -1, -1):\underline{6} \end{aligned}$$

But we have shown that $S + \bar{S} = \dfrac{k^3 - k}{3}$ for $n = k$. Therefore, for $n = 3$, we have $S + \bar{S} = 8$, so sums of 0 and 6 indicate corresponding sums of 8 and 2 when the order of the elements is reversed. Thus we have, for $n = 3$.

Sums	0	2	4	6	8
Frequencies	1	2	0	2	1

For $n = 4$ we have

$$\begin{aligned} D_{4,1,1} &= (0, \ \ 0, \ \ 0, \ \ 0) \\ D_{4,2,1} &= (1, \ \ 1, -2, \ \ 0) \\ D_{4,3,1} &= (2, -1, -1, \ \ 0) \end{aligned}$$

where these are obtained from $D_{3,1}$, $D_{3,2}$ and $D_{3,3}$ respectively by inserting a zero as a fourth difference. We operate on each of these four times with O_4. For example,

$$\begin{aligned} D_{4,2,1} &= \phantom{D_{4,2,2} =} (1, \ \ 1, -2, \ \ 0): \underline{6} \\ O_4(D_{4,2,1}) &= D_{4,2,2} = (2, -1, \ \ 1, -2):\underline{10} \\ O_4(D_{4,2,2}) &= D_{4,2,3} = (0, \ \ 2, -1, -1): \underline{6} \\ O_4(D_{4,2,3}) &= D_{4,2,4} = (3, \ \ 0, \ \ 0, -3):\underline{18} \\ O_4(D_{4,2,4}) &= D_{4,2,1} = (1, \ \ 1, -2, \ \ 0) \end{aligned}$$

As a check on computation, we notice, first, that the set is closed by the reappearance of $D_{4,2,1}$; and, second, that 10, 6, 18 and 6 are congruent, modulus 4. In like fashion, one of the sets for $n = 5$, is the following;

$$\begin{aligned} D_{5,2,4,1} &= (3, \ \ 0, \ \ 0, -3, \ \ 0):\underline{18} \\ D_{5,2,4,2} &= (1, \ \ 1, -2, \ \ 1, -1): \underline{8} \\ D_{5,2,4,3} &= (2, -1, \ \ 2, \ \ 0, -3):\underline{18} \end{aligned}$$

$$D_{5,2,4,4} = (0,\ \ 3,\ \ 1,\ -2,\ -2):\underline{18}$$
$$D_{5,2,4,5} = (4,\ \ 2,\ -1,\ -1,\ -4):\underline{38}$$
$$D_{5,2,4,1} = (3,\ \ 0,\ \ 0,\ -3,\ \ 0)$$

Of course the sums for $n = 5$ can be obtained from those for $n = 4$ by making use of (4). For $D_{4,2,4} = (3, 0, 0, -3):\underline{18}$ we have $S_1 = 18, k = 4, d_1 = 3, d_2 = 0, d_3 = 0, d_4 = -3$. Then

$$\begin{aligned}
S_1 &= 18 \\
S_2 &= S_1 + (4 - 2\cdot 3)(5) = 8 \\
S_3 &= S_2 + (4 - 2\cdot 0 - 2)(5) = 18 \\
S_4 &= S_3 + (4 - 2\cdot 0 - 4)(5) = 18 \\
S_5 &= S_4 + (4 - 2\cdot - 3 - 6)(5) = 38 \\
S_1 &= S_5 - 4\cdot 5 = 18.
\end{aligned}$$

However, results obtained by this latter method do not help with the case of $n = 6$. If we desire to obtain results for $n = 6$ we will need to exhibit the complete sets of differences for $n = 5$ as we did by the former method.

An alternative method for obtaining frequencies of sums of squares is of some interest. It will be illustrated for $n = 4$. Let us consider the square array

$$\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{pmatrix}$$

If we form all possible products $a_ib_jc_kd_l(i, j, k, l = 1, 2, 3, 4; i \neq j \neq k \neq l)$, the subscripts give the 4! permutations of 1, 2, 3, 4. Now let us form a new array

$$\begin{pmatrix} a_0 & b_1 & c_2 & d_3 \\ a_{-1} & b_0 & c_1 & d_2 \\ a_{-2} & b_{-1} & c_0 & d_1 \\ a_{-3} & b_{-2} & c_{-1} & d_0 \end{pmatrix}$$

where subscripts in each column represent the vertical distance of the term above the principal diagonal. Since the original terms had subscripts giving all possible arrangements of 1, 2, 3, 4, terms formed in a similar fashion from the new array will give all possible arrangements of the differences. Now form a third array

$$\begin{pmatrix} x^0 & x^1 & x^4 & x^9 \\ x^1 & x^0 & x^1 & x^4 \\ x^4 & x^1 & x^0 & x^1 \\ x^9 & x^4 & x^1 & x^0 \end{pmatrix}$$

where the exponent of x is the square of the corresponding subscript in the

TABLE I

Frequencies of sums of squares of rank differences

Σd^2 N	2	3	4	5	6	7
0	1	1	1	1	1	1
	*					
2	1	2	3	4	5	6
4		*0	1	3	6	10
6		2	4	6	9	14
8		1	2	7	16	29
10			*2	6	12	26
12			2	4	14	35
14			4	10	24	46
16			1	6	20	55
18			3	10	21	54
20			1	*6	23	74
22				10	28	70
24				6	24	84
26				10	34	90
28				4	20	78
30				6	32	90
32				7	42	129
34				6	29	106
					*	
36				3	29	123
38				4	42	134
40				1	32	147
42					20	98
44					34	168
46					24	130
48					28	175
50					23	144
52					21	168
54					20	144
56					24	*184

* The asterisk shows the location of the mean. The frequencies for $n = 6, 7$ extend beyond the limits of the table but may easily be obtained by symmetry.

second array. It is easy to see that, if terms are formed from the new array by the same method as before, our terms are powers of x where the exponents represent sums of squares of differences. If we now define the array to be equal to the sum of the terms formed from the array, then

$$\begin{pmatrix} x^0 & x^1 & x^4 & x^9 \\ x^1 & x^0 & x^1 & x^4 \\ x^4 & x^1 & x^0 & x^1 \\ x^9 & x^4 & x^1 & x^0 \end{pmatrix} = k_1x^0 + k_2x^2 + \cdots k_6x^{10} + \cdots k_2x^{18} + k_1x^{20},$$

and the k's give the desired frequencies for sums of squares corresponding to exponents of x. For example $\Sigma d^2 = 0$ occurs k_1 times, $\Sigma d^2 = 2$ occurs k_2 times, etc.

It can be readily verified that, for $n < 5$, the array can be expanded as a determinant and the values of the k's can be obtained by taking the absolute values of the coefficients in the expansion. Also, considering the arrays as determinants, their values for $n = 2, 3, 4$ are, respectively, $(1 - x^2)$, $(1 - x^2)^2$ $(1 - x^4)$, $(1 - x^2)^3$ $(1 - x^4)^2$ $(1 - x^6)$. If it were possible to obtain a general form of this type it might be possible to greatly reduce the labor which is involved in expanding the arrays. At present, however, this method of attack does not seem feasible on account of the lack of adequate sub-checks, the amount of work involved, and its inappropriateness for use by inexperienced clerical help.

Hotelling and Pabst[5] have given exact results in terms of n for the cases $\Sigma d^2 = 0, 2, 4, 6$. It is certainly possible to follow their method to obtain general results for Σd^2 larger than 6, but, as they suggest, the work becomes very laborious. For $\Sigma d^2 = 8$ we need the sets of possible integral values for x_1, x_2, $\cdots$ x_n, under the following conditions: (a) $\sum_{i=1}^{n} x_i = 0$, (b) $\sum_{i=1}^{n} x_i^2 = 8$, (c) $1 + x_1$, $2 + x_2$, $3 + x_3$, $4 + x_4$, $\cdots$ $n + x_n$ are the numbers 1, 2, 3, $\cdots$ n, (but not necessarily in that order).

Possible solutions are:

(a) $x_{i-2} = 2, x_{i-1} = 0, x_i = -2$ $(i = 3, 4, \cdots n)$ and the other x's zero,

(b) $x_{b-2} = 2, x_{b-1} = -1, x_b = -1, x_{a-1} = 1, x_a = -1$ $(a = 5, 6, \cdots n;$ $b = 3, 4, \cdots a - 2)$,

(c) $x_{b-1} = 1, x_b = -1, x_{a-2} = 2, x_{a-1} = -1, x_a = -1$ $(a = 5, 6, \cdots n; b = 2, 3, \cdots a - 3)$,

(d) $x_{b-2} = -2, x_{b-1} = 1, x_b = 1, x_{a-1} = 1, x_a = -1$ $(a = 5, 6, \cdots n; b = 3, 4, \cdots a - 2)$,

(e) $x_{b-1} = 1, x_b = -1, x_{a-2} = -2, x_{a-1} = 1, x_a = 1$ $(a = 5, 6, \cdots n; b = 2, 3, \cdots a - 3)$,

(f) $x_{d-1} = x_{c-1} = x_{b-1} = x_{a-1} = 1; x_d = x_c = x_b = x_a = -1$ $(a = 8, 9, \cdots n;$ $b = 6, 7, \cdots a - 2; c = 4, 5, \cdots b - 2; d = 2, 3, \cdots c - 2)$

Frequencies for each of these types must be considered separately. The

[5] Loc. cit. p. 35.

method of evaluation will be illustrated for type (f), since this type yields the polynomial of highest degree. It is apparent that the required frequency is obtained by computing $\sum_{8}^{n}\left(\sum_{6}^{a-2}\left(\sum_{4}^{b-2}\left(\sum_{2}^{c-2} 1\right)\right)\right)$. It can be verified that the result is

$$\frac{(n-4)^{(4)}}{4!} = \frac{(n-4)(n-5)(n \perp 6)(n-7)}{24}$$

The total of (a), (b), (c), (d), (e), and (f) is

$$(n-2) + 2(n-3)^{(2)} + \frac{(n-4)^{(4)}}{4!}$$

For $\Sigma d^2 = 10$, the result seems to be

$$2(n-3) + (n-3)^{(2)} + (n-4)^{(3)} + \frac{(n-5)^{(5)}}{5!}.$$

For sums greater than 8 the method becomes quite uninviting, not only because of the intricacy of the necessary analysis, but also because of the opportunities for mechanical errors and the absence of satisfactory checks. Besides, if the exact distribution for a particular value of n is desired, we need expressions for $\Sigma d^2 = 0, 2, 4, \cdots \frac{n^3 - n}{6} - 2$. For n as small as 8, this means the requirement of 42 formulas. It is fairly evident that these formulas will comprise polynomials ranging in degree from 0 to 41.

III. APPROXIMATIONS

Since the exact distributions of sums of squares are not easily obtained, we next consider the problem of finding approximations for them. Hotelling and Pabst[6] have given a method of deriving the even moments of the distribution of r', (the odd moments being zero), and have recorded the values of the second and fourth moments. They have also remarked that the kurtosis, $\beta_2 = \mu_4/\mu_2^2$, approaches 3 and that the distribution of r' approaches normality as n approaches infinity. These are valuable and interesting results. Because of them the normal curve suggests itself as an approximating function. Its use has been considered a little later in this investigation.

But a distribution with a finite range causes trouble at the tails when a normal fit is attempted, and, for this problem, we are particularly interested in the tails. It seems more feasible to attempt an approximation with the Pearson type II curve, $y = y_0\left(1 - \frac{x^2}{a^2}\right)^m$. This has the advantage of a finite range and three

[6] Loc. cit. p. 32 et seq.

constants to be determined. The values of these constants, as given by Elderton[7] are

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)}, \qquad a^2 = \frac{2\mu_2\beta_2,}{3 - \beta_2}, \qquad y_0 = \frac{N \times \Gamma\,(2m + 2)}{a \times 2^{2m+1} \times [\Gamma\,(m + 1)]^2} \tag{8}$$

(where N is the total frequency).

If we use this distribution to approximate the distribution of sums of squares, it proves convenient to define x as equal to one-half the deviation of Σd^2 from its mean, i.e.,

$$x = \frac{\Sigma d^2}{2} - \frac{n^3 - n}{12}$$

Then the relative frequency of $\Sigma d^2 = k$ is approximated by

$$\int_{x_s - \frac{1}{2}}^{x_s + \frac{1}{2}} f(x)\,dx \doteq f(x_s) \qquad \text{where } x_s = \frac{k}{2} - \frac{n^3 - n}{12}$$

(Of course, closer approximations may be obtained, if desired). The approximation used is clear if we remember that only even values of k are possible and that the range is now $\dfrac{n^3 - n}{6}$.

The moments for x are now obtained from the moments for r' by multiplying by the proper powers of $\dfrac{n^3 - n}{12}$. We have

$$\mu_2(x) = (n - 1)\left[\frac{n(n + 1)}{6}\right]^2$$

The value of β_2 is unchanged. For r' or x it is

$$\beta_2 = \frac{3(25n^4 - 13n^3 - 73n^2 + 37n + 72)}{25n(n + 1)^2(n - 1)}$$

For $n = 5$, $\mu_2 = 25$, $\beta_2 = 2.0720$, $N = 5!$. Using these values and equations (8), we obtain $a \doteq 10.566$, $m = .73276$, $y_0 = 7.8545$. The approximating function is $y = 7.8545\left(1 - \dfrac{x^2}{111.64}\right)^{.73276}$ In table II the computed values of y and the true frequencies are listed for comparison.

When testing the significance of a particular value of Σd^2 our principal interest is in the probability that $\Sigma d^2 \leq k$, rather than in the probability that $\Sigma d^2 = k$. The probability that $\Sigma d^2 \leq k$ requires cumulation of frequencies, followed by division by the total frequency. If results, given in table II, are compared it is noticed that the maximum error in using the type II function is .0194 and the average error is .0072. Comparisons for other values of n are given in table III.

[7] Elderton, W. P., *Frequency Curves and Correlation*, Layton, London, 2nd ed., 1927, p. 84.

TABLE II

Comparison of exact and approximate frequencies for n = 5

(Approximations obtained by computing ordinates of

$$y = 7.845\left(1 - \frac{x^2}{111.64}\right)^{.73276}\Bigg)$$

Σd^2	Frequencies		Cumulative (expressed as percent of 120)		Difference of cumulatives
	Exact	Approximate	Exact	Approximate	
0	1	1.50	.0083	.0125	−.0042
2	4	3.04	.0417	.0378	+.0039
4	3	4.21	.0667	.0729	−.0062
6	6	5.14	.1167	.1157	+.0010
8	7	5.91	.1750	.1650	+.0100
10	6	6.52	.2250	.2193	+.0057
12	4	7.01	.2583	.2777	−.0194
14	10	7.39	.3417	.3393	+.0024
16	6	7.65	.3917	.4031	−.0114
18	10	7.80	.4750	.4681	+.0069
20	6	7.85	.5250	.5335	−.0085
					average of absolute values = .0072

TABLE III

Approximating functions, with errors involved

n	Approximating functions		Average and maximum absolute values of differences of cumulatives		
	Type II	Normal	Exact—type II	Exact—normal	Type II—normal
5	$7.8545\left(1 - \frac{x^2}{111.64}\right)^{.73276}$	$\frac{5!}{\sqrt{50\pi}} e^{-\frac{x^2}{50}}$	.0072 .0194	.0200 .0415	.0210 .0357
6	$31.652\left(1 - \frac{x^2}{351.75}\right)^{1.3715}$	$\frac{6!}{\sqrt{122.5\pi}} e^{-\frac{x^2}{122.5}}$	.0030 .0126	.0131 .0273	.0136 .0270
7	$156.33\left(1 - \frac{x^2}{918.84}\right)^{2.0160}$	$\frac{7!}{\sqrt{261.33\pi}} e^{-\frac{x^2}{261.33}}$	.0017 .0067	.0106 .0221	.0108 .0209
8	$918.72\left(1 - \frac{x^2}{2098.4}\right)^{2.6635}$	$\frac{8!}{\sqrt{504\pi}} e^{-\frac{x^2}{504}}$			.0086 .0175
9	$6276.3\left(1 - \frac{x^2}{4332.6}\right)^{3.3140}$	$\frac{9!}{\sqrt{900\pi}} e^{-\frac{x^2}{900}}$			
10	$64515\left(1 - \frac{x^2}{8266.6}\right)^{3.9655}$	$\frac{10!}{\sqrt{1512.5\pi}} e^{-\frac{x^2}{1512.5}}$			

It would be very convenient if the cumulative frequencies could be approximated by the use of normal curves. In table III are listed the proper normal curves, along with comparisons with the exact values and with the values obtained from the type II curves. For the values of n investigated the normal curve is not as satisfactory as the type II. This, of course, is to be expected because of the lack of agreement between the fourth moment of the normal curve and of the exact distribution. However, in view of the fact that, for values of n investigated, the maximum and average errors decrease as n increases, it seems satisfactory to sacrifice accuracy to expedience and use the normal curve as an approximating function for cases of n greater than 10. This has been done in constructing table V. In further justification it might be noted that β_2, which approaches 3 as n approaches infinity is an increasing function of n for n greater than 3.

IV. TABLES TO TEST THE SIGNIFICANCE OF THE RANK CORRELATION COEFFICIENT, WITH EXAMPLES OF THEIR USE

Table IV gives the probability that, for any given value of n and a computed value of Σd^2 less than or equal to the mean, the value will not be exceeded by chance. For a value of Σd^2 greater than or equal to the mean, it gives the probability that the value will be equalled or exceeded. The values for $n = 2, 3, 4, 5, 6, 7$ are computed from exact frequencies; those for $n = 8, 9, 10$ are computed from type II curves.

Table V is constructed by the use of normal curves. It gives the limits of Σd^2 for a few of the more useful probabilities.

It seems desirable to explain why values of Σd^2 were tabled rather than values of r'. It was done for two reasons: first, to avoid the difficulties arising from discrete variates; and, second, because the tables seem more useful in the form given since the labor of completing the calculation of r' can be avoided if the computed value of Σd^2 tests as not significant.

Example 1. Seven individuals are ranked by two criteria, as indicated below. Are the results significantly alike?

A	1	2	3	4	5	6	7		
B	2	1	6	3	4	7	5		
d	1	-1	3	-1	-1	1	-2	/	0
d^2	1	1	9	1	1	1	4	/	18

Solution: Rows 3 and 4 give the differences and squared differences, respectively. If we enter table IV with $n = 7$ and $\Sigma d^2 = 18$, we find $P = .0548$, so we would expect that a value as small as 18 would occur by chance more than 5% of the time. This does not usually indicate significance so it is useless to compute the value of r'. It is interesting to notice that r' actually does prove to be equal to .68 and that, if we had used the formula, $\sigma_{r'} = 1.0471 \left(\frac{1 - r'^2}{\sqrt{n}}\right)$ we might have

TABLE IV

The probability that $\Sigma d^2 \geq S$ for $S \geq \Sigma_M$, or that $\Sigma d^2 \leq S$ for $S \leq \Sigma_M$ (where Σ_M represents mean value of sum of squares)

	$N = 2$	3	4	5	6	7	8	9	10
Σ_M	1	4	10	20	35	56	84	120	165
S									
0	.5000	.1667	.0417	.0083	.0014	.0002	.0003	.0001	.0000
2	.5000	.5000	.1667	.0417	.0083	.0014	.0006	.0002	.0001
4		.5000	.2083	.0667	.0167	.0034	.0011	.0003	.0001
6		.5000	.3750	.1167	.0292	.0062	.0018	.0005	.0001
8		.1667	.4583	.1750	.0514	.0119	.0028	.0007	.0002
10			.5417	.2250	.0681	.0171	.0042	.0010	.0003
12			.4583	.2583	.0875	.0240	.0059	.0015	.0004
14			.3750	.3417	.1208	.0331	.0081	.0020	.0005
16			.2083	.3917	.1486	.0440	.0108	.0027	.0007
18			.1667	.4750	.1778	.0548	.0141	.0035	.0009
20			.0417	.5250	.2097	.0694	.0179	.0045	.0011
22				.4750	.2486	.0833	.0224	.0057	.0014
24				.3917	.2819	.1000	.0275	.0071	.0018
26				.3417	.3292	.1179	.0331	.0087	.0022
28				.2583	.3569	.1333	.0396	.0106	.0027
30				.2250	.4014	.1512	.0469	.0127	.0032
32				.1750	.4597	.1768	.0550	.0152	.0039
34				.1167	.5000	.1978	.0639	.0179	.0046
36				.0667	.5000	.2222	.0736	.0210	.0054
38				.0417	.4597	.2488	.0841	.0244	.0064
40				.0083	.4014	.2780	.0956	.0281	.0075
42					.3569	.2974	.1078	.0323	.0086
44					.3292	.3308	.1207	.0368	.0100
46					.2819	.3565	.1345	.0417	.0114
48					.2486	.3913	.1491	.0470	.0130
50					.2097	.4198	.1645	.0528	.0148
52					.1778	.4532	.1806	.0589	.0168
54					.1486	.4817	.1974	.0656	.0189
56					.1208	.5183	.2150	.0726	.0212
58					.0875	.4817	.2332	.0802	.0237

TABLE IV—*Continued*

N	6	7	8	9	10
Σ_M	35	56	84	120	165
S					
60	.0681	.4532	.2520	.0882	.0264
62	.0514	.4198	.2715	.0966	.0293
64	.0292	.3913	.2915	.1056	.0324
66	.0167	.3565	.3120	.1149	.0358
68	.0083	.3308	.3330	.1248	.0394
70	.0014	.2974	.3544	.1351	.0432
72		.2780	.3761	.1459	.0472
74		.2488	.3982	.1571	.0515
76		.2222	.4205	.1688	.0561
78		.1978	.4431	.1809	.0609
80		.1768	.4657	.1935	.0659
82		.1512	.4885	.2065	.0713
84		.1333	.5113	.2198	.0769
86		.1179	.4885	.2336	.0828
88		.1000	.4657	.2477	.0889
90		.0833	.4431	.2622	.0954
92		.0694	.4205	.2770	.1021
94		.0548	.3982	.2922	.1091
96		.0440	.3761	.3077	.1164
98		.0331	.3544	.3234	.1239
100		.0240	.3330	.3394	.1318
102		.0171	.3120	.3557	.1399
104		.0119	.2915	.3721	.1483
106		.0062	.2715	.3888	.1570
108		.0034	.2520	.4056	.1659
110		.0014	.2332	.4226	.1751
112		.0002	.2150	.4397	.1846
114			.1974	.4568	.1944
116			.1806	.4741	.2044
118			.1645	.4914	.2146
120			.1491	.5086	.2251
122			.1345	.4914	.2358
124			.1207	.4741	.2468

TABLE IV—*Concluded*

N	8	9	10	N	8	9	10
Σ_M	84	120	165	Σ_M	84	120	165
S				S			
126	.1078	.4568	.2580	168	.0003	.1459	.4865
128	.0956	.4397	.2694	170		.1351	.4731
130	.0841	.4226	.2810	172		.1248	.4596
132	.0736	.4056	.2928	174		.1149	.4462
134	.0639	.3888	.3048	176		.1056	.4328
136	.0550	.3721	.3169	178		.0966	.4196
138	.0469	.3557	.3293	180		.0882	.4063
140	.0396	.3394	.3418	182		.0802	.3931
142	.0331	.3234	.3545	184		.0726	.3802
144	.0275	.3077	.3673	186		.0656	.3673
146	.0224	.2922	.3802	188		.0589	.3545
148	.0179	.2770	.3932	190		.0528	.3418
150	.0141	.2622	.4063	200		.0470	.3293
152	.0108	.2477	.4196	202		.0417	.3169
154	.0081	.2336	.4328	204		.0368	.3048
156	.0059	.2198	.4462	206		.0323	.2928
158	.0042	.2065	.4596	208		.0281	.2810
160	.0028	.1935	.4731	210		.0244	.2694
162	.0018	.1809	.4865	212		.0210	.2580
164	.0011	.1688	.5000	214		.0179	.2468
166	.0006	.1571	.5000	216		.0152	.2358

(Tables for cases 9 and 10 can be completed by symmetry.)

judged the value of r' significant, since $\sigma_{r'} = .213$, and .213 is less than one-third of .68.

Example 2. Six golfers found, upon ranking their scores and also ranking their respective amounts of sleep for the previous night, that the two orders were the reverse of one another except that the two ranking 1, 2 in sleep ranked 5, 6 in score. Is the negative correlation too great to be reasonably attributed to chance?

Solution: We find $\Sigma d^2 = 68$ and, upon consulting table IV, $P = .0083$, so we conclude that more sleep might mean fewer strokes.

Example 3. Before an examination a teacher ranked his class of 13 members.

After the examination he found that the sum of the squares of the deviations of rank on examination from rank estimated was 144. Should he consider the agreement satisfactory?

TABLE V

Pairs of values between which Σd^2 has a probability, P, of being included

N	$P = .99$		.98		.96		.90		.80	
11	40.8	399.2	58.2	381.8	77.1	362.9	105.6	334.4	130.8	309.2
12	60.9	505.1	82.4	483.6	105.9	460.1	141.2	424.8	172.5	393.5
13	93.3	634.7	119.6	608.4	148.2	579.8	191.2	536.8	229.3	498.7
14	125.9	780.1	161.4	748.6	195.8	714.2	247.4	662.6	293.3	616.7
15	174.5	945.5	211.8	908.2	252.6	867.4	313.8	806.2	368.2	751.8
16	227.8	1132.2	271.6	1088.4	319.4	1040.6	391.2	968.8	455.0	905.0
17	290.5	1341.4	341.4	1290.6	397.0	1235.0	480.4	1151.6	554.6	1077.4
18	363.6	1574.4	422.3	1515.7	486.3	1451.7	582.4	1355.6	667.8	1270.2
19	447.9	1832.1	514.9	1765.1	588.2	1691.8	698.0	1582.0	795.6	1484.4
20	544.1	2115.9	620.2	2039.8	703.4	1956.6	828.1	1831.9	939.0	1721.0
21	653.0	2427.0	738.9	2341.1	832.8	2247.2	973.6	2106.4	1098.7	1981.3
22	775.5	2766.5	872.0	2670.0	977.3	2564.7	1135.3	2406.7	1275.7	2266.3
23	912.5	3135.5	1020.2	3027.8	1137.8	2910.2	1314.2	2733.8	1471.0	2577.0
24	1064.7	3535.3	1184.3	3415.7	1315.1	3284.9	1511.1	3088.9	1685.4	2914.6
25	1233.0	3967.0	1365.4	3834.6	1510.1	3689.9	1727.0	3473.0	1919.8	3280.2
26	1418.2	4431.8	1564.1	4285.9	1723.6	4126.4	1962.7	3887.3	2175.3	3674.7
27	1621.1	4930.9	1781.5	4770.5	1956.5	4595.5	2219.2	4332.8	2452.6	4099.4
28	1842.7	5465.3	2018.1	5289.9	2209.8	5098.2	2497.3	4810.7	2752.8	4555.2
29	2083.7	6036.3	2275.1	5744.9	2484.3	5635.7	2797.9	5322.1	3076.7	5043.3
30	2345.0	6645.0	2553.2	6436.8	2780.8	6209.2	3122.0	5868.0	3425.2	5564.8

Solution: Entering table V with $n = 13$ we see that $P = .96$ for a value between 148.2 and 579.8, and that $P = .98$ for a value between 119.6 and 608.4. Therefore the probability of not exceeding 144 by chance is between .02 and .01. It would seem that the teacher showed considerable knowledge of his class.

Carnegie Institute of Technology.