



The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance  
Author(s): Milton Friedman

Source: *Journal of the American Statistical Association*, Vol. 32, No. 200, (Dec., 1937), pp. 675-701

Published by: American Statistical Association

Stable URL: <http://www.jstor.org/stable/2279372>

Accessed: 08/07/2008 13:40

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=astata>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

# THE USE OF RANKS TO AVOID THE ASSUMPTION OF NORMALITY IMPLICIT IN THE ANALYSIS OF VARIANCE

BY MILTON FRIEDMAN  
*National Resources Committee*

**M**OST projects involving the collection and analysis of statistical data have for one of their major aims the isolation of factors which account for variation in the variable studied. The statistical tool ordinarily employed for this purpose is the analysis of variance. Frequently, however, the data are sufficiently extensive to indicate that the assumptions necessary for the valid application of this technique are not justified. This is especially apt to be the case with social and economic data where the normal distribution is likely to be the exception rather than the rule. This difficulty can be obviated, however, by arranging each set of values of the variate in order of size, numbering them 1, 2, and so forth, and using these ranks instead of the original quantitative values. In this way no assumption whatsoever need be made as to the distribution of the original variate.

The utilization of ranked data is thus frequently a desirable device to avoid normality assumptions; in addition, however, it may be inescapable either because the data available relate solely to order, or because we are dealing with a qualitative characteristic which can be ranked but not measured.

The possibility of using ranked data in problems involving simple correlation and thereby avoiding assumptions of normality has recently been emphasized in an article by Harold Hotelling and Margaret Richards Pabst.<sup>1</sup> It is the purpose of the present article to outline a procedure whereby the analysis of ranked data can be employed in place of the ordinary analysis of variance when there are two (or more) criteria of classification. This procedure has two major advantages. As already indicated, it is applicable to a wider class of cases than the ordinary analysis of variance. In addition, it is less arduous than the latter technique, requiring but a fraction as much time. The loss of information through utilizing the procedure outlined below when the analysis of variance could validly be applied may thus be more than compensated for by its greater economy. This consideration is likely to be especially important with those large scale collections of social and economic data which have become increasingly frequent in recent years and for which the funds available for analysis are limited.

<sup>1</sup> "Rank Correlation and Tests of Significance Involving No Assumption of Normality," *Annals of Mathematical Statistics*, VII (1936) 29-43.

## THE PROCEDURE

The procedure, which I shall call the *method of ranks*, involves first ranking the data in each *row* of a two-way table and then testing to see whether the different *columns* of the resultant table of ranks can be supposed to have all come from the same universe. This test is made by computing from the mean ranks for the several columns a statistic,  $\chi_r^2$ , which tends to be distributed according to the usual  $\chi^2$  distribution when the ranking is, in fact, random, i.e., when the factor tested has no influence.

The details of the procedure can best be explained by presenting an example. Table I gives the standard deviations of expenditures on different categories of expenditure for seven income levels.<sup>2</sup> It is de-

TABLE I  
STANDARD DEVIATIONS AT DIFFERENT INCOME LEVELS\* OF EXPENDITURES ON  
THE MAJOR CATEGORIES DURING 1935-36 OF 246 MINNEAPOLIS AND  
ST. PAUL FAMILIES OF WAGE-EARNERS AND LOWER  
SALARIED CLERICAL WORKERS†

Category of expenditure	Annual family income						
	\$750- 1,000	\$1,000- 1,250	\$1,250- 1,500	\$1,500- 1,750	\$1,750- 2,000	\$2,000- 2,250	\$2,250 2,500
Housing	\$103.3	\$68.42	\$89.53	\$77.94	\$100.0	\$108.2	\$184.9
Household operation	42.19	44.31	60.91	73.90	43.87	61.74	102.3
Food	71.27	81.88	100.71	86.52	100.3	90.75	100.6
Clothing	37.59	60.05	56.97	60.79	71.82	83.04	117.1
Furnishings and equip- ment	58.31	52.73	96.04	60.42	104.33	89.78	85.77
Transportation	46.27	82.18	129.8	181.0	172.33	164.8	246.8
Recreation	19.00	23.07	38.70	45.81	59.03	50.69	55.18
Personal care	8.31	8.43	9.16	14.28	10.63	15.84	12.50
Medical care	20.15	33.48	60.08	69.35	114.34	45.28	101.6
Education	3.16	4.12	12.73	18.95	8.89	41.52	66.33
Community welfare	4.12	18.87	8.54	12.92	25.30	19.85	16.76
Vocation	7.68	11.18	10.44	10.95	10.54	13.96	14.39
Gifts	5.29	10.91	11.22	25.26	42.25	48.80	69.38
Other	6.00	5.57	22.23	2.45	6.24	1.00	4.00

\* In computing the standard deviations the influence of family composition (in terms of number of members and their age) was eliminated by grouping the families at each income level into similar family types and computing the sums of squares within such income-family type groups. These sums of squares were summed for the family types at each income level and divided by the number of degrees of freedom. This gave the variance at each income level. It is the square roots of the variances which are entered in the table.

† The figures in this table are based on schedules collected by the Cost of Living Division of the U. S. Bureau of Labor Statistics. These schedules were loaned to the National Resources Committee for special analyses, of which this is one.

<sup>2</sup> The figures given in Table I were obtained from schedules on the receipts and disbursements of families of wage earners and lower salaried clerical workers during 1935-36 collected in Minneapolis and St. Paul by the Cost of Living Division of the U. S. Bureau of Labor Statistics. These schedules were loaned to the National Resources Committee for special analyses, several of which are used in this article.

sired to determine whether the standard deviations differ significantly for the different income levels.

The first step is to form Table II from Table I by ranking the standard deviations for each category, giving the lowest value a rank of 1,

TABLE II  
RANKING OF INCOME LEVELS BY SIZE OF STANDARD DEVIATION FOR EACH CATEGORY OF EXPENDITURE\*

Category of expenditure	Annual family income						
	\$750- 1,000	\$1,000- 1,250	\$1,250- 1,500	\$1,500- 1,750	\$1,750- 2,000	\$2,000- 2,250	\$2,250- 2,500
Housing	5	1	3	2	4	6	7
Household operation	1	3	4	6	2	5	7
Food	1	2	7	3	5	4	6
Clothing	1	3	2	4	5	6	7
Furnishings and equip- ment	2	1	6	3	7	5	4
Transportation	1	2	3	6	5	4	7
Recreation	1	2	3	4	7	5	6
Personal care	1	2	3	6	4	7	5
Medical care	1	2	4	5	7	3	6
Education	1	2	4	5	3	6	7
Community welfare	1	5	2	3	7	6	4
Vocation	1	5	2	4	3	6	7
Gifts	1	2	3	4	5	6	7
Other	5	4	7	2	6	1	3
a. Total	23	36	53	57	70	70	83
b. Mean rank	1.643	2.571	3.786	4.071	5.000	5.000	5.929
c. Deviation	-2.357	-1.429	-.214	.071	1.000	1.000	1.929

Sum of squared deviations = 13.3692  
 $\chi^2 = 40.108$

\* The figures in this table are derived from Table I.

the next lowest rank of 2, etc.<sup>3</sup> Thus, in each row of Table II, we have a set of numbers from 1 to 7, since there are seven income levels.

On the hypothesis that for any one category the value of the standard deviation is the same at all income levels, differences among the values in each row of Table I will arise solely from sampling fluctuations. The rank entered for a particular income level would then be a matter of chance; in repeated samples each of the numbers from 1 to 7 would appear with equal frequency.<sup>4</sup>

<sup>3</sup> It is, of course, immaterial whether the ranking is from the lowest to the highest or the reverse, i.e., from the highest to the lowest.

<sup>4</sup> This statement is strictly valid only if the different entries in the same row are assumed to come from the same universe—no matter, of course, what its nature. In the present example it requires some qualification since the standard deviations in each row are not all based on the same number of cases. In this case, while two entries in the same row of the original table (e.g., Table I) will have the same expected value, one will exceed the other more than half the time. The reason for this is that the

If, therefore, the standard deviation were independent of the income level, the set of ranks in each *column* would represent a random sample of 14 items (that being the number of categories of expenditure) from the discontinuous rectangular universe—1, 2, 3, 4, 5, 6, 7. The mean of this universe is 4, or, in general,  $\frac{1}{2}(p+1)$ , where  $p$  is the number of ranks. The variance is also 4, or in general  $(p^2-1)/12$ .<sup>5</sup>

The next step in the procedure is to obtain the mean rank for each column. These are given on line b of Table II. In the absence of a relation between the standard deviations and income level, these means are all estimates of the same thing, namely of the mean of the rectangular universe. Moreover, the sampling distribution of the means will be approximately normal so long as the number of rows is not too small.<sup>6</sup>

The sampling distribution of the mean ranks (where  $\bar{r}_j$  is the mean rank of the  $j$ -th column) will have a mean value ( $\rho$ ) of  $\frac{1}{2}(p+1)$  and a variance  $\sigma^2$  of  $(p^2-1)/(12n)$ , where  $n$  is the number of rows, i.e., the number of ranks averaged.<sup>7</sup>

Since the true mean and true standard deviation of the chance universe are known, the hypothesis that the means come from a single homogeneous normal universe can be tested by computing

$$\chi_r^2 = \frac{p-1}{p\sigma^2} \sum_{i=1}^p (\bar{r}_i - \rho)^2 = \frac{12n}{p(p+1)} \sum_{i=1}^p \left\{ \bar{r}_i - \frac{1}{2}(p+1) \right\}^2 .$$

sampling distribution of the ratio of two variances is not symmetrical unless both variances are based on the same number of degrees of freedom. The mean value of the ratio is approximately unity, but the median is not equal to one—it is less than one if the numerator is based on fewer degrees of freedom than the denominator, and conversely. In ranking *two* standard deviations, therefore, the one based on the smaller number of cases would receive a rank of 1 more than half the time. When more than two standard deviations are ranked this tendency is somewhat compensated for by the greater probability that those based on the fewest cases will receive relatively high ranks, and thus the average rank will be less affected. This difficulty does not, however, affect the validity of the illustrative analysis presented here, since the two highest income classes contain the smallest numbers of families but have the highest average ranks.

More generally, when the entries in different columns of the same row come from symmetrical universes with the same mean but different variances, the several ranks will have the same *expected* value, but the probability distribution for each cell will not be exactly rectangular. This condition of symmetry is a sufficient condition for the ranks to have the same expected value; it is, however, more stringent than is necessary. This difficulty clearly calls for further analysis.

<sup>5</sup> The sum of the numbers from 1 to  $p$  is  $\frac{1}{2}p(p+1)$ . The mean is therefore  $\frac{1}{2}(p+1)$ . The sum of the squares of the numbers from 1 to  $p$  is  $(2p+1)(p+1)p/6$ . The variance is, therefore,  $(2p+1)(p+1)/6 - \frac{1}{4}(p+1)^2 = (p^2-1)/12$ .

<sup>6</sup> That the sampling distribution of samples drawn from a rectangular universe approaches normality quite rapidly is, of course, well known. The distribution of means for samples of two is a triangle; for samples of three it is made up of three parabolic segments, the first and third concave upwards, and the middle one concave downward. An empirical distribution for samples of ten is given by Hilda Frost Dunlap, "An Empirical Determination of the Distribution of Means, Standard Deviations and Correlation Coefficients Drawn from Rectangular Populations," *Annals of Mathematical Statistics*, II (1931), 66-81. The universe sampled was a discontinuous rectangular universe, including the integers from 1 to 6. The empirical distribution shows extremely close conformity to the normal curve.

<sup>7</sup> This follows from the fact that the variance of a mean of  $n$  observations of equal weight is  $1/n$  times the variance of an individual observation.

So long as the number of rows and columns is not too small,  $\chi_r^2$  computed in this way will be distributed according to the usual  $\chi^2$  distribution with  $p-1$  degrees of freedom.<sup>8</sup> If, now,  $\chi_r^2$  is significantly greater than might reasonably have been expected from chance, the implication is that the mean ranks differ significantly, i.e., that the size of the standard deviation depends on the income level.

The computation of  $\chi_r^2$  is extremely simple. The mean of the seven mean ranks is, of necessity, equal to the true mean of 4. The difference between the mean rank for each column and 4 is given on line c of Table II. The sum of the squares of these differences is 13.3692 and  $\chi_r^2 = 40.1076$ .

This illustrative computation has been made using a formula that makes clear the nature of  $\chi_r^2$ . In actual practice the following alternative formula which involves only integers and makes unnecessary the computation of the actual mean ranks will be found more convenient:

$$\chi_r^2 = \frac{12}{np(p+1)} \sum_{j=1}^p \left( \sum_{i=1}^n r_{ij} \right)^2 - 3n(p+1),$$

where  $r_{ij}$  is the rank entered in the  $i$ -th row and  $j$ -th column.

The number of degrees of freedom on which this estimate is based is  $p-1=6$ . For six degrees of freedom the value of  $\chi^2$  which would be exceeded by chance once in 20 times is 12.592, and once in a hundred times, 16.812.<sup>9</sup> The probability of a value greater than 40 is .000001.<sup>10</sup> There can thus be little question that the observed mean ranks differ significantly, i.e., that the standard deviation is related to the income level. From the mean ranks it is seen that with but one minor exception the standard deviations consistently increase with income.

Since the value of  $\chi_r^2$  is invariant under transpositions of the columns of ranks under their captions this information—that the ranks increase with income—has not been utilized. Whenever the columns themselves can be ranked, the additional information supplied by the relationship between the order of the mean ranks and the order of the columns can be used by computing a rank difference correlation between the two corresponding sets of ranks, determining the probability that the correlation coefficient obtained would have been equalled or exceeded by chance, converting this probability into the value of  $\chi^2$

<sup>8</sup> For a justification of the formula for  $\chi_r^2$  and of the statement that  $\chi_r^2$  tends to be distributed like  $\chi^2$ , as well as for some indication of the number of columns and rows necessary, see pp. 687-694 and the mathematical appendix.

<sup>9</sup> Fisher, R. A., *Statistical Methods for Research Workers*, Table III.

<sup>10</sup> Pearson, Karl, *Tables for Statisticians and Biometricians*, 3rd Edition, London, 1930, Part, I Table XII.

which corresponds to it for two degrees of freedom, and pooling the resultant value of  $\chi^2$  with  $\chi_r^2$ .<sup>10a</sup> In the present illustrative example the evidence is so clear that this additional information will obviously not affect the conclusion. It will, however, serve to exemplify the procedure. The rank difference correlation between the mean rank and the income level is .991. (In deriving this coefficient the tied ranks were treated in the manner suggested below, i.e., they were assigned the average value of the ranks for which they were tied.) The probability of securing a value as great as or greater than this is between .00277 and .00040. The value of  $\chi^2$  corresponding to the larger of these figures for two degrees of freedom is  $-2 \log_e .0277 = 11.77$ . Adding this to  $\chi_r^2$  gives 51.88 as the value to be entered in the  $\chi^2$  table for 8 degrees of freedom. The probability associated with this value is smaller than that for  $\chi_r^2$  and, indeed, is so small that it cannot be determined from the published tables.

In order to test whether the standard deviations are related to the type of expenditure it is only necessary to repeat the above analysis; this time, however, treating the columns in the way in which the rows were previously treated, and *vice versa*. Thus the standard deviations would be ranked for each income level, and the mean ranks obtained for each type of expenditure.

It might appear offhand as if the procedure used to study the relation between standard deviations and income level does not make use of all of the information provided by Table II, that it neglects the distribution of the ranks within the columns, and that this supplies additional information about the consistency of the ranking. This, however, is not the case. Since Table II must contain  $n$  1's,  $n$  2's, . . . ,  $n$   $p$ 's, the total sum of the squared deviations from the grand mean is the same no matter what the arrangement of the ranks within the table—it is, in fact, equal to  $np(p^2 - 1)/12$ . The sum of squares within columns plus the sum of squares between columns must add up to this total. Knowledge of one of these sums of squares thus implies knowledge of the other. In the above example we have used the sum of squares between columns; no additional information is thus supplied by the sum of squares within columns.

It should be noted that in testing the significance of the differences among the columns no assumption whatsoever needs to be made as to the similarity of the distribution of the original variate for the different rows. The test takes the form of comparing the mean ranks for the several columns; essentially, however, the null hypothesis tested is

<sup>10a</sup> See Hotelling and Pabst, *op. cit.*, pp. 35 and 40, and Fisher, *op. cit.* art., 21.1.

that the original entries in each row are from the same universe; whether or not this universe is the same for the different rows is entirely irrelevant to the validity of the test.

The method of ranks does not provide for testing "interaction." It is of the very nature of the method that it cannot do so. Without exact quantitative measurement, "interaction," in the sense used in the ordinary analysis of variance, is meaningless.

It should further be noted that the method of ranks may not provide a test of the influence of a factor if there is reason to suspect that this influence is in a different *direction* for the different rows; if, for example, the standard deviation increases with income for certain types of expenditure and decreases with income for others. For in such a case the mean ranks of the  $p$  columns may all have the same expected value, although the  $p$  ranks for each of the rows do not. Thus, if  $\chi_r^2$  is significant, the conclusion is that the ranking is not random. But  $\chi_r^2$  may not be significant, not because the ranking is random, or because the differences in the mean ranks are too small for the observed sample to display significance, but because the influence of the factor tested is different in direction for the different rows. In this connection, however the general point should be emphasized that non-significant results do not establish the validity of the null hypothesis in the same way that significant results tend to contradict it.

In some cases two (or more) of the values of the variate in a row will be identical, i.e., there will be "tied" ranks. Two procedures can be followed: first, the ranks tied for can be assigned to the two (or more) values at random; or second, each value can be given the average value of the ranks tied for (e.g., if two values are tied for the ranks 2 and 3 each can be given the rank of 2.5). In general, the second of these procedures seems to be preferable, since it uses slightly more of the information provided by the data.<sup>11</sup> The substitution of the average rank for the tied values does not affect the validity of the  $\chi_r^2$  test.<sup>12</sup>

#### THE EFFICIENCY OF THE METHOD OF RANKS RELATIVE TO THE ANALYSIS OF VARIANCE

It is evident that the method of ranks does not utilize all of the information furnished by the data, since it relies solely on order and

<sup>11</sup> This alternative method of handling tied ranks and its advantages were brought to my attention by Mr. W. Allen Wallis, who has developed a simple adjustment to the usual formula for the rank-difference correlation to allow for the treatment of tied ranks in this fashion.

<sup>12</sup> Its only effect is to change very slightly the "true" value of the variance. In the extreme case when tied ranks are as probable as untied ranks, the variance of an individual observation is changed from  $(p^2-1)/12$  to  $p(p-1)/12$ , i.e., it is reduced by  $(p-1)/12$  or in the ratio of 1 to  $p+1$ . The reduction is thus relatively small when  $p$  is moderately large.



makes no use of the quantitative magnitude of the variate. It is this very fact that makes it independent of the assumption of normality. At the same time, it is desirable to obtain some notion about the amount of information lost, that is, about the efficiency of the method of ranks in situations where the analysis of variance provides the proper test.<sup>13</sup>

For the special case of  $p=2$  (i.e., of two ranks) the method of ranks is equivalent to the binomial series test of significance of a mean difference, that is, it is equivalent to testing whether the proportion of positive differences between the pairs of values in each row of the  $2 \times n$  table (the proportion of 2's in the first column of the table of ranks) differs significantly from  $\frac{1}{2}$ .<sup>14</sup> Now, W. G. Cochran recently showed<sup>15</sup> that the binomial series test of a mean difference has an efficiency of 63.7 per cent. It follows that the method of ranks, for the special case of  $p=2$ , likewise has an efficiency of 63.7 per cent.

<sup>13</sup> By the "efficiency" of a statistic  $m$  used to estimate a parameter  $\mu$  is meant the ratio of the variance of the maximum likelihood estimate of  $\mu$  to the variance of  $m$ . The difference between this ratio and unity multiplied by 100 gives the percentage of "information" lost. (R. A. Fisher, *op. cit.*, Chapter IX.)

In the present instance, since the analysis of ranks and the analysis of variance provide estimates of different parameters—in the one case, of  $\chi^2$ , and in the other, of the analysis of variance ratio—it is first necessary to secure a relationship between the two parameters which can be used to estimate one from the other. In this way both methods can be used to estimate the same parameter.

<sup>14</sup> The analogous method for  $p$  greater than 2, while it provides a method for analyzing a table of ranks and seems superficially closely related to the method of ranks, is essentially very different.

This alternative procedure involves the formation from the basic table of ranks of a  $p \times p$  contingency table giving the number of ranks of each size in each column. Thus, one of the classifications is by column number, the other by the value of the rank. Such tables can then be analyzed by computing  $\chi^2$  in the usual manner and testing its significance. Unless the number of rows is large relative to the number of columns, the usual  $\chi^2$  tables will, of course, not be applicable. Exact distributions can, however, be obtained in the manner indicated by F. Yates ("Contingency Tables Involving Small Numbers and the  $\chi^2$  Test," *Journal of the Royal Statistical Society*, Supplement, Volume I (1934), pp. 217-35).

This procedure does not, however, test the same hypothesis as the method of ranks. The reason is that with the contingency table method the numerical values of the ranks in no way affect the result, whereas in the method of ranks they do. Thus, consider the following  $3 \times 3$  tables of ranks:

A.	1	2	3		B.	1	2	3
	1	2	3			1	2	3
	3	2	1			1	3	2

It is clear that B indicates greater departure from the hypothesis that the ranking is random than does A. Both tables contain one column in which all three ranks are identical and two columns in which two out of three ranks are the same. But in B these latter two columns contain ranks which vary less than for the corresponding columns of A. Stated differently, in B every rank in the last two columns is greater than any rank in the first; no comparable statement is valid for A.

The contingency analysis would indicate, however, that A and B diverge equally from expectation, since both will give contingency tables which, except for permutations of rows and columns, are identical. The method of ranks, on the other hand, will indicate that B diverges more from expectation than A;  $\chi^2$  is  $4\frac{2}{3}$  for B, but only  $\frac{2}{3}$  for A.

For the purpose of determining whether one variable has a significant influence on another, it seems clear that the method of ranks is definitely preferable to the contingency analysis just outlined.

The reason why the two methods are equivalent for  $p=2$  is evident; when there are only two ranks, there is no possibility of different ranks diverging by varying amounts.

<sup>15</sup> "The Efficiencies of the Binomial Series Tests of Significance of a Mean and of a Correlation Coefficient," *Journal of the Royal Statistical Society*, C (1937), 69-73.

Moreover, this provides a measure of the *minimum* efficiency of the method of ranks. When  $p=2$ , a classification in terms solely of greater or smaller is substituted for the exact quantitative measurements; as  $p$  increases a more and more finely subdivided scale is substituted for the exact measurements. It seems reasonable, therefore, that the loss in information through using ranks decreases as  $p$  increases.

For the special case of  $n=2$ , it is shown below that  $\chi_r^2$  and the rank difference correlation are essentially equivalent. On the assumption that the true correlation is zero, Hotelling and Pabst have shown that the efficiency of the rank difference correlation approaches 91.19 per cent as  $p$  increases. In their words, "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."<sup>16</sup>

For the more general case, when  $p$  and  $n$  are greater than 2 I have not been able to determine the efficiency of the method of ranks. It seems clear, however, that the loss of information is less than the 36 per cent lost when  $p=2$  and probably greater than the 9 per cent lost when  $n=2$ .

In the absence of the theoretical analysis there are presented here the results of applying both the analysis of variance and the method of ranks to the same data. A comparison of these results will, of course, offer no conclusive evidence as to the relative efficiency of the two methods; but it should at least suggest whether the loss of information in using the method of ranks is so great as to vitiate completely its usefulness.

The data analyzed are the same as those utilized in the illustrative analysis summarized in Tables I and II above, i.e., they are data on the expenditures and savings during 1935-36 of 246 Minneapolis and St. Paul families of wage earners and lower salaried clerical workers. In the present instance, however, the analysis is directed toward determining whether income and family composition have a significant influence on the expenditures for the various categories and on savings. The analysis given above, it will be recalled, attempted to determine whether income had a significant influence on the *standard deviations* of expenditure.

The 246 families have been grouped into seven income classes,<sup>17</sup>

<sup>16</sup> *Op. cit.*, pp. 42-43.

<sup>17</sup> The total income of a family is defined as including not only money income, but also the imputed value of gifts in kind, of food produced at home and of the use of a home owned by the family.

each \$250 in range, and five family composition types.<sup>18</sup> This gives 35 groups in all.

For each of the major categories of expenditures, for savings, and for certain sub-groups of items, three variances have been computed: the variance (1) between income levels, (2) between family types, and (3) within groups.<sup>19</sup>

For 14 major categories of expenditure, for 13 sub-groups of several of these categories, and for savings, there was computed the ratios of the variance between income levels and the variance between family types to the variance within groups. These ratios, designated as  $F_i$  and  $F_f$  respectively, are given in Table III.

To each of the 28 items considered, the method of ranks was also applied to test the influence of income and family type.

In testing the influence of income, the seven mean expenditures for each family type were ranked. This gave five sets of seven ranks. In testing the influence of family type the procedure was reversed; the five mean expenditures at each income level were ranked, giving seven sets of five ranks.

The results of the method of ranks are likewise given in Table III. This table gives the values of  $\chi^2$  computed in testing for the influence of income, as well as those obtained in testing for the influence of family type.

For both the analysis of variance and the method of ranks the values which are significant at the .01 level are indicated with a double star; those which are significant only at the .05 level, with a single star.

The two methods yield measures of the influence of income and family type for 28 items. There are thus 56 independent analyses by each method. In Table III these measures are classified into three

<sup>18</sup> The family types are defined as follows:

Type 1 Husband, wife, and one child under 16

2 Husband, wife, and two children under 16

3 Husband, wife, one person 16 or over, and one or no other persons

4 Husband, wife, one child under 16, one person 16 or over, and one or two other persons

5 Husband, wife, and three or four children under 16.

<sup>19</sup> There was also computed the variance due to interaction. Since the method of ranks can give no measure of interaction, this variance is of no interest here. It is worth pointing out, however, that interaction was significant for only three out of 28 cases; for one of those the probability was between .05 and .01 and for two it was less than .01.

Since the numbers of items in the subclasses are neither equal nor proportionate, there is some difficulty in decomposing the variation between groups. The variances between income levels and between family types were computed by the method of weighted squares of means. This method does not give an estimate of interaction when there are more than two classes for both of the factors. Consequently, the variance due to interaction was computed by the method of unweighted means.

For an excellent statement of the difficulties raised by disproportionate subclass numbers and of the available methods of analysis, see G. C. Snedecor, and G. M. Cox, "Disproportionate Subclass Numbers in Tables of Multiple Classification," Research Bulletin 180, Agricultural Experiment Station, Iowa State College of Agriculture and Mechanic Arts (March 1935).

TABLE III  
RESULTS OF ANALYSIS OF VARIANCE AND METHOD OF RANKS

Measures of the Influence of Income and Family Type on Expenditures for the Major Categories of Expenditure and for Sub-Groups of Items, and on Savings, Based on Data on the Expenditures and Savings During 1935-36 of 246 Minneapolis and St. Paul Families<sup>1</sup>

Item	Analysis of variance Ratios of variances <sup>2</sup>		Method of ranks $\chi_r^2$	
	Income $F_i$	Family type $F_f$	Income	Family type
<i>Major categories of expenditure<sup>3</sup></i>				
Food	15.33**	5.75**	27.02**	19.09**
Household operation	9.95**	1.01	24.24**	4.94
Housing	9.50**	1.63	21.94**	6.17
Clothing	9.40**	1.38	25.54**	9.46
Recreation	4.25**	1.98	23.83**	11.89*
Personal care	4.10**	.80	21.11**	4.14
Transportation	3.78**	1.97	24.00**	10.06*
Gifts	3.36**	.96	21.17**	3.74
Community welfare	2.95**	.45	17.04**	.49
Education	2.93**	1.79	17.31**	8.11
Medical care	2.51*	.80	18.69**	6.51
Vocation	.69	1.01	4.71	1.51
Furnishings and equipment	.42	.37	6.96	3.69
Other	.25	.30	5.74	5.40
<i>Savings (or deficit)</i>	2.50*	1.25	14.74*	4.57
<i>Sub-groups of items</i>				
Food <sup>4</sup> :				
Dairy products	6.71**	9.41**	23.66**	21.83**
Fruit	4.87**	.38	12.69*	3.31
Food away from home	3.49**	3.94**	17.34**	10.09*
Meat	2.59*	2.02	9.34	3.77
Miscellaneous foods	2.01	1.21	15.00*	5.49
Fish	.98	2.43*	4.11	1.91
Vegetables	.73	2.11	6.69	8.80
Grain products	.71	4.76**	3.26	9.71*
Sweets	.20	1.05	3.96	9.94*
Poultry	.20	.99	.30	1.89
Personal care:				
Personal service	4.31**	.70	19.80**	4.71
Personal supplies	3.38**	.75	14.34*	1.49
Household operation:				
Fuel and light <sup>5</sup>	7.26**	1.56	23.25**	6.74

\* Indicates that observed figure is "significant," i.e., greater than the value which would be exceeded by chance once in twenty times. For the ratios of variances this value is 2.14 for income and 2.42 for family type. For  $\chi_r^2$  it is 12.592 for income and 9.488 for family type. The difference between the values for income and family type is a result of a difference in the number of degrees of freedom on which the respective estimates are based.

\*\* Indicates that observed figure is "highly significant," i.e., greater than the value which would be exceeded but once in a hundred times by chance. For the ratios of variances this value is 2.89 for income and 3.41 for family type. For  $\chi_r^2$  it is 15.033 for income and 13.277 for family type.

<sup>1</sup> The figures in this table are based on schedules collected by the Cost of Living Division of the U. S. Bureau of Labor Statistics. These schedules were loaned to the National Resources Committee for special analyses, one of which is presented here.

<sup>2</sup>, <sup>3</sup>, <sup>4</sup>, <sup>5</sup> See next page.

groups: those which would have been exceeded by chance (a) in more than five per cent of random samples, (b) in between five per cent and one per cent of random samples, and (c) in less than one per cent of random samples. An indication of the relative efficiency of the two methods is provided by Table IV, which gives a comparison of the two classifications.

From the entries in the diagonal of Table IV, it is seen that for 45 out of the 56 analyses the two methods lead to similar conclusions. In no case does one of the methods indicate a probability of less than .01 while the other indicates a probability greater than .05.

TABLE IV  
COMPARISON OF RESULTS OF ANALYSIS OF VARIANCE AND METHOD OF RANKS

Method of ranks Probability of $\chi_r^2$	Analysis of variance Number of $F$ 's with probability			Total
	Greater than .05	Between .05 and .01	Less than .01	
Greater than .05	28	2	0	30
Between .05 and .01	4	1	4	9
Less than .01	0	1	16	17
Total	32	4	20	56

In this example, it seems clear that the loss of information in using the method of ranks is not very great. Indeed, on the basis of Table IV alone, it would be difficult, if not impossible, to choose between the two methods.

A comparison of the ranking of the 28 items by the size of  $F$  and by the size of  $\chi_r^2$ , provides one further indication that the hypotheses tested by the two methods are essentially the same except for the inclusion of the normality assumption in that tested by the analysis of variance. The rank difference correlation between  $F_i$  and the corresponding  $\chi_r^2$  is .88; between  $F_f$  and the corresponding  $\chi_r^2$ , .66. Both correlations are very large in comparison with their standard error of .19.

<sup>2</sup>  $F_i$  is the ratio of the variance between income levels to the variance within classes.  $F_f$  is the ratio of the variance between family types to the variance within classes.

<sup>3</sup> Expenditures include not only money expenses but also the imputed value of gifts in kind. For food, the imputed value of home produced food, and for housing, the imputed value of the use of an owned home, are also included.

<sup>4</sup> The original data give the expenditures on the sub-groups of food only for a seven day period. The remaining ratios in the table are all based on data for annual expenditures.

<sup>5</sup> Fuel and light is, of course, but one of the sub-groups under household operation.

It should be noted that the illustrative comparison just presented is to some extent weighted against the analysis of variance. The distribution of expenditure data departs considerably from normality.<sup>20</sup> In addition, the analysis summarized in Tables I and II indicated that the standard deviation of expenditures is related to the income level; the assumption of uniform variance is, therefore, not justified. However, the body of data analyzed represents no more extreme a departure from the assumptions of normality and uniform variance than is frequently met with.

#### THE RELATION BETWEEN THE DISTRIBUTION OF $\chi_r^2$ AND $\chi^2$

The statement was made above without proof that  $\chi_r^2$  tends to be distributed as  $\chi^2$  with  $p-1$  degrees of freedom. This statement requires justification.

It is well known that the sum of the squares of  $m$  independent observations drawn from a normal universe with unit variance and zero mean is distributed according to the  $\chi^2$  distribution with  $m$  degrees of freedom. In the present instance, when the number of rows is not too small, the mean ranks can be treated as observations from a normal universe with a true mean  $\frac{1}{2}(p+1)$ . However, only  $p-1$  of the  $p$  mean ranks are independent, since the sum of the  $p$  mean ranks must equal  $\frac{1}{2}p(p+1)$ . If  $(p-1)$  of them were selected at random, the sum of the squared deviations from the true mean of  $\frac{1}{2}(p+1)$  would seem to be distributed as  $\chi^2$ . However, to discard one of the mean ranks would neglect some of the information; in addition, there is no criterion for deciding which to discard. Instead we can compute the mean squared deviation and multiply it by the number of degrees of freedom,  $(p-1)$ . This gives<sup>21</sup>

$$\frac{p-1}{p} \sum_{i=1}^p (\bar{r}_i - \rho)^2$$

as the numerator of  $\chi_r^2$ . The denominator must be  $\sigma^2$ , the variance of  $\bar{r}$ .

<sup>20</sup> On the question of the effect of departure from normality on the analysis of variance, see Egon S. Pearson, "The Analysis of Variance in Cases of Non-normal Variation," *Biometrika*, Vol. 23, 1931, and T. Eden and F. Yates, "On the Validity of Fisher's  $z$  Test When Applied to an Actual Example of Non-Normal Data," *Journal of Agricultural Science*, Vol. 23, 1933. The conclusion of both papers is that moderate departure from normality does not seriously affect the analysis of variance.

<sup>21</sup> By analogy with  $\chi^2$  as ordinarily defined, the multiplier  $(p-1)/p$  seems unnecessary. The difference is this. In the ordinary case we have a sum of squares *artificially lessened* because the deviations are computed from the observed *rather than* the true mean. Here, the observed mean is, of necessity, equal to the true mean. We thus have the sum of  $p$  squared deviations from the true mean, one of these, however, being essentially a duplication. This is evident when there are only two columns and the two deviations must be equal in absolute value; it is less obvious when there are more than two columns and the duplication is, as it were, spread among all of the deviations. A rigorous demonstration that  $(p-1)/p$  is the multiplier needed to correct for this duplication is provided by the proof in the mathematical appendix that the  $\chi_r^2$  distribution approaches that of  $\chi^2$ .

This statement, of course, is not a rigorous proof that the distribution of  $\chi_r^2$  approaches the distribution of  $\chi^2$  as  $n$  increases. A rigorous proof has, however, been provided by Dr. S. S. Wilks and is reproduced in the mathematical appendix.

In addition, the exact values of the first three moments of  $\chi_r^2$  have been derived.<sup>22</sup> The mean value is  $p - 1$ ; the variance,  $2(p - 1)(n - 1)/n$ ; and the third moment about the mean,  $8(p - 1)(n - 1)(n - 2)/n^2$ . The

TABLE V  
EXACT DISTRIBUTION OF  $\chi_r^2$  FOR TABLES WITH FROM 2 TO 9 SETS OF THREE RANKS

( $p=3$ ;  $n=2, 3, 4, 5, 6, 7, 8, 9$ )

$P$  is the probability of obtaining a value of  $\chi_r^2$  as great as or greater than the corresponding value of  $\chi_r^2$

n=2		n=5		n=7		n=8		n=9	
$\chi_r^2$	P	$\chi_r^2$	P	$\chi_r^2$	P	$\chi_r^2$	P	$\chi_r^2$	P
0	1.000	0.0	1.000	0.000	1.000	0.00	1.000	0.000	1.000
1	.833	0.4	.954	0.286	.964	0.25	.967	0.222	.971
3	.500	1.2	.691	0.857	.768	0.75	.794	0.667	.814
4	.167	1.6	.522	1.143	.620	1.00	.654	0.889	.865
n=3		2.8	.367	2.000	.486	1.75	.531	1.556	.569
$\chi_r^2$	P	3.6	.182	2.571	.305	2.25	.355	2.000	.398
0.000	1.000	4.8	.124	3.429	.237	3.00	.285	2.667	.328
0.667	.944	5.2	.093	3.714	.192	3.25	.236	2.889	.278
2.000	.528	6.4	.039	4.571	.112	4.00	.149	3.556	.187
2.667	.361	7.6	.024	5.429	.085	4.75	.120	4.222	.154
4.667	.194	8.4	.0085	6.000	.052	5.25	.079	4.667	.107
6.000	.028	10.0	.00077	7.143	.027	6.25	.047	5.556	.069
n=4		n=6		7.714	.021	6.75	.038	6.000	.057
$\chi_r^2$	P	$\chi_r^2$	P	8.000	.016	7.00	.030	6.222	.048
0.0	1.000	0.00	1.000	8.857	.0084	7.75	.018	6.889	.031
0.5	.931	0.33	.956	10.286	.0036	9.00	.0099	8.000	.019
1.5	.653	1.00	.740	10.571	.0027	9.25	.0080	8.222	.016
2.0	.431	1.33	.570	11.143	.0012	9.75	.0048	8.667	.010
3.5	.273	2.33	.430	12.286	.00032	10.75	.0024	9.556	.0060
4.5	.125	3.00	.252	14.000	.000021	12.00	.0011	10.667	.0035
6.0	.069	4.00	.184			12.25	.00086	10.889	.0029
6.5	.042	4.33	.142			13.00	.00026	11.556	.0013
8.0	.0046	5.33	.072			14.25	.000061	12.667	.00066
		6.33	.052			16.00	.0000036	13.556	.00035
		7.00	.029					14.000	.00020
		8.33	.012					14.222	.000097
		9.00	.0081					14.889	.000054
		9.33	.0055					16.222	.000011
		10.33	.0017					18.000	.0000006
		12.00	.00013						

corresponding values for the  $\chi^2$  distribution with  $p - 1$  degrees of freedom are  $p - 1$ ,  $2(p - 1)$ , and  $8(p - 1)$ , respectively. It follows that  $\chi_r^2$  and  $\chi^2$  always have the same mean value, and that the variance and

<sup>22</sup> I am indebted to Mr. William C. Shelton for the derivation of the mean value and for suggesting the method of deriving the other moments.

third moment of  $\chi_r^2$  approach the variance and third moment of  $\chi^2$  as  $n$  increases.

For the special case of  $p=3$ , the exact distribution of  $\chi_r^2$  has been derived for  $n$  from 2 to 9; and for  $p=4$ , for  $n$  equal to 2, 3 and 4.<sup>23</sup> Table V gives the distributions for  $p=3$ , and Table VI for  $p=4$ . These distributions give some empirical indication of how rapidly the  $\chi_r^2$  distribution approaches the  $\chi^2$  distribution; in addition, they can be used to make exact tests for small values of  $n$  and  $p$ .

TABLE VI  
EXACT DISTRIBUTION OF  $\chi_r^2$  FOR TABLES WITH FROM 2 TO 4 SETS OF FOUR RANKS

( $p=4; n=2, 3, 4$ )

$P$  is the probability of obtaining a value of  $\chi_r^2$  as great as or greater than the corresponding value of  $\chi_r^2$ .

$n=2$		$n=3$		$n=4$			
$\chi_r^2$	$P$	$\chi_r^2$	$P$	$\chi_r^2$	$P$	$\chi_r^2$	$P$
0.0	1.000	0.2	1.000	0.0	1.000	5.7	.141
0.6	.958	0.6	.958	0.3	.992	6.0	.105
1.2	.834	1.0	.910	0.6	.928	6.3	.094
1.8	.792	1.8	.727	0.9	.900	6.6	.077
2.4	.625	2.2	.608	1.2	.800	6.9	.068
3.0	.542	2.6	.524	1.5	.754	7.2	.054
3.6	.458	3.4	.446	1.8	.677	7.5	.052
4.2	.375	3.8	.342	2.1	.649	7.8	.036
4.8	.208	4.2	.300	2.4	.524	8.1	.033
5.4	.167	5.0	.207	2.7	.508	8.4	.019
6.0	.042	5.4	.175	3.0	.432	8.7	.014
		5.8	.148	3.3	.389	9.3	.012
		6.6	.075	3.6	.355	9.6	.0069
		7.0	.054	3.9	.324	9.9	.0062
		7.4	.033	4.5	.242	10.2	.0027
		8.2	.017	4.8	.200	10.8	.0016
		9.0	.0017	5.1	.190	11.1	.00094
				5.4	.158	12.0	.000072

The tables show that if we adopt .01 as a level of significance, then for  $p=3$ , it is impossible to obtain a significant value for  $n$  less than 4, and for  $n=4$  only perfect consistency will yield a significant value; for  $n=5$ , two values will satisfy the criterion and for  $n=6$ , four values. If .05 is adopted as a level of significance, only perfect consistency is significant for  $n=3$ , while 2 values are significant for  $n=4$ , and four values for  $n=5$ .

For  $p=4$ , the .01 criterion cannot be satisfied for  $n=2$ , is satisfied by perfect consistency for  $n=3$ , and by 6 values for  $n=4$ . The .05

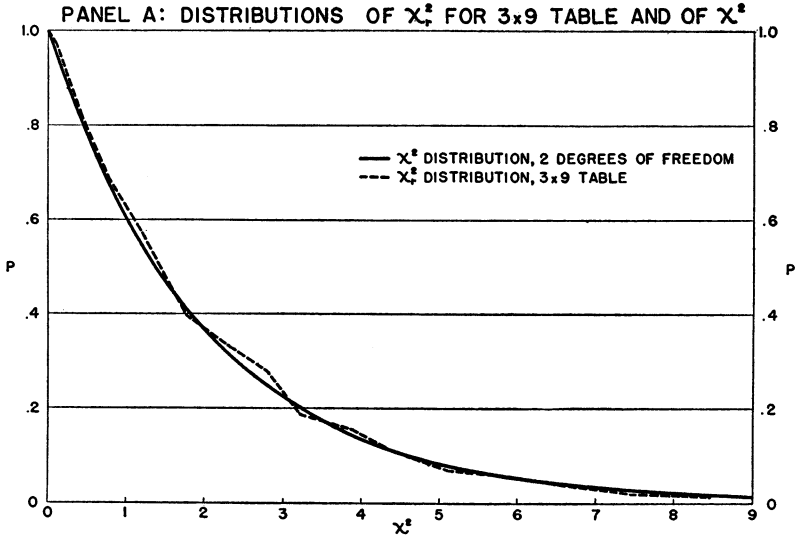
<sup>23</sup> These distributions were derived by the rather laborious process of building up the distribution for each value of  $n$  from the distribution for the next smaller value. The labor involved increases greatly as  $n$  increases, and even more rapidly as  $p$  increases.



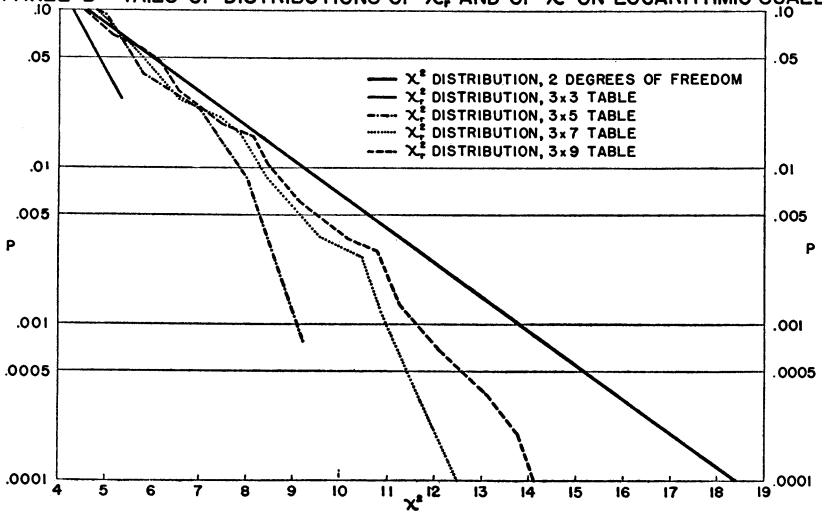
criterion is satisfied by one value for  $n=2$ , three values for  $n=3$ , and 11 values for  $n=4$ .

CHART 1

COMPARISON OF DISTRIBUTIONS OF  $\chi_r^2$  AND  $\chi^2$  FOR TWO DEGREES OF FREEDOM



PANEL B: TAILS OF DISTRIBUTIONS OF  $\chi_r^2$  AND OF  $\chi^2$  ON LOGARITHMIC SCALE



The comparison of the  $\chi_r^2$  distribution with the  $\chi^2$  distribution is shown in Chart 1 for  $p=3$  and in Chart 2 for  $p=4$ . In making this com-

parison it is necessary to allow for the discontinuity of the  $\chi_r^2$  distribution. Only a discrete number of finite values of  $\chi_r^2$  are possible while  $\chi^2$  is continuous. The probability associated with any  $\chi_r^2$  in Tables V and VI must thus be considered as corresponding to a value of  $\chi^2$  intermediate between that value of  $\chi_r^2$  and the immediately preceding value. This intermediate value has been arbitrarily chosen as halfway between the two values of  $\chi_r^2$ . It is these intermediate values which form the abscissas of the points plotted in the charts.

Panel A of Chart 1 compares the  $\chi_r^2$  distribution for a  $3 \times 9$  table with the  $\chi^2$  distribution for 2 degrees of freedom. For convenience, the distributions have been compared in cumulative form. The ordinate gives the probability of securing a value of  $\chi^2$  or  $\chi_r^2$  as great as or greater than the abscissa. The solid line gives the  $\chi^2$  distribution, the dotted line, the  $\chi_r^2$  distribution. The agreement between the two distributions is very close. The cumulative curve for the  $\chi_r^2$  distribution tends to be somewhat above that for the  $\chi^2$  distribution for low values of  $\chi_r^2$  and below it for high values. This is to be expected since  $\chi_r^2$  must be less than a fixed finite value (that corresponding to perfect consistency) while  $\chi^2$  is not so limited.

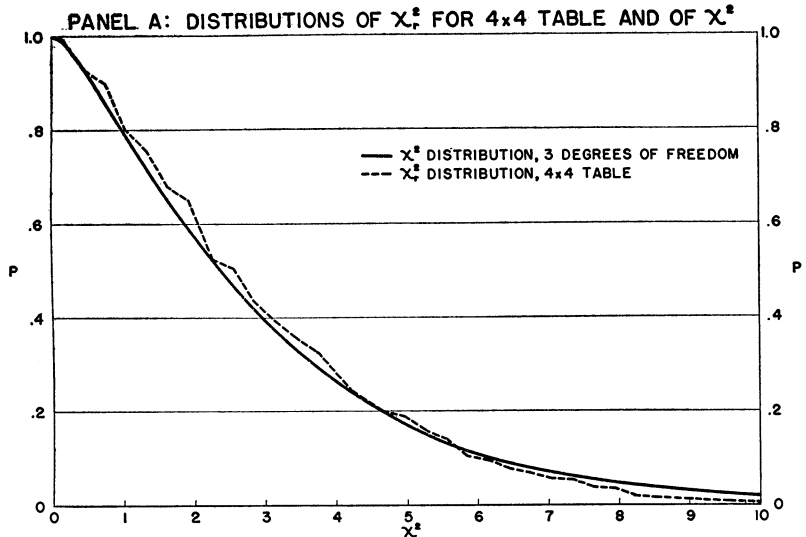
In utilizing tests of significance, it is the small values of  $P$ , i.e., one of the tails of the distribution, with which we are ordinarily concerned. In order to bring out more clearly the behavior of this part of the  $\chi_r^2$  distribution, a logarithmic scale is used for the probability in Panel B of Chart 1. This panel gives the cumulative distribution of  $\chi_r^2$  to the right of  $P = .10$  for  $p = 3$ , and  $n = 3, 5, 7, 9$ ; as well as the corresponding portion of the  $\chi^2$  distribution. The chart shows the tendency of this portion of the  $\chi_r^2$  curve to be below the  $\chi^2$  curve. It also clearly indicates the tendency for the  $\chi_r^2$  distribution to approach the  $\chi^2$  distribution and suggests that it does so fairly rapidly.

Panels A and B of Chart 2 are similar to the corresponding panels of Chart 1, but relate to  $p = 4$ . Panel A compares the  $\chi_r^2$  distribution for a  $4 \times 4$  table with the  $\chi^2$  distribution for three degrees of freedom. The agreement is good, although, because of the smaller value of  $n$ , the discrepancies are somewhat greater than in Panel A of Chart 1. It indicates the same consistent tendency for the  $\chi_r^2$  distribution to be above the  $\chi^2$  distribution for small values of  $\chi_r^2$  and below it for large values. Panel B gives the portion of the cumulative distribution of  $\chi_r^2$  to the right of  $p = .10$ , for  $p = 4$  and  $n = 2, 3, 4$ , as well as the corresponding portion of the  $\chi^2$  distribution. Once again the vertical scale is logarithmic. The tendency for the  $\chi_r^2$  distribution to approach the  $\chi^2$  distribution very rapidly as  $n$  increases, is plain.

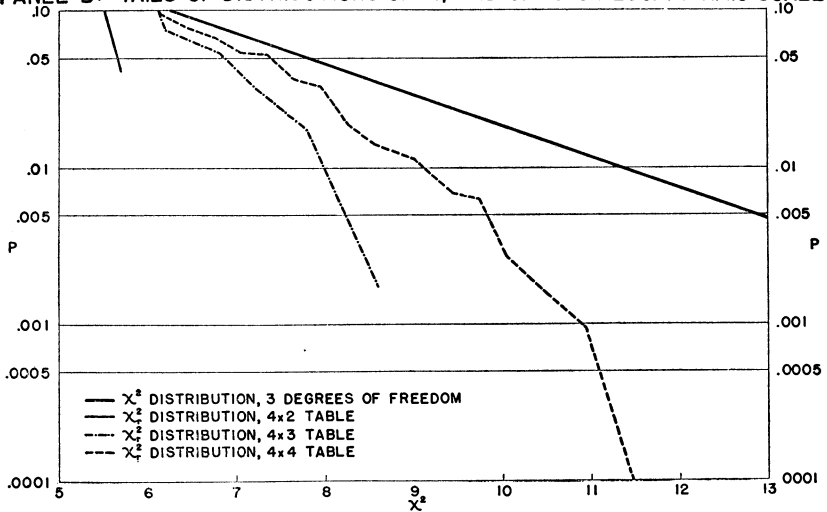
The tendency for the values of  $\chi_r^2$  adjusted for discontinuity to be less than  $\chi^2$  for small probabilities suggests that any errors resulting

CHART 2

COMPARISON OF DISTRIBUTIONS OF  $\chi_r^2$  AND  $\chi^2$  FOR THREE DEGREES OF FREEDOM



**PANEL B: TAILS OF DISTRIBUTIONS OF  $\chi_r^2$  AND OF  $\chi^2$  ON LOGARITHMIC SCALE**



from using the  $\chi^2$  distribution as an approximation to the  $\chi_r^2$  distribution are likely to be in the proper direction—that is, the significance of

results will be understated rather than exaggerated. This tendency toward under-statement is compensated—indeed, in some cases over-compensated—by the fact that the values of  $\chi_r^2$  which can be observed (i.e., the values of  $\chi_r^2$  not adjusted for discontinuity) are always greater than the adjusted values. This factor is of minor significance, however, since the number of possible values of  $\chi_r^2$  increases very rapidly—and hence the interval between them decreases very rapidly—as either  $p$  or  $n$  increases. Even for  $p$  and  $n$  both as small as 4, the difference between the adjusted and unadjusted values of  $\chi_r^2$  is, for practical purposes, negligible. It is .15 in all but four cases, .3 in three of these, and .45 in the remaining one.

A comparison of the  $\chi^2$  and  $\chi_r^2$  distributions at the critical points sheds further light on this problem. For  $p=3$  the value of  $\chi^2$  corresponding to  $P=.05$  is 5.991. From Table V, the nearest value which  $\chi_r^2$  can have for  $p=3$ ,  $n=9$  is 6, and this has a probability of .057 associated with it. Thus, by using the  $\chi^2$  distribution we should be led to overestimate slightly the significance of a value of  $\chi_r^2=6$ . The next higher value of  $\chi_r^2$  is 6.22, and its significance we should estimate properly, since the probability associated with it is .048. The value of  $\chi^2$  corresponding to  $P=.01$  is 9.21. From Table V, the nearest values which  $\chi_r^2$  can have are 8.67, with a probability of .0103, and 9.55 with a probability of .0060. In this case, the use of the  $\chi^2$  distribution would yield the correct results; 8.66 would be attributed a probability greater than .01 and 9.55 one less than .01.

For  $p=4$ , the value of  $\chi^2$  corresponding to  $P=.05$  is 7.815. The nearest values of  $\chi_r^2$  for  $p=4$  and  $n=4$ , as given in Table VI, are 7.5 with a probability of .052, and 7.8 with a probability of .036. The .01 value of  $\chi^2$  is 11.341. From the table, 9.3 has a probability of .012, 9.6 of .0069, and 11.1 of .00094. Here, the use of  $\chi^2$  would in each case understate the significance of  $\chi_r^2$ .

While no definitive conclusions can be drawn from these comparisons they suggest that for  $p=3$ , the use of the  $\chi^2$  distribution is likely to give sufficiently accurate results for  $n$  greater than 9; while for  $p=4$ , the use of the  $\chi^2$  distribution is likely to understate the significance of large values of  $\chi_r^2$  unless  $n$  is somewhat larger than 4. In view of the apparent rapidity with which the  $\chi_r^2$  distribution approaches the  $\chi^2$  distribution when  $p=4$ , it seems reasonable that for  $n$  equal to or greater than 6, the  $\chi^2$  distribution will give sufficiently accurate results. For  $p$  greater than 4 it is more difficult to make any general statement; but it seems safe to say that the  $\chi^2$  distribution will give fairly accurate

results for  $n$  equal to or greater than 6.<sup>24</sup> A procedure that seems applicable when  $p$  is quite large and  $n$  less than 6 is discussed below.

#### RELATION BETWEEN $\chi_r^2$ AND THE RANK DIFFERENCE CORRELATION

When only two sets of ranks are available, the appropriate measure of relationship is the rank difference correlation coefficient,  $r'$ . This coefficient is computed by the usual product-moment formula with the ranks serving as the variables, or by the equivalent, but more convenient, formula

$$r' = 1 - \frac{6 \sum d^2}{p^3 - p},$$

where  $d$  is the difference between two paired values, and  $p$ , as above, is the number of pairs of ranks.<sup>25</sup>

For  $n=2$ ,  $\chi_r^2$  uses the same data as the rank difference correlation and is designed to test the same hypothesis. The two are, therefore, essentially equivalent. It is shown in the mathematical appendix that the relation between them is

$$\chi_r^2 = (p - 1)(1 + r').$$

For  $n=2$  testing the significance of  $r'$  is thus equivalent to testing the significance of  $\chi_r^2$ .

Under the hypothesis of homogeneity, the mean value of  $r'$  is zero, and its variance,  $1/(p-1)$ .<sup>26</sup> It follows from the last equation that, for  $n=2$ , both the mean value and variance of  $\chi_r^2$  are  $(p-1)$ . These results agree, of course, with the more general formulae given above.

#### THE APPROACH TO NORMALITY

Hotelling and Pabst have shown that  $r'$  tends to become normally distributed as  $p$  increases. It follows that for  $n=2$ ,  $\chi_r^2$  tends also to become normally distributed as  $p$  increases.

When  $n$  is large the distribution of  $\chi_r^2$  approaches that of  $\chi^2$  and the latter approaches normality as the number of degrees of freedom increases.

Since, for the smallest value of  $n$  as well as for large values, the distribution of  $\chi_r^2$  tends to normality as  $p$  increases, it seems reasonable to assume that for intermediate values of  $n$  it behaves similarly. As-

<sup>24</sup> It is worth recalling that the rapidity with which the variance of the  $\chi_r^2$  distribution approaches the variance of the  $\chi^2$  distribution depends solely on  $n$  and not at all on  $p$ . On the other hand, the number of distinct values of  $\chi_r^2$  depends on both  $p$  and  $n$ .

<sup>25</sup> This is the usual notation except that the number of pairs of ranks is ordinarily designated as  $n$ . The present notation is used in order to preserve consistency with the preceding analysis.

<sup>26</sup> Hotelling and Pabst, *op. cit.*, p. 36.

suming this to be the case, then for small values of  $n$  and large values of  $p$  the significance of  $\chi_r^2$  can be tested by considering

$$\frac{\chi_r^2 - (p - 1)}{2 \frac{n - 1}{n} (p - 1)}$$

as a normally distributed variate with zero mean and unit standard deviation.

Further study is clearly needed on this point, both in order to obtain a rigorous proof that for small values of  $n$  the  $\chi_r^2$  distribution tends to normality as  $p$  increases, and also to determine the rapidity with which it approaches normality.

#### CONCLUSION

The method of ranks is a method which can be applied to data classified by two (or more) criteria to determine whether the factors used as criteria of classification have a significant influence on the variate classified. Stated differently, the method tests the hypothesis that the values of the variate corresponding to each subdivision by one of the factors are homogeneous, i.e., from the same universe. The method uses solely information on "order" and makes no use of the quantitative values of the variate as such. For this reason no assumption need be made as to the nature of the underlying universe or as to whether the different sets of values come from similar universes. The method is thus applicable to a wide class of problems to which the analysis of variance, designed to test a similar hypothesis, cannot validly be applied.

The basic step in the application of the method of ranks is the computation of a statistic,  $\chi_r^2$ , from a table of ranks. The sampling distribution of this statistic approaches the  $\chi^2$  distribution as the number of sets of ranks increases. When the number of sets of ranks is moderately large (say greater than 5 for four or more ranks) the significance of  $\chi_r^2$  can be tested by reference to the available  $\chi^2$  tables. When the number of ranks in each set is 3, and the number of sets 9 or less, or when the number of ranks in each set is 4, and the number of sets 4 or less, the significance of  $\chi_r^2$  can be tested by reference to the exact tables given above. When, however, both the number of ranks and the number of sets of ranks are very small, it is impossible to obtain significant results.

When the number of ranks is large, but the number of sets of ranks small, there is reason to suppose—though no rigorous proof is avail-

able—that  $\chi_r^2$  is normally distributed about a mean of  $p-1$  and with a variance of  $2(p-1)(n-1)/n$ , where  $p$  is the number of ranks and  $n$  the number of sets of ranks. In such cases, then, the significance of  $\chi_r^2$  can be tested by reference to tables of the normal curve.

The theoretical discussion of the efficiency of the method of ranks relative to the analysis of variance indicates that in situations when the latter method can validly be applied and when the number of sets of ranks is large the maximum loss of information through using the analysis of ranks is 36 per cent. The minimum loss is probably 9 per cent. The amount of information lost appears to be greatest when there are only two ranks in each set, and decreases as the number of ranks increases.

The application of the two methods to the same body of data provides further evidence as to their relative efficiency. The data employed were classified into five groups by one of the factors and into seven by the other. The results suggested that in this instance the loss of information through using the method of ranks was not very great, that both methods tended to yield the same result.

The method of ranks requires less than one-fourth as much time as the analysis of variance. In the light of the conclusions just stated concerning their relative efficiency, this suggests that even though the assumptions necessary for the latter method are known to be satisfied, if the problem of computation is a serious one, the method of ranks might profitably be used as an alternative to the analysis of variance or, at least, as a preliminary method to suggest fruitful hypotheses which might then be more accurately tested by the analysis of variance.

#### MATHEMATICAL APPENDIX

1. *Proof that the  $\chi_r^2$  distribution approaches the  $\chi^2$  distribution as  $n$  increases.*<sup>27</sup>

Let  $r_{ij}$  = the rank in the  $i$ -th row and  $j$ -th column, ( $i=1, \dots, n; j=1, \dots, p$ )

$$(1) \quad r'_{ij} = r_{ij} - \frac{1}{2}(p+1)$$

and

$$(2) \quad \bar{r}_j' = \frac{1}{n} \sum_{i=1}^n r'_{ij}.$$

The characteristic function of the quantities  $\bar{r}_j'$  ( $j=1, \dots, p$ ) is given by

$$(3) \quad \phi = E \left( \exp i \sum_{j=1}^{p-1} \theta_j \bar{r}_j' \right)$$

<sup>27</sup> This proof is adapted from one given by Dr. S. S. Wilks in a letter to the author.

where  $E$  stands for expected value. Only  $p-1$  of the  $\bar{r}'_j$ 's are included because  $\bar{r}'_p$  is expressible in terms of  $\bar{r}'_1, \dots, \bar{r}'_{p-1}$ . This in turn follows from the fact that, for each value of  $i$ ,  $r_{ij}$  takes all the values from 1 to  $p$  as  $j$  varies. Substituting from (2) into (3) we have

$$(4) \quad \phi = E \left( \exp \frac{i}{n} \sum_{i=1}^n \sum_{j=1}^{p-1} \theta_j r'_{ij} \right).$$

Since the set of ranks in each row is independent of the set of ranks in any other row

$$(5) \quad \phi = \left[ E \left( \exp \frac{i}{n} \sum_{j=1}^{p-1} \theta_j r'_j \right) \right]^n$$

where  $r'_j$  stands for any of the sets of  $r'_{ij}$ . Expanding,

$$(6) \quad \phi = \left\{ E \left[ 1 + \frac{i}{n} \sum_{j=1}^{p-1} \theta_j r'_j + \frac{i^2}{2n^2} \left( \sum_{j=1}^{p-1} \theta_j r'_j \right)^2 + \frac{1}{n^3} R' \right] \right\}^n$$

or

$$(7) \quad \phi = \left\{ E \left[ 1 + \frac{i}{n} \sum_{j=1}^{p-1} \theta_j r'_j + \frac{i^2}{2n^2} \left( \sum_{j=1}^{p-1} \theta_j^2 r_j'^2 + 2 \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \theta_j \theta_{j'} r'_j r'_{j'} \right) + \frac{1}{n^3} R' \right] \right\}^n.$$

But since  $r'_j$  takes all of the  $p$  values differing by unity from  $-\frac{1}{2}(p-1)$  to  $\frac{1}{2}(p-1)$  with equal probability

$$(8) \quad E r'_j = 0,$$

$$(9) \quad E r_j'^2 = \frac{1}{p} \sum_{r_j' = -(p-1)/2}^{(p-1)/2} r_j'^2 = (p^2 - 1)/12.$$

Further

$$(10) \quad E r'_j r'_{j'} = - (p + 1)/12$$

since

$$(11) \quad \left( \sum_{r_j' = -(p-1)/2}^{(p-1)/2} r_j' \right)^2 = \sum_{r_j' = -(p-1)/2}^{(p-1)/2} r_j'^2 + 2 \sum_{r_j' = -(p-1)/2}^{(p-3)/2} \sum_{r'_j = r'_j + 1}^{(p-1)/2} r'_j r'_{j'} = 0$$

and hence

$$(12) \quad p(p - 1) E r'_j r'_{j'} = - p E r_j'^2 = - p(p^2 - 1)/12.$$



Using these results in (7) gives

$$(13) \quad \phi = \left\{ 1 - \frac{1}{2n^2} \left[ \frac{p^2 - 1}{12} \sum_{j=1}^{p-1} \theta_j^2 - 2 \frac{p+1}{12} \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \theta_j \theta_{j'} \right] + \frac{1}{n^3} R \right\}^n$$

where  $R$  is a bounded function of  $p, r_1', \dots, r_{p-1}'$  for  $n = 1, 2, 3, \dots$ .

Allowing  $n$  to approach infinity, we have

$$(14) \quad \phi \cong \exp - \frac{p^2 - 1}{24n} \left( \sum_{j=1}^{p-1} \theta_j^2 - \frac{2}{p-1} \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \theta_j \theta_{j'} \right).$$

This, however, is the characteristic function for a multivariate normal distribution. It follows that  $\bar{r}_1', \dots, \bar{r}_{p-1}'$  are asymptotically normally distributed with a matrix of variances and covariances given by the matrix of the quadratic in  $\theta_1, \dots, \theta_{p-1}$ .

Taking the reciprocal of the matrix of the  $\theta$ 's and associating it with the  $\bar{r}_j$ 's we have as the distribution function of the  $\bar{r}_j$ 's:

$$(15) \quad C \exp \left\{ - \frac{1}{2} \frac{12n}{p(p+1)} \left( 2 \sum_{j=1}^{p-1} \bar{r}_j'^2 + 2 \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \bar{r}_j' \bar{r}_{j'}' \right) \right\} d\bar{r}_1' d\bar{r}_2' \dots d\bar{r}_{p-1}'$$

where  $C$  is a constant

Since  $\sum_{j=1}^p \bar{r}_j' = 0$  it follows that  $\bar{r}_p' = -\sum_{j=1}^{p-1} \bar{r}_j'$  and hence

$$(16) \quad \bar{r}_p'^2 = \left( \sum_{j=1}^{p-1} \bar{r}_j' \right)^2 = \sum_{j=1}^{p-1} \bar{r}_j'^2 + 2 \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \bar{r}_j' \bar{r}_{j'}'.$$

Substituting (16) in the exponent of (15) we have, finally, for the distribution function

$$(17) \quad C \exp \left\{ - \frac{1}{2} \frac{12n}{p(p+1)} \sum_{j=1}^p \bar{r}_j'^2 \right\} d\bar{r}_1' d\bar{r}_2' \dots d\bar{r}_{p-1}' \\ = C \exp \left( - \frac{1}{2} \chi_r^2 \right) d\bar{r}_1' d\bar{r}_2' \dots d\bar{r}_{p-1}',$$

by the definition of  $\chi_r^2$ . It follows that for  $n$  large  $\chi_r^2$  is distributed like  $\chi^2$  with  $p-1$  degrees of freedom.

### 2. Derivation of the exact moments<sup>28</sup> of $\chi_r^2$ .

By definition

$$(18) \quad \chi_r^2 = \frac{12n}{p(p+1)} \sum_{j=1}^p \bar{r}_j'^2 = \frac{12n}{p(p+1)} \frac{1}{n^2} \sum_{j=1}^p \left( \sum_{i=1}^n r'_{ij} \right)^2.$$

<sup>28</sup> The derivation of the mean value of  $\chi_r^2$  is adapted from one communicated to the author by Mr. William C. Shelton, who also suggested the method used for deriving the other moments. The method employed is essentially that developed by J. Splawa-Neyman, "Contributions to the Theory of Small Samples Drawn from a Finite Population," *Biometrika*, Vol. 17, 1925, pages 472-79.

Expanding, replacing  $\sum_{j=1}^p \sum_{i=1}^n r_{ij}^2$  by its value  $np(p^2-1)/12$ , and rearranging the order of the summation signs we have

$$(19) \quad \chi_r^2 = (p-1) + \frac{24}{p(p+1)n} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j}.$$

Taking the expected value of both sides gives

$$(20) \quad E\chi_r^2 = (p-1) + \frac{24}{p(p+1)n} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p E(r'_{ij} r'_{i'j}).$$

But since one row of the table of ranks is entirely independent of any other row,  $E r'_{ij} r'_{i'j} = E r'_{ij} E r'_{i'j} = 0$ .

$$(21) \quad \therefore E\chi_r^2 = p-1.$$

From (19) and (21)

$$(22) \quad \chi_r^2 - E\chi_r^2 = \frac{24}{p(p+1)n} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j}.$$

The  $k$ -th moment of  $\chi_r^2$  about its mean value can therefore be obtained by evaluating the expected value of the  $k$ -th power of the right hand side of (22).

To determine the variance of  $\chi_r^2$  first note that  $\sum_{j=1}^p r'_{ij} r'_{i'j}$  is independent of  $\sum_{j=1}^p r'_{i'j} r'_{i''j}$ . This can be proved by multiplying the two expressions. The expected value of the resultant product is easily shown to be zero. Likewise,  $\sum_{j=1}^p r''_{ij} r'_{i'j}$  is independent of  $\sum_{j=1}^p r'_{i'j} r'_{i''j}$ . It follows that

$$(23) \quad E \left[ \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j} \right]^2 = \sum_{i=1}^{n-1} \sum_{i'=i+1}^n E \left( \sum_{j=1}^p r'_{ij} r'_{i'j} \right)^2.$$

But

$$(24) \quad \begin{aligned} E \left( \sum_{j=1}^p r'_{ij} r'_{i'j} \right)^2 &= E \left( \sum_{j=1}^p r'_{ij}{}^2 r'_{i'j}{}^2 + 2 \sum_{j=1}^{p-1} \sum_{j'=j+1}^p r'_{ij} r'_{i'j} r'_{ij'} r'_{i'j'} \right) \\ &= \sum_{j=1}^p E r'_{ij}{}^2 E r'_{i'j}{}^2 + 2 \sum_{j=1}^{p-1} \sum_{j'=j+1}^p E(r'_{ij} r'_{ij'}) E(r'_{i'j} r'_{i'j'}). \end{aligned}$$

Substituting (9) and (10) into (24) and the resultant expression into (23) gives

$$(25) \quad E \left[ \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j} \right]^2 = \frac{n(n-1)}{2} \frac{p^2(p-1)(p+1)^2}{12^2}.$$

Multiplying this by  $[p(p+1)n/24]^{-2}$  gives, finally, the variance of  $\chi_r^2$

$$(26) \quad \sigma^2 = 2 \frac{n-1}{n} (p-1).$$

To determine the third moment of  $\chi_r^2$  about its mean note that the only term in the expansion of

$$\left[ \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j} \right]^3$$

whose expected value is not zero is

$$(27) \quad 6 \sum_{i=1}^{n-2} \sum_{i'=i+1}^{n-1} \sum_{i''=i'+1}^n \left[ \sum_{j=1}^p r'_{ij} r'_{i'j} \sum_{j=1}^p r'_{ij} r'_{i''j} \sum_{j=1}^p r'_{i'j} r'_{i''j} \right].$$

Expanding the expression in brackets gives

$$(28) \quad \sum_{j=1}^p r'_{ij}{}^2 r'_{i'j}{}^2 r'_{i''j}{}^2 + 3 \sum_{j=1}^{p-1} \sum_{j'=j+1}^p r'_{ij}{}^2 r'_{i'j} r'_{i''j} r'_{i'j'} r'_{i''j'} r'_{i'j''} \\ + 6 \sum_{j=1}^{p-2} \sum_{j'=j+1}^{p-1} \sum_{j''=j'+1}^p r'_{ij} r'_{i'j} r'_{i''j} r'_{i'j'} r'_{i''j'} r'_{i'j''} r'_{i''j''}.$$

Taking expected values, and substituting from (9) and (10) gives  $p^3(p-1)(p+1)^3/12^3$  as the expected value of (28). Substituting this in (27) gives

$$(29) \quad E \left[ \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \sum_{j=1}^p r'_{ij} r'_{i'j} \right]^3 = \frac{6n(n-1)(n-2)}{1 \cdot 2 \cdot 3} \frac{p^3(p-1)(p+1)^3}{12^3}.$$

Multiplying this by  $[p(p+1)n/24]^{-3}$  gives as the third moment of  $\chi_r^2$  about its mean

$$(30) \quad \mu_3 = \frac{8(n-1)(n-2)}{n^2} (p-1).$$

3. *Derivation of relationship between  $\chi_r^2$  and rank difference correlation ( $r'$ ) when  $n=2$ .*

From (19) we have, for  $n=2$

$$(31) \quad \chi_r^2 = (p - 1) + \frac{12}{p(p + 1)} \sum_{i=1}^p r'_{1i} r'_{2i} .$$

But, using the product moment formula, the rank difference correlation coefficient is defined as<sup>29</sup>

$$(32) \quad r' = \frac{\sum_{i=1}^p r'_{1i} r'_{2i}}{\sqrt{\sum_{i=1}^p r'_{1i}{}^2 \sum_{i=1}^p r'_{2i}{}^2}} = \frac{12}{p(p^2 - 1)} \sum_{i=1}^p r'_{1i} r'_{2i} .$$

Substituting in (31) this gives

$$(33) \quad \chi_r^2 = (p - 1)(1 + r')$$

as the relation between  $\chi_r^2$  and the rank difference correlation coefficient.

<sup>29</sup> The notation used in (32) may be somewhat confusing. The symbol  $r'$  which stands for the rank difference correlation coefficient is to be distinguished from  $r'_{ij}$  which stands for the deviate of a rank from its expected value.