Rank Correlation and Population Models
Author(s): H. E. Daniels

Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 12, No. 2 (1950), pp. 171-191
Published by: Blackwell Publishing for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983980
Accessed: 01/07/2010 14:17

Rank Correlation and Population Models

By H. E. Daniels

Statistical Laboratory, University of Cambridge

1. The purpose of this paper is to examine the relation between rank correlation and specific features of the population about which information is required. The paper was written after those by P. A. P. Moran and J. W. Whitfield were in draft form, and with their permission I have commented on one or two points arising out of their work.

The first question I wish to consider is how far Kendall's τ and Spearman's ρ give similar information about a bivariate population of ranks. The population may be ranked from that of an underlying variable, or it may be a population of ranks in its own right. Both measures of rank correlation are usually considered reasonable in that they measure in some sense the "degree of agreement" between two rank orders, and I have given a similar justification of the general coefficient Γ (1948). It is sometimes assumed that because the sample estimates* T and R of τ and ρ are highly correlated in the case of independence, τ and ρ describe more or less the same aspect of a separate bivariate population of ranks when the correlation is not zero. The following inequality shows that the assumption may be far from true.

2. It is convenient to start with a finite sample of n which is subsequently made indefinitely large.

Let $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$ be two rankings, and assign scores $a_{ij} = \operatorname{sgn}(p_j - p_i)$, $b_{ij} = \operatorname{sgn}(q_j - q_i)$ to the corresponding pairs of ranks, taking $a_{ii} = b_{ii} = 0$. Then

$$\sum_{j=1}^{n} a_{ij} = -2(p_i - \bar{p})$$

and similarly for b_{ij}, so that

$$T = \frac{1}{n(n-1)} \Sigma\, a_{ij}\, b_{ij}, \quad R = \frac{3}{n(n^2-1)} \Sigma\, a_{ij}\, b_{ik},$$

all suffixes being summed from 1 to n. Since relations like $p_i > p_j > p_k > p_i$ are impossible, a_{ij}, a_{jk}, a_{ki} cannot be all positive or all negative, hence $a_{ij} + a_{jk} + a_{ki} = \pm 1$, and similarly $b_{ij} + b_{jk} + b_{ki} = \pm 1$. Moreover, we can write

$$(a_{ij} + a_{jk} + a_{ki})(b_{ij} + b_{jk} + b_{ki}) = \varepsilon_{ijk},$$

where ε_{ijk} has the following meaning. Let p_i, p_j, p_k and q_i, q_j, q_k be renamed according to their order of magnitude, becoming p'_i, p'_j, p'_k; q'_i, q'_j, q'_k, where p', q' are now permutations of 123. For example the ranks 472 become 231. Then $\varepsilon_{ijk} = +1$ or -1 according as p' is an even or odd permutation of q', and it is zero if any pair of suffixes is equal.

Summing over all values of the suffixes we find

$$3n\, \Sigma\, a_{ij}\, b_{ij} - 6\, \Sigma\, a_{ij}\, b_{ik} = \Sigma\, \varepsilon_{ijk}.$$

* I prefer the symbol T to t, which is standard for Student's ratio. As Whitfield and Moran both use different notations, there seems no harm in introducing a third one.

The quantity

$$U = \frac{1}{n(n-1)(n-2)} \Sigma \, \varepsilon_{ijk}$$

is the fraction of "even" triplets out of all possible triplets. The equation may then be written

$$3nT - 2(n+1)R = (n-2)U,$$

and since $-1 \leqslant U \leqslant 1$ we obtain the inequality

$$-(n-2) \leqslant 3nT - 2(n+1)R \leqslant (n-2).$$

Writing ν for the population value of U, and making n large, the relations for the population become

$$3\tau - 2\rho = \nu$$

$$-1 \leqslant 3\tau - 2\rho \leqslant 1.$$

The similarity is worth noting between the T, R, U relation and the approximate regression formula $3T = 2R$ established by Kendall (1948, p. 68) for the case of independence. U may be regarded in this sense as a residual.

3. The inequality is of interest only if populations exist for which $3\tau - 2\rho$ has values near the limits. I now show that if $\tau > 0$ the upper limit can be attained but not the lower limit: when $\tau < 0$ the reverse is true, and when $\tau = 0$ both limits are attainable.

Start as before with rankings of n. The q ranking can always be chosen in the natural order $123 \ldots n$. Then if the p ranking is a cyclic permutation of the natural order, that is, of the form $m+1, m+2, \ldots, n, 1, 2, \ldots, m$, all ε_{ijk}'s are $+1$, $U = 1$ and the upper limit is attained. It is, moreover, not difficult to show that no other p ranking has every $\varepsilon_{ijk} = 1$. On the other hand if, and only if, the p ranking is a cyclic permutation of $n, n-1, \ldots 3, 2, 1$, all ε_{ijk}'s are -1 and the lower limit is attained.

For the ranking $m+1, m+2, \ldots n, 1, 2, \ldots m$ the value of T is

$$T = \frac{(n-2m)^2 - n}{n(n-1)},$$

which has the minimum value $-\dfrac{1}{(n-1)}$ at $m = \frac{1}{2}n$ if n is even, or $-\dfrac{1}{n}$ at $m = \frac{1}{2}(n \pm 1)$ if n is odd. When n is made indefinitely large, the range of τ for populations of this type is effectively $0 \leqslant \tau \leqslant 1$, both limits being attained. Thus for such values of τ there are populations having $3\tau - 2\rho = 1$, and in the same way cyclic permutations of $n, n-1, \ldots 3, 2, 1$ lead to populations with $-1 \leqslant \tau \leqslant 0$ having $3\tau - 2\rho = -1$.

The fact that $\nu = -1$ implies $\tau \leqslant 0$ is sufficient to show that the lower limit is never actually attained when $\tau > 0$, but except for small values of τ it is not even a good lower bound, for the fact that $\rho \leqslant 1$ means that $\nu \geqslant 3\tau - 2$, which may be far from -1 when $\tau > \frac{1}{3}$. The determination of a precise lower bound for ν when $\tau > 0$ is a problem of some difficulty, and I should be interested to see it solved.

When $\tau = 0$ the two types of population mentioned will provide examples of $\nu = 1$ and $\nu = -1$, and we arrive at the rather surprising conclusion that it is possible to have populations in which $\tau = 0$ and $\rho = \frac{1}{2}$ or $-\frac{1}{2}$. Clearly in such cases τ and ρ are describing very different aspects of the population. The examples are admittedly extreme, but similar less violent discrepancies between τ and ρ probably exist in other cases.

4. The practising statistician will object, with some justification, that while such a result is theoretically possible, it is highly unlikely that populations met with in the world of experience are of this form. I think situations approximating to the extreme cases may occur. For example, suppose an inquiry is conducted into the relation between intelligence test score and salary in a

mixed population which, unknown to the investigator, consists of two groups following different professions, one group being of markedly lower average intelligence than the other but receiving a considerably higher average salary. If there is a high positive correlation between intelligence and salary within each group, the population would be more or less of the kind discussed.

Nevertheless, the criticism is a reasonable one, and implies that in most cases some types of population are *a priori* more likely than others. Even when the data are available only in ranked form, the investigator usually has a fair understanding of the underlying situation. When allowance is made for the extra information, more sensitive criteria can, of course, be obtained. In recent American work on non-parametric tests this aspect of the problem is emphasized, test criteria being chosen which are most powerful with respect to the relevant alternative hypotheses. In a different way, the assumption of an underlying bivariate normal population from which the data are randomly sampled can be introduced to justify considerably reduced standard error formulae for T and R.

It is possible, moreover, to go further in many cases. T and R by no means exhaust the information available in the ranked data, and there is no reason why the assumptions should not be tested on the data themselves. I later discuss in detail one simple situation of this kind.

5. If it is known that correlation exists between two ranked attributes, two extreme types of underlying population of ranks may be distinguished. The actual situation will in most cases be somewhere in between.

(i) The sample is regarded as having been randomly chosen from a bivariate population of ranks. For example, a random sample of individuals may be ranked in order of darkness of hair and darkness of eyes, all individuals being assessed by an infallible judge, so that randomness enters only in the sampling of individuals.

(ii) There is a fixed set of individuals being assessed by a population of judges, or by the same judge in repeated trials, on a particular attribute whose ranking is known *a priori*. The random element is uncertainty of preference, the correlation being the result of real differences between the individuals, and the population is one of rankings conditional on a given objective order.

Intermediate situations arise where, for example, in (i) the judge's assessment is not infallible, or in (ii) the objective order is not accurately known.

Methods of dealing with case (i) have been given by Professor Kendall and myself (1947), without introducing further assumptions about the form of a possible parent population of variables subsequently ranked. We first showed that in such cases the variance of T cannot exceed $\frac{2}{n}(1-\tau^2)$ in samples of n. (Hoeffding (1948) has shown this to be a special case of an extremely general result.) In practice the utility of the formula is doubtful because of the very wide confidence limits it gives for τ, though we show that cases may occur where the upper limit is nearly attained. An idea of the extent of information being thrown away by its use can be got in the following way. In the notation of §2, write $c_{ij} = a_{ij} b_{ij}$ so that c_{ij} has the value $+1$ or -1 according as $p_j - p_i$ and $q_j - q_i$ have the same or opposite sign, and $c_{ii} = 0$. Let n be even, say $2m$, and divide the sample arbitrarily into m pairs 12, 34, 56 . . . Then c_{12}, c_{34}, c_{56}, . . . , $c_{m-1,m}$ are m independent variables taking the values 1 or -1 with probability $\frac{1}{2}(1+\tau)$, $\frac{1}{2}(1-\tau)$ respectively (we suppose there are no ties). Hence $c_{12} + c_{34} + \ldots + c_{m-1,m}$ is a binomial variable ranging from $-\frac{n}{2}$ to $\frac{n}{2}$ in steps of 2, with mean $\frac{n}{2}\tau$ and variance $\frac{n}{2}(1-\tau^2)$, and we have the unbiased estimate

$$T_0 = \frac{2}{n}(c_{12} + c_{34} + \ldots + c_{m-1,\,m})$$

for τ with variance $\frac{2}{n}(1-\tau^2)$. If n is odd, one of the sample members has to be omitted, and T_0 has variance $\frac{2}{n-1}(1-\tau^2)$.

So the use of the maximum variance formula is more or less equivalent to estimating τ from the maximum number $\frac{n}{2}$ of independent comparisons between pairs, which can be chosen arbitrarily in $n!/2^{n/2}$ ways. Indeed, the estimate T_0 has the advantage over T that its distribution is exactly known. Since T is the average of the T_0's obtained by all possible selections of pairs, and since the correlation between any pair of T_0's is $\leqslant 1$, the upper limit to the variance is again established, at least for even n. When n is odd a slightly weaker result is obtained.

The alternative procedure suggested by us is to estimate the variance of T from the sample itself, an unbiased estimate being

$$v = \frac{1}{(n-2)(n-3)}\left\{\frac{4n}{(n-1)}\sum_{i=1}^{n} T_i^2 - 2(2n-3)\,T^2 - 2\right\},$$

where the quantities $T_i = \frac{1}{n}\sum_{j=1}^{n} c_{ij}$ are easily calculated from the row sums of the score matrix c_{ij}. An approximate test for τ is then to treat $(T-\tau)/\sqrt{v}$ as a standardized normal deviate. This is probably the best one can do without further assumptions. The approximation is similar in spirit to that involved in χ^2 tests for contingency tables, and appears to work well in practice.

Unfortunately, the example used by us to illustrate the calculations is not relevant to the theory of the paper. It is in fact a good example of a situation approximating to the second case mentioned. A set of 30 wool samples were measured for mean fibre diameter, and were also ranked visually by three assessors. Apart from the inaccuracy of measurement, which may result in some contiguous samples being interchanged, the state of affairs is precisely that described under (ii).

The essential difference between (i) and (ii) is that in the former correlation is usually the relevant concept, both rankings being as it were on an equal footing, whereas in (ii) the objective order is a predetermined variable and the idea of regression is more appropriate. We are really using T or R to test for significant regression, or to measure regression in some way when it is present, and it is not obvious that either T or R is necessarily the best available statistic for the purpose. H. B. Mann (1945), in an important paper, discusses the properties of T when considered as a test for regression, and shows that while it is an optimum criterion for revealing slight trends, large trends are better detected by an alternative statistic K, which is the smallest number such that all pairs of ranks separated by not less than K are in the correct order. We now examine T and R as measures of trend, making certain assumptions about the population.

6. The process of assessment involved in (ii) is likely in many cases to be that the judge selects individuals in pairs and decides his preferences by rating each individual on an imaginary continuous scale y. The simplest population model is for the frequency function of the ratings $y_1 y_2 \ldots y_n$ to be

$$f(y_1 - m_1) f(y_2 - m_2) \ldots f(y_n - m_n),$$

where $m_1, m_2 \ldots m_n$ are the expectations of his ratings. The model is that envisaged by Babington Smith, as mentioned by Moran in Part II of his paper, and introduced by Thurstone (1926) in connection with the method of paired comparisons. A more elaborate model would allow f to vary in some way with individuals.

It may also happen that the objective order corresponds to increasing values on some scale x. The regression problem is then to find the relation between m and x, the available information being the ranking of the y's, and the values of x, known except perhaps for origin.

The discussion will be confined to the case where the x values are equally spaced if the scale is suitably chosen. Many experiments are designed in this way—for example, the weighing experiment described by Whitfield, though the manner in which it was conducted introduces complications. Some experiments, such as that on wool grading mentioned in §5, approximate to this model sufficiently for a similar method of treatment to be applicable.

As a further simplification, which appears to be valid for both experiments mentioned, it is assumed that the regression of m on x is linear. The discussion can be easily extended if required to curvilinear regression.

The successive x values may be chosen at unit intervals so that $x_j = j$. It is convenient to use the symbol h_{jk} which is 1 if $x_k - x_j$ and $y_k - y_j$ have the same sign and 0 otherwise, with $h_{jj} = \frac{1}{2}$. Ties are excluded.

Then

$$Eh_{jk} = \int_{-\infty}^{\infty} f(y_j - m_j)\,dy_j \int_{y}^{\infty} f(y_k - m_k)\,dy_k$$
$$= \int_{-\infty}^{m_k - m_j} g(w)\,dw,$$

where $g(w)$ is the symmetrical frequency function for the difference of two independent individuals sampled from $f(y)$. Since all statistics giving information on regression must depend only on rank differences, only distributions involving the w's are relevant, and the conditions may be relaxed to allow the assessor to alter the origin of his y scale with each pair compared, as may well happen in practice. The regression is assumed linear so that $m_k - m_j = (k - j)\,\theta$ say. The scale of y is arbitrary and the standard deviation of $g(w)$ may be taken as unity; θ then measures the accuracy of discrimination.

7. Let us first consider θ to be small.

Ignoring θ^3,

$$Eh_{jk} = \tfrac{1}{2} + \theta g(0)(k - j) = \tfrac{1}{2}\beta(k - j),$$

Only $\beta = \theta g(0)$, the regression coefficient of h_{jk} on $k - j$, can be estimated without further assumptions. The variables h_{jk} are not independent, in fact ignoring θ^2 and always taking the suffixes of h to be in increasing order we have

$$\text{var } h_{jk} = \tfrac{1}{4}, \qquad \text{cov } h_{jk}\,h_{jl} = {}^{1}$$
$$\text{cov } h_{jk}\,h_{kl} = -\tfrac{1}{12}, \qquad \text{cov } h_{jk}\,h_{lm} = 0,$$

the second and third results following from the fact that $Eh_{jk}\,h_{jl}$ is the chance that $y_k - y_j \geqslant 0$, $y_l - y_j \geqslant 0$, and similarly for $Eh_{jk}\,h_{kl}$.

The unbiased "least-squares" estimate of β can be obtained by standard methods and justified by its minimum variance property, but it is most easily got by minimizing the variance directly. Let

$$W = \sum_{j<k} \lambda_{jk}\,h_{jk}$$

be any linear function of the h's. We require the λ's which minimize var W subject to $\sum_{j<k} \lambda_{jk}(k - j)$ constant. For all $j < k$,

$$\frac{\partial(\text{var } W)}{\partial\lambda_{jk}} = \tfrac{1}{12}\,\lambda_{jk} + \tfrac{1}{12}\sum_{m=k+1}^{n} (\lambda_{jm} - \lambda_{km}) + \tfrac{1}{12}\sum_{l=1}^{j-1} (\lambda_{lk} - \lambda_{lj})$$
$$+ \tfrac{1}{12}\sum_{m=j+1}^{k} \lambda_{jm} + \tfrac{1}{12}\sum_{l=j}^{k-1} \lambda_{lk}.$$

It is found on substitution that $\lambda_{jk} = k - j$ reduces this to $\frac{1}{12}\,(n + 1)(k - j)$, which is of the right form. Thus

$$W_1 = \sum_{j<k} (k - j)\,h_{jk}$$

can be used to obtain the best unbiased estimate of β, which turns out to be*

$$b_1 = \frac{12W_1}{n^2(n^2 - 1)} - \frac{1}{n}.$$

Now W_1 is in fact a simple function of the Spearman coefficient R. For in the notation of §2, $h_{jk} = \frac{1}{2}(1 + a_{jk}), j < k$; $a_{kj} = -a_{jk}$, and since the x's are in natural order, $b_{jk} = \text{sgn}(k - j)$. Using the formula for R in §2, we find

* Mr. D. V. Lindley has pointed out to me that if $f(y - m)$ is not symmetrical the neglected terms in cov $h_{jk}h_{jl}$ and cov $h_{jk}h_{kl}$ are $0(\theta)$. But with equidistant x's, b_1 must be an odd function of θ, since renumbering the x's in the reverse order is equivalent to writing $-\theta$ for θ. The approximation error in b_1 is therefore $0(\theta^3)$ even when f is not symmetrical.

$$W_1 = \frac{n(n^2-1)}{12}(1+R).$$

So the best unbiased estimate of $\beta = \theta g(0)$ for small θ is simply $b_1 = R/n$. There ought to be an obvious reason for this in terms of regression of the original ranks on the objective ranking, but it escapes me. The trouble is, of course, that the ranks are not independent.*

For values of β whose square is negligible,

$$\operatorname{var} b_1 = \frac{1}{n^2(n-1)}.$$

If the sample is large enough for b_1 to be nearly normally distributed, the minimum variance property implies that in testing $\beta = 0$, b_1 has the usual optimum power properties with respect to alternative β's which are small. Mann's statement that T has similar optimum properties for such alternatives does not really conflict with our result, since T and R are nearly perfectly correlated in samples of moderate size when there is almost complete independence. Writing W_2 for W when all the λ's are unity, $W_2 = \frac{1}{4}n(n-1)(1+T)$ and the corresponding estimate of β is $b_2 = 3T/2(n+1)$ which has variance

$$\operatorname{var} b_2 = \frac{(2n+5)}{2n(n+1)(n^2-1)}.$$

Even in samples of 3, var $b_2 = 0{\cdot}0573$ is only slightly larger than var $b_1 = 0{\cdot}0556$. There appears to be considerable latitude in the choice of λ's for about the same accuracy in estimating β.

8. When θ is not small the problem cannot be handled very satisfactorily without further knowledge of the form of $g(w)$. Moran (Pt. II of his paper) develops the relevant theory of T to some extent, obtaining its mean and variance in a form which can be calculated when $g(w)$ is normal, and Mann's proposal of the alternative statistic K has already been mentioned. Following the present line of argument, we seek the best estimate of β or θ and compare its efficiency of estimation with that of T and R.

The assumption that $g(w)$ has a known form, for example that it is normal, is of such assistance that it is worth while examining the data for evidence on the points. Fortunately this is not a difficult matter. Since $Eh_{j\,j+r} = \int_{-\infty}^{r\theta} g(w)\,dw$ the quantities

$$S_r = \sum_{j=1}^{n-r} h_{j\,j+r}$$

provide estimates of $(n-r)\int_{-\infty}^{r\theta} g(w)\,dw$, and the customary probit device of transforming $z_r = \frac{1}{n-r} S_r$ to a normal deviate ζ_r and plotting ζ_r against r can be used. A straight line through the origin will indicate normality of $g(z)$ if the other assumptions are valid. I have tried out the method on a weighing experiment conducted by Mr. W. A. Donaldson using 24 weights increasing uniformly from 200 gm. to 246 gm., divided into 3 successive groups of 8 and each ranked by 12 judges, the average of the 36 values of S_r being taken (Fig. 1), and the straight line obtained confirms quite well the hypothesis of normality. The three wool assessments mentioned in §5 are also plotted† (Figs. 2A, 2B, 2C) and apart from the slightly anomalous behaviour of assessment A for small values of r, probably due to the errors in measurement and assessment being of comparable magnitude in that region, the agreement with hypothesis is quite good. I propose, therefore, to assume normality when necessary, though a similar method applies for any known $g(w)$.

* I have since realized that when β is small the ordinary regression procedure can be applied to the original ranks provided the variance of the estimate of β is suitably adjusted, since to the order considered all pairs of ranks are equally correlated to an extent $-1/(n-1)$.

† The plotting was discontinued where S_r differed from $n - r$ by one or zero.

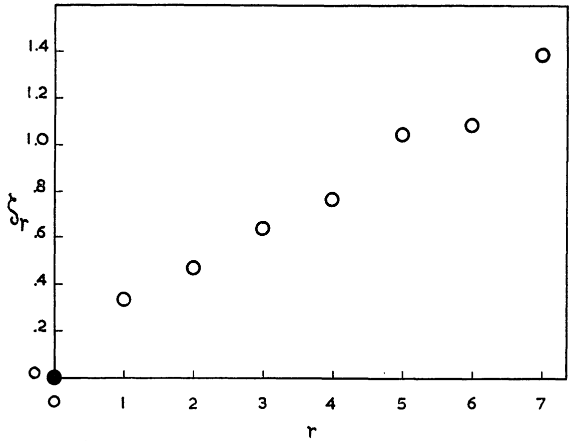

FIG. 1.—Weighing experiment. (W. A. Donaldson.)

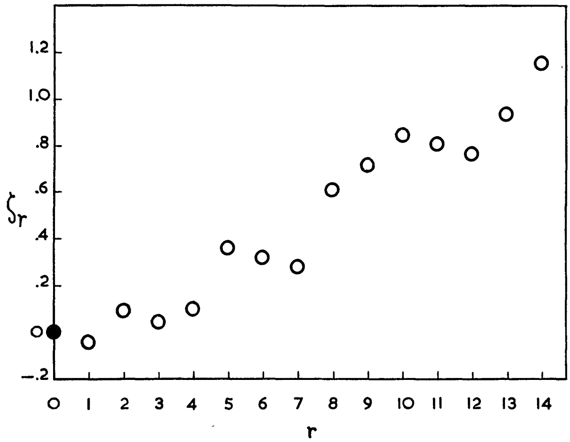

FIG. 2A.—Wool grading experiment.

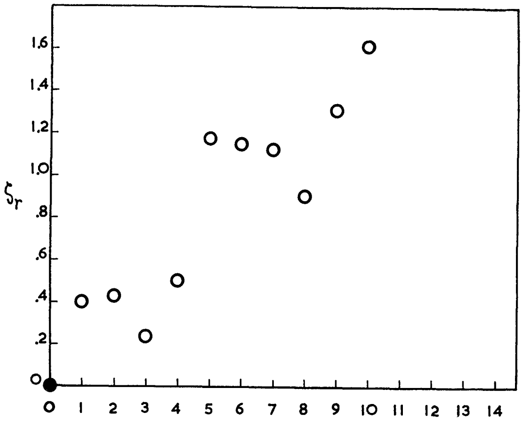

FIG. 2B.—Wool grading experiment.

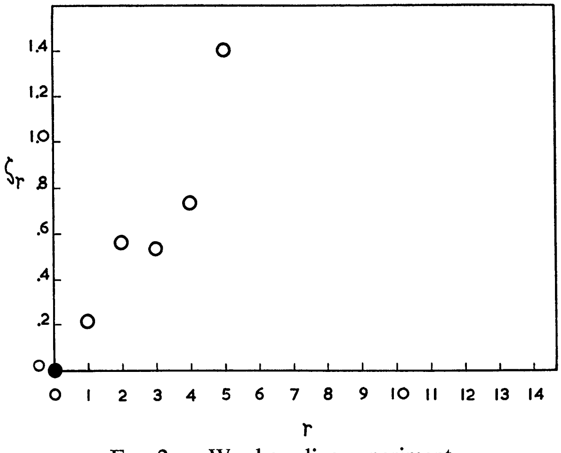

FIG. 2c.—Wool grading experiment.

The standard deviation of $g(w)$ can be taken as unity, and the gradient of the line then estimates θ. A similar scheme was used by Thurstone (1927) for independent paired comparisons; the S_r's are then independent and θ can be estimated by the standard methods of probit analysis. In the present case the S_r's are correlated and the procedure is more complicated.

9. It is first of interest to examine the relation of τ and ρ to θ for the present simple model. The expressions W_1, W_2 of §6 are $W_1 = \sum\limits_{r=1}^{n-1} rS_r$, $W_2 = \sum\limits_{r=1}^{n-1} S_r$, and $E(W_1) = \frac{1}{12}n(n^2 - 1)(1 + \rho)$, $E(W_2) = \frac{1}{4}n(n - 1)(1 + \tau)$. Hence we get

$$\tau = 1 - \frac{4}{n(n-1)} \sum_{r=1}^{n-1} (n - r) \int_{r\theta}^{\infty} g(w)dw$$

$$\rho = 1 - \frac{12}{n(n^2-1)} \sum_{r=1}^{n-1} r(n - r) \int_{r\theta}^{\infty} g(w)dw.$$

Useful approximations can be found if θ is small but $n\theta$ is large enough for the integral to be negligible for $r \geqslant n$. This is true for experiments arranged to give slight discrimination between contiguous individuals but perfect discrimination between extreme ones, as is often the case. The τ formula is then rearranged in the form

$$\tau = 1 - \frac{2}{n(n-1)} \sum_{s=0}^{\infty} s(2n - s - 1) \int_{s\theta}^{(s+1)\theta} g(w)dw,$$

which is approximately

$$\tau = 1 - \frac{2}{n(n-1)} \int_{0}^{\infty} (\alpha + \tfrac{1}{2})(2n - \alpha - \tfrac{1}{2})g(\theta\alpha)\theta d\alpha$$

$$= 1 + \frac{1}{(n-1)}\left\{1 - \frac{1}{4n} - \frac{2g_1}{\theta} + \frac{g_2}{n\theta^2}\right\},$$

where g_m is the m^{th} absolute moment of $g(w)$, and g_2 can be taken as unity. The corresponding approximation for ρ is

$$\rho = 1 + \frac{1}{(n^2-1)}\left\{\frac{3}{4} - \frac{g_1}{2n\theta} - \frac{3g_2}{\theta^2} + \frac{2g_3}{n\theta^3}\right\}.$$

In the normal case,

$$g_1 = \sqrt{\frac{2}{\pi}},\ g_2 = 1,\ g_3 = 2\sqrt{\frac{2}{\pi}}.$$

The approximation ought to work moderately well for $\theta < 0{\cdot}5$, $n\theta > 2{\cdot}5$.

Since τ and ρ depend quite markedly on n as well as θ, caution is necessary in interpreting results such as those obtained by Whitfield for his weighing experiment, where the number of individuals varies from one experiment to another.

10. The estimation of θ from the ζ_r's can be carried out by regression methods if their variance matrix is known. Since $S_r = (n-r)\int_{-\infty}^{\zeta_r} g(w)\,dw$ we have, if $1 - z_r$ is not too small,

$$E(\zeta_r) \sim r\theta,$$

$$\text{cov}\; S_r S_s = (n-r)(n-s)g(r\theta)g(s\theta)\;\text{cov}\;\zeta_r\zeta_s.$$

Taking $g(w)$ to be normal it is found that

$$\text{cov}\; S_r S_s = 2(n-s)\{F(r\theta, s\theta;\; \tfrac{1}{2}) - F(r\theta)\,F(s\theta)\}$$
$$+\, 2[n-r-s]\{F(r\theta, s\theta;\; -\tfrac{1}{2}) - F(r\theta)\,F(s\theta)\}$$

if $r < s$, and

$$\text{var}\; S_r = (n-r)\,F(r\theta)\{1 - F(r\theta)\} + 2[n-r]\{F(r\theta, r\theta;\; \tfrac{1}{2}) - F^2(r\theta)\},$$

where $[x] = x$ or 0 according as x is positive or negative, $F(x)$ is the chance that $X \geqslant x$, and $F(x, y;\; \pm\tfrac{1}{2})$ is the chance that $X \geqslant x$, $Y \geqslant y$, X and Y being bivariate normal with correlation $\pm\tfrac{1}{2}$ and unit variance. Denote the variance matrix of the S_r's by V_{rs} and its inverse by V^{rs}.

Ignoring difficulties near $r = n - 1$, when z_r is nearly unity and $n - r$ is small, the regression estimate of θ is

$$D = \Sigma\, \mu_s \zeta_s,$$

where

$$\mu_s = \frac{\sum_r r(n-r)(n-s)g(r\theta)g(s\theta)\,V^{rs}}{\sum_{r,s} rs(n-r)(n-s)g(r\theta)g(s\theta)\,V^{rs}}$$

and

$$\text{var}\; D = \{\sum_{r,s} rs(n-r)(n-s)g(r\theta)g(s\theta)\,V^{rs}\}^{-1}.$$

The parameter θ in μ_s would be sufficiently accurately estimated from a line fitted by eye.

In practice trouble will certainly arise from the behaviour of z_r when $n - r$ is small, particularly if a single ranking is used, and the following equally efficient method of estimation is to be preferred. An estimate D_0 is obtained from the value W_0 taken by $W = \Sigma\lambda_r S_r$ when the λ's are chosen to minimize var $W\Big/\left(\dfrac{\partial E(W)}{\partial\theta}\right)^2$. The appropriate coefficients are

$$\lambda_r = \sum_s s(n-s)g(s\theta)\,V^{rs},$$

so that D_0 is obtained from

$$W_0 = \sum_{r,s} s(n-s)(n-r)g(s\theta)\,V^{rs}.\int_{-\infty}^{rD_0} g(w)dw$$

and var D_0 = var D. As before, θ is taken from an approximate probit line.

Estimates D_ϱ, D_τ are also available from R and T, using formulae analogous to those relating τ, ρ and θ, and their relative efficiency can be examined. The variance of D_ϱ is var $R\Big/\left(\dfrac{\partial\rho}{\partial\theta}\right)^2$ and similarly for D_τ. Since var $W_1 = \Sigma\, rs\, V_{rs}$ and var $W_2 = \Sigma V_{rs}$ we find

$$\text{var}\; D_\varrho = \frac{\Sigma\, rs\, V_{rs}}{\{\Sigma\, r^2(n-r)g(r\theta)\}^2}$$

$$\text{var}\; D_\tau = \frac{\Sigma\, V_{rs}}{\{\Sigma\, r(n-r)g(r\theta)\}^2}.$$

As might be expected from §7, D, D_0 and D_ϱ become identical when θ is small if θ^2 is ignored,* and all the estimates have practically the same efficiency. I have calculated the variances for the case $\theta = 0\cdot 3$, $n = 8$, for which the approximate formula gives $\tau = 0\cdot 58$, and find that

$$\text{var } D_0 = 0\cdot 02281, \quad \text{var } D_\varrho = 0\cdot 02294, \quad \text{var } D_\tau = 0\cdot 02394.$$

The three estimates are still remarkably alike in accuracy, and little is lost by using T or R. It was not possible with the available tables of $F(x, y;\ \pm \frac{1}{2})$ to examine higher values of θ, but the result suggests T or R may be highly efficient for a wide range of θ. As pointed out in §8, it is still necessary to transform to D_τ or D_ϱ if discrimination is being compared from rankings of different size. I am having tables prepared for this purpose which will be of use over the whole range of 7 (see Appendix). For small samples there is, of course, a bias in $D\tau$ and $D\varrho$ which remains to be investigated.

So far no mention has been made of the direct maximum likelihood approach, which is a comparatively difficult one. If $p_1, p_2, \ldots p_n$ are the ranks of the known x's when the assessed y's are taken in order $123 \ldots n$, the likelihood of the ranking is

$$P = \int_{u_1 \leqslant u_2 \leqslant \ldots \leqslant u_n} \ldots \int f(u_1 - p_1\theta) f(u_2 - p_2\theta) \ldots f(u_n - p_n\theta) du_1 \ldots du_n,$$

where f has an arbitrary mean. When θ is small and the integrand is expanded to θ^2, $\dfrac{\partial P}{\partial \theta} =$ leads to the statistic R as before. The general case is very complicated even if f is taken as normal (with variance $\frac{1}{2}$), but I doubt whether the resulting estimate will be much more accurate than D. In view of Mann's result, however, it would be of interest to see whether his K leads to a more accurate estimate of θ than D, in which event I may be wrong about maximum likelihood.

11. A numerical check on the assumptions of normality and linearity of regression is obtained by comparing the residual quadratic form for the ζ_r's,

$$Q = \sum_{r,s} (n - r)(n - s) g(r\theta) g(s\theta)\, V^{rs}\, \zeta_r \zeta_s - \frac{D^2}{\text{var } D}$$

with its expectation $n - 2$. If the ζ_r's near $r = n - 1$ are too erratic they may be omitted and the calculation of D and Q carried out on the remainder, using a suitably truncated variance matrix and correspondingly reduced expectation for Q. But I do not think a χ^2 test for Q with $n - 2$ degrees of freedom is legitimate if only a single ranking is used. The ζ_r's are individually nearly normally distributed for moderate n in the sense that each ζ_r distribution has the form

$$C \exp \left\{ \frac{-(\zeta_r - r\theta)^2}{2 \text{ var } \zeta_r} \right\} . [1 + O(n^{-\frac{1}{2}})].$$

For Q to be distributed as χ^2, all the ζ_r's are required to be *jointly* normally distributed, which certainly does not follow.

A more justifiable procedure is to isolate additional single degrees of freedom, such as those representing the quadratic, cubic, . . . components of the regression, and testing the corresponding squares as χ^2. Significant non-linear regression might be interpreted in a variety of ways, which I shall not discuss here.

12. The model considered is the simplest of its kind, and the next step is to discover how to cope with cases where there is uncertainty in both rankings. They may be considered as assessments of an unknown true rank order with discriminations θ and φ respectively. When both θ and φ are of comparable magnitude I believe it is possible to estimate only $\theta + \varphi$. On the other hand, when φ is known to be large compared with θ, as for example in Fig. 2 (A), it is undoubtedly possible to estimate at least θ. A crude method is to fit a probit line to the points for which the objective ranking is certain. The optimum estimation of θ and φ in the general case remains an intriguing problem.

* A direct proof is quite difficult.

Acknowledgments

I am grateful to Mr. P. A. P. Moran, Mr. J. W. Whitfield and Mr. B. Babington-Smith for their useful comments on the paper.

References

Daniels, H. E. (1948), "A property of rank correlations," *Biometrika*, **35**, 416.

—— and Kendall, M. G. (1947), "The significance of rank correlation when parental correlation exists," *ibid.*, **34**, 197.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distributions," *Ann. Math. Statist.*, **19**, 293.

Kendall, M. G. (1948), *Rank Correlation Methods.* London: Griffin & Co.

Mann, H. B. (1945), "Nonparametric tests against trend," *Econometrica*, **13**, 245.

Thurstone, L. L. (1927), "The method of paired comparisons for social values," *J. Abn. and Soc. Psychol.*, **21**, 384.

Appendix

I am indebted to Mr. D. A. East for calculating the following table giving θ in terms of τ and n according to the exact formula of §9. It can be used to convert T into an estimate D_τ of θ.

Table of θ

	τ								
n	·1	·2	·3	·4	·5	·6	·7	·8	·9
2	·126	·253	·385	·524	·674	·842	1·036	1·282	1·645
3	·094	·191	·292	·401	·522	·663	·837	1·072	1·446
4	·076	·153	·235	·324	·426	·547	·703	·925	1·300
5	·063	·128	·197	·273	·360	·467	·607	·815	1·184
6	·054	·110	·169	·235	·312	·407	·535	·730	1·090
7	·047	·096	·148	·207	·275	·361	·478	·661	1·010
8	·042	·085	·132	·184	·246	·324	·432	·604	·943
9	·038	·077	·119	·166	·223	·294	·395	·557	·884
10	·034	·070	·108	·152	·203	·269	·363	·517	·833
11	·032	·064	·099	·139	·187	·248	·336	·482	·788
12	·029	·059	·092	·129	·173	·231	·313	·451	·747
13	·027	·055	·085	·120	·161	·215	·293	·425	·711
14	·025	·051	·080	·112	·151	·202	·275	·401	·678
15	·024	·048	·075	·105	·142	·190	·260	·380	·649
16	·022	·045	·070	·099	·134	·180	·246	·361	·621
17	·021	·043	·067	·094	·126	·169	·233	·344	·596
18	·020	·041	·063	·089	·120	·161	·222	·328	·573
19	·019	·039	·060	·084	·114	·153	·212	·314	·552
20	·018	·037	·057	·080	·109	·146	·202	·301	·532
21	·017	·035	·054	·077	·104	·140	·194	·289	·514
22	·016	·034	·052	·073	·099	·134	·186	·278	·497
23	·016	·032	·050	·070	·095	·129	·178	·267	·481
24	·015	·031	·048	·068	·092	·124	·172	·258	·466
25	·015	·030	·046	·065	·088	·119	·166	·249	·452

Discussion on Symposium on Ranking Methods

Mr. R. L. Plackett: Our gratitude is due to the authors for their interesting survey. It is also, I think, due to Professor Kendall for writing a valuable monograph on the subject, and then encouraging the speakers to demonstrate that it is now incomplete. The papers presented to-night are widely different in some respects, being the work of three people in a field of statistics where much remains to be done. Criticism and comment, therefore, appear to be forced into a rather grasshopper approach with considerable use of the speakers' names. Let us examine what matters there are in common.

Firstly, all three papers use a different notation for the rank correlation coefficients of Spearman and Kendall. Dr. Daniels notes that Mr. Whitfield and Mr. Moran use different notations and he therefore sees no harm in introducing two additional symbols, one of which is associated with the name of Hotelling while the other is standard for the multiple correlation coefficient. This question of notation is undoubtedly a nuisance, but I feel there should be some consideration for research workers in subjects requiring a statistical approach who tend to associate a particular symbol with a particular procedure and may be thrown into great confusion by this evening's proliferation. As in industry, if you want to sell, you must standardize. I suggest that r_S and r_K, with population values ρ_S and ρ_K, are unambiguous if it is made clear that the suffices are capital letters and there is no connection with serial correlation. By an extension of Dr. Daniels's argument I see no harm in using this fourth notation myself.

Secondly, all three papers discuss that approach to ranking which assumes an underlying variate. Here we find a certain dissimilarity of opinion. Dr. Daniels devotes the last two-thirds of his paper to discussing the relation between a fixed variate and a ranking based on a rating scale. Mr. Moran employs the same theoretical basis in his Part II. On the other hand, Mr. Whitfield in section 2 brings forward arguments against the rating scale to support the use of Kendall's coefficient r_K. It is true that Dr. Daniels is concerned with a rating scale in the judge's head and Mr. Whitfield with one written down on paper, but there seems no logical distinction between the two, and I think that Whitfield's arguments will often tend to invalidate Daniels's assumptions, though not always, for there are certainly examples where the assumptions are reasonable. Students in their final year at a University will have their papers marked in the usual way, but the only results which appear are that a particular student is in Class 1, 2, 3 or 4. Here a more or less continuous variate is transformed into a ranking. One may then wish to examine the relation between the known H.S.C. grades and the final ranking, which is a rather awkward case of Daniels's problem—awkward because of the numerous ties. In general, I think the imaginary rating scale is too imaginary—why not a rating area, or a rating volume, or a subjective weighting of several scales? these seem more appropriate for judging objects to which we react in a complicated way. What happens when people rank—do they consciously perform a series of paired comparisons, or ranking of sub-groups? Statistics, in fact, are not enough. We want a psychological study of human behaviour to determine how the numerical information reached us, but not one in which the human observer is considered as a crude measuring instrument. Similar problems arise in the study of the action of poisons on insects—it is possible to invent mathematical models which seem reasonable to the mathematician, because tractable, but what is really required is a physiological study of insect behaviour.

Turning to the papers separately, I congratulate Dr. Daniels on his ingenious proof that $-1 \leqslant 3\rho_K - 2\rho_S \leqslant 1$. He clearly shows that ρ_K and ρ_S describe different aspects of the population, but can he be more precise about what these aspects are? In other words, do circumstances exist in which for some reason r_K is preferable to r_S, or another of his Γ series of coefficients is preferable to either—are a_{ij} and b_{ij} related to anything in the information required? Mr. Moran says these circumstances do exist, so perhaps he can supply some examples. An evident development of Dr. Daniels's work on the estimation of the regression coefficient θ would be to determine which Γ coefficient gives the most efficient estimate, although this seems a difficult problem, the solution of which might not result in much increase in efficiency. Another development is that required in probit analysis when the individual normal equivalent deviations ζ_r are correlated. How far, if at all, this correlation accounts for the linearity in Fig. 1 and the optimistic assertion of it in Fig. 2 would provide an interesting investigation, but I certainly think that linearity here should be cautiously interpreted—which is a way of saying we cannot quite decide what the interpretation is.

I have no comments on Mr. Moran's paper beyond those already mentioned, and pass now to Mr. Whitfield's survey. The weight experiment (section 5) designed to discover a relation between the quality of a ranking and the size of the sub-sample on which it is based, appears to have suffered somewhat from the amount of ranking required from each subject introducing an interaction between subject and order injurious to the latin square layout.

A similar effect will still tend to occur even when the conditions are presented on different occasions, and the variation in subject performance seems to present a tricky problem. Perhaps Mr. Whitfield can pick solid untemperamental types.

I think also that complications arise in this analysis because of the different variances of the mean coefficients r_K, unknown but bounded by ρ_K. A separate estimate of variance, calculated as Dr. Daniels indicates on p. 174, might have been useful. As regards Mr. Whitfield's problems in sections 6 and 7, I would further advocate choosing suitable a_{ij} and b_{ij}, at least until Dr. Daniels develops his estimation of the discrimination coefficients θ and φ. Finally, on the porridge-bowl problem (section 9) I was not clear how exactly to apply the test of significance, which, I presume, was the purpose of determining the distribution of the number of interchanges. In the first place one would require to find the minimum number of interchanges of pairs of adjacent items needed to transform the order actually obtained into a "good order"; experience of a few permutations suggests that unless one has a complete list, this may not always be accomplished with ease or even correctness.

Many diverse and intriguing problems have been revealed in this symposium. I would like to propose a vote of thanks to Dr. Daniels for his original and important contribution to the theory, to Mr. Moran for the careful way in which he has given the recent history of the subject, and to Mr. Whitfield for his useful survey of the practical problems which face the experimenter in this field.

Mr. B. Babington Smith: I have great pleasure in seconding a vote of thanks to these three symposiasts for their interesting and stimulating papers. Mr. Moran and I are occasionally able to discuss such problems as these, and I am glad to find that the defeatist note which creeps into his conversation has not permeated his contribution to-night. Of one of his unsolved problems, partial τ, I propose to say something later. Mr. Whitfield hunts in much the same country as I do, and there seems no doubt that we serve a useful purpose in finding and formulating fresh difficulties for the theorist. I congratulate Mr. Whitfield on the crop he has produced to-night. Of Dr. Daniels my feeling is that after i have been trying to see through brick walls, I find Dr. Daniels is looking over the top.

Of the problem of ranking methods in general, I would urge that more attention might be paid to paired comparisons. Observation of experiments has taught me that in a wide range of situations a subject's procedure, when called upon to produce a rank order, varies roughly according to the difficulty of the task. The greater the differences between the items, the more the procedure is one of assessing the absolute values of the items; these being settled, the ordering of the items follows. The smaller the differences the more the procedure is one of comparison between pairs, and there seems to be a case, therefore, for making the statement that a subject never adopts a method of ranking; the fundamental processes are either direct assessment or direct comparison. In either case the rank order is an end result, achieved, without confusion, from the results of paired comparison, where the differences are large, and when they are small by imposing the condition that there shall be no circular triads,

It was partly to draw attention to this distinction that I adopted the Thurstonian scheme in ranking and paired comparison and made the suggestion I did to Mr. Moran. One supposes an underlying linear scale on which the items lie distributed in some way. A normal distribution is obviously the most convenient assumption. In ranking, each item is given a value and the ranking then follows from these values. In paired comparison the item may have a different value (drawn, of course, from the hypothetical distribution) each time it appears in a comparison. It is this possibility of taking a different value which allows circular triads to arise in a linear model and may answer Mr. Whitfield's point.

So far as I know, this scheme has only been considered for linear scales, but there is no difficulty in conceiving either scheme in any number of dimensions. In practice it is easy enough to produce a situation which probably involves several dimensions, and in this way it is easy to envisage the circular triads. Another possibility which model makers may have to consider is that in paired comparison the unit may be "the pair" and not the "item" alone.

I have another reason for preferring "paired comparison" as a model. In psychological experiment, if one is concerned to order a set of items or people with respect to an attribute or attributes, one is concerned with assessing the differences between them. It is only in a limited class of case that one is concerned with either a bivariate population in which the value of ρ is $\rho' \neq 0$, or of which we may say that items spaced equally 3 units apart will not be reversed in order. Models such as considered by Kendall and Daniels or Mann give an answer which is so intangible. What I want to know is more often "is coffee preferred to tea ?", or "is gymnastics preferred to commercial subjects ?", and what is the chance of reversal? It may help me to visualize relations in a group of items if I convert such proportions into distances on a linear scale. Again, if I were choosing a candidate for a job the answer required would be of that type, and I doubt if an answer $\theta = \frac{1}{2}$ would satisfy me: perhaps Dr. Daniels can reassure me on this point. The probit method seems to have distinct possibilities; once again it is free of tiresome constraints if applied to paired comparisons.

There is a point about the variance of T. It has been shown that the variance of T in ranking on the null hypothesis is given by

$$\text{var } T_{(r)} = \frac{2(2n+5)}{9n(n-1)} \rightarrow \frac{4}{9n} \text{ when } n \text{ is large.}$$

(May I say in passing that I regard ranking methods as becoming altogether unwieldy as n becomes large.)

For paired comparisons, on the null hypothesis, the variance of $T_{(pc)}$ is $1/4\binom{n}{2}$. In other words

$$\text{var } T_{(r)} = \frac{4}{9}(2n+5) \text{ var } T_{(pc)}$$

$$\sim \frac{8}{9}n \text{ var } T_{(pc)} \text{ for large } n.$$

It is clear from this that the variance of $T_{(pc)}$ is of order n^{-1} times the variance of $T_{(r)}$, an advantage which is, of course, gained by making $\binom{n}{2}$ judgments instead of n.

Looking at it another way, the variance of mean T from n pairs of rankings of n items would be 8/9 the variance of T from two sets of paired comparisons, when n is large. It is not clear, however, on the face of it which method will be most economical when the time taken to make judgments is taken into account: especially when we note that the numerical factor $4(2n + 5)/9n > 1$ when $n < 20$.

From observations, I am quite doubtful whether ranking actually has the advantage. I recently watched a subject produce a rank order from nine items. In order to do so he made more than 70 comparisons between pairs of items, which is roughly the equivalent of two sets of paired comparisons. Of course, the 70 judgments were not evenly distributed because he paid much more attention to the more difficult decisions.

I should like to make some comment on the experiment on weights which Mr. Whitfield mentions. If he had carried out the experiment by means of paired comparisons on the same material, and if (*pace* the Weber Law) we regard his weights as equally spaced at intervals corresponding to $\cdot 5\sigma$ in a Thurstonian scale, I find that the expected values of T between pairs of judges would be:

Size of Group	*Expected value of* T
3	·22
4	·32
6	·49
8	·60
12	·72
24	·85

This is interesting in connection with the point raised by Dr. Daniels about the change in the size of T with n. This approach is not the same as Mr. Whitfield's but the difference in trend is more obvious, and I am driven to ask myself whether more than the difference between ranks and paired comparisons is not required to account for that. In fact, his experimental procedure is very harsh. By making all the items alike, I should say he has made his subjects' task almost impossible, and I think this is shown by the falling off in correlation with the size of group. I find it difficult to reconcile his description of the set of weights with any sort of interviewing except, perhaps, in some brave new world, interviewing a set of Bokanovsky twins in the dark! In interviewing candidates one attaches one's impressions to names, colour of hair, clothes or gait, although such attributes may be strictly irrelevant and indeed only serve as aids to memory.

I can well imagine that his subjects preferred to work with smaller groups: but even so Mr. Whitfield has not explained how the results from smaller groups were run together to give a final order. Failing this step it seems to me that the results with the smaller groups are not directly comparable with those from the larger as estimates of T, and the inference that there is no loss of information till the very small groups are reached is not warranted.*

E. G. Chambers raised a question about paired comparisons some years ago which turned on the great effort involved in making paired comparisons. I toyed then with the idea of breaking the main block down into sets of overlapping groups, but there were difficulties about pooling them. Mr. Moran suggested one might work with balanced incomplete blocks, but this does not involve any saving in the number of judgments made, although there may be an appreciable saving in the amount of effort expended at one time.

May I make a note in passing about the coefficient of inconsistency? I suggest there would be certain advantages in re-defining the coefficient of inconsistency (and with Mr. Plackett's permission I will suggest a new symbol): $\Xi = 1 - \left\{4d \middle/ \binom{n}{2}\right\}$ which gives it the same form, but a different minimum value, for n odd and even. The advantage is that it is then in line with other coefficients; that when the chance expectation of circular triads is realized, the value of the coefficient is zero. There is a certain rough justice in attaching a negative sign to the situation where the number of circular triads exceeds chance expectation, and I feel inclined to say that when this reaches a maximum we have reached a state of wilful inconsistency.

I have one more point: Mr. Moran on p. 157 says that partial τ for more than three variables has not yet been discussed, nor has there been any definition of multiple rank correlation. Kendall's partial τ has always struck me as being a sort of cuckoo in the nest, that is different in kind from total τ which is always derived from an additive process, since it can be derived from procedure of product moments. Even Kendall speculates that the resemblance is a coincidence. If we consider the situation in which we wished to partial out two or more variables, it is clear that we cannot adopt Kendall's procedure directly, because he arranges the independent variable in the natural order. We may, however, produce an intelligible and reasonable question and some sort of answer if we regard the matter thus:

Consider four judges, A, B, C and D. We may reasonably ask what is the relationship between A and B in different situations, namely: (i) when C and D agree; (ii) when C and D disagree. Similarly, for m judges we may consider what happens to the relationship between two of them when the remaining $m - 2$ all agree, all but one agree, and so on. Two points may be noted here: (i) that a situation is freed from internal constraints if we consider paired comparisons rather than ranking, and (ii) there is a difference between the case where we consider what they agree about and the case where we only consider whether they agree—a distinction which is not overt in the case of the three variables.

* As Mr. Whitfield has explained in a footnote, I misunderstood the procedure which he adopted. I supposed, when he spoke of breaking the set of 24 up into, say, 6 sets of 4 each, that the four in each set were weights adjacent to one another in the scale. I have therefore tried to estimate afresh the expected values of T for a model rather closer to his experiment.

If we consider all possible samples of m items from a group of n, which are distributed along a hypothetical linear scale at equal intervals of $a\sigma$, the mean separation between items in the sample will be $(n + 1)a\sigma/(m + 1)$.

Taking $a = 0{\cdot}2$ to give a result of the right order of size, the values for samples of m scattered at uniform intervals of $(n + 1)(0{\cdot}2\sigma)/(m + 1)$ would be for groups of various sizes:

$n = 24$	$m = 24$	·63
	12	·64
	8	·65
	6	·67
	4	·70
	3	·74
	2	·82

The trend here is *prima facie* quite close to Mr. Whitfield's. More exact values could be obtained given the variate values of any sample.

I still feel that Mr. Whitfield's model only goes half the way; because, if we are seeking the most economical or most efficient way of assessing a group of n candidates by interview, we must not forget the following point: If we separate the group into two or more blocks it may be easier to make judgments among members of a block, but we may find we lose what we have gained when we come to put the sets together again to give a single final order.

I will give as an example of the sort of information which could be derived by this approach the table of preference from Kendall's book—*Rank Correlation Methods*—p. 127, where the preferences of 21 schoolboys for 13 school subjects are set out. My wife and I went through the list of school subjects expressing our preferences in paired comparisons. The results are set out in three tables below. The results for comparisons where the partition of preference among the schoolboys was 17 to 4 (or more extreme) are shown in Table 1. Table 2 includes comparisons where the schoolboys divided 13:8, 14:7, 15:6 or 16:5, while in Table 3 are those comparisons which gave partitions 12:9 or less extreme.

Tables Showing the Distribution of Agreements between Two Judges A and B

Table 1.—17 *cases where* 17 *or more boys agreed in one direction.*

	B's judgments agree with majority	*B's judgments disagree*
A's judgments agree with majority	4	6
A's judgments disagree	3	4

Table 2.—47 *cases where* 13, 14, 15, 16 *boys agreed in one direction.*

	B agrees	*B disagrees*
A agrees	18	7
A disagrees	7	15

Table 3.—14 *cases where* 12 *or less boys agreed in one direction.*

	B agrees	*B disagrees*
A agrees	6	5
B disagrees	0	3

One can see how agreement between the two judges A and B changes with the level of agreement among the main group of 21. In this case the agreement (at any rate as measured by an assoriation coefficient) runs counter to the agreement between the boys.

This is a question which may well be of interest and has clear analogies to partial association.

I have great pleasure in seconding the vote of thanks.

Mr. A. Stuart: From Mr. Moran's remarks on sampling from a ranked population and Dr. Daniels's discussion, in section 5 of his paper, of alternative population models, it is clear that the work already done on sampling in the non-null case has been confined to Dr. Daniels's type (i), i.e. sampling from a population of ranks. As an illustration of Dr. Daniels's type (ii) model, sampling from a population of rankers, it may be of interest to the meeting to hear of some work recently done in the m ranking case.

An enquiry carried out by the Social Research Division of the London School of Economics, which has been fully reported in the *British Journal of Sociology* (Vol. I, No. 1, March, 1950), suggests the following problem: If a random sample of m rankings is drawn from a population of M rankings (ties not being permitted) and the mean rank allotted to each object is calculated, which of the differences of mean ranks in the sample will be significant of differences in the population?

Let the frequency of allotment of the r^{th} rank to the i^{th} object be α_{ir} in the population, and a_{ir} in the sample. Then it follows at once, by an application of Vandermonde's Theorem, that

$$E(\bar{a}_i) = \alpha_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where $\bar{\alpha}_i$, $\bar{a}_i$ are respectively the population and sample mean rankings for the i^{th} object.

Similarly, by repeated application of the same theorem,

$$\text{var } (\bar{a}_i) = \frac{M-m}{M-1}\frac{\text{var } \alpha_i}{m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where var α_i is the variance of the rankings of the i^{th} object in the population.

These are the usual results for sampling from a finite population, as one would expect. When one considers the covariance of the means of two objects, the situation becomes more complicated, for it is necessary then to take account not only of the α's of the population, but also of the number of rankings in the population which place *both* the i^{th} object in the r^{th} rank and the j^{th} object in the s^{th} rank. If that number is $\alpha_{ir.js}$, after some lengthy algebra we find:

$$\text{cov}\,(\bar{a}_i, \bar{a}_j) = \frac{M-m}{M-1}\frac{1}{m}\left\{\frac{1}{M}\sum_r \Sigma\, rs\alpha_{ir.js} - \bar{\alpha}_i\,\bar{\alpha}_j\right\} \quad . \quad . \quad . \quad (3)$$

The expression in brackets is, of course, the covariance of the rankings allotted to the two objects in the population which may be called cov $\alpha_{i.j}$. Cov $a_{i.j}$ is the corresponding value in the sample.

Putting these results in the formula

$$\text{var}\,(x - y) = \text{var}\,x + \text{var}\,y - 2\,\text{cov}\,(x, y)$$

we find

$$\text{var}\,(\bar{a}_i - \bar{a}_j) = \frac{M-m}{M-1}\frac{1}{m}\{\text{var}\,\alpha_i + \text{var}\,\alpha_j - 2\,\text{cov}\,\alpha_{i.j}\} \quad . \quad . \quad . \quad (4)$$

Substituting the unbiased estimators

$$\frac{m}{m-1}\,\frac{M-1}{M}\,\text{var}\,a_i \text{ for var } \alpha_i$$

and

$$\frac{m\,M-1}{m-1\,M}\,\text{cov}\,a_{i.j} \text{ for cov } \alpha_{i.j}$$

we obtain the formula in terms of sample statistics

$$\text{var}\,(\bar{a}_i - \bar{a}_j) = \frac{M-m}{M}\,\frac{1}{m-1}\,(\text{var}\,a_i + \text{var}\,a_j - 2\,\text{cov}\,a_{ij}) \quad . \quad . \quad 5$$

In virtue of the tendency of the distribution to normality, this expression provides a standard error for tests of significance between the sample mean rankings of any two objects. If, further, M is large compared to m, and m itself is large, it reduces to

$$\frac{1}{m-1}\,(\text{var}\,a_i + \text{var}\,a_j - 2\,\text{cov}\,a_{i.j}) \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Where only two objects are being ranked, i.e. $n = 2$, this reduces, as is to be expected, to the binomial test for $p = \frac{1}{2}$.

It must be remembered, however, that the test only gives the probability, for each pair of objects, that their mean ranks are in the same order in the sample as in the population. But there are $\binom{n}{2}$ pairs of objects, and as the tests are not independent, we cannot compound the individual probabilities directly.

Further, it does not follow from the significance of the difference of a given object mean $\bar{a}_i$ from another that a third mean, further away from $\bar{a}_i$ than the second was, would also be significantly different: in fact, the variance and covariance terms in (6) would be different for each pair of objects.

These are severe restrictions on the usefulness of the test, and further work is being done in the hope of improving upon it. But even as it stands, it might perhaps be of some value for particular purposes.

Mr. J. I. Mason: I will try to outline the problem facing the practical business statistician. As members of the symposium have stressed, the assumptions necessary at present to provide a mathematical model for ranking assessments, although fascinating and elegant, are not yet adequate for much practical business work. Even where they are adequate they are by no means sensitive enough. They are inconclusive, and do not, in themselves, help the practical man to assert anything positive; he has still to rely on his feeling of the background, and will always have to do so.

Let us consider the following simple case:

Expected	*Actual*
1	2
2	1
3	3
4	4
5	5
6	6

Here the difference shown may be very real, yet the null hypothesis tests get nowhere. If, further, the hypothesis of equiprobable permutations is unlikely or if, say, the last four ranks are correlated, nothing can be done by pure statistics as yet. It may even happen in practice, paradoxically, that the unchanged ranking of the last four heightens the probability that inversion of the first pair is highly significant. Just as with χ^2, too high a "P-value" may be a danger signal.

In order to deal with permutations having different probabilities the line of attack is, perhaps, to try to rank the rankings. But this involved the whole unsolved problem of weights. If we take the same example, as before and try to weight by the squares of the inverse order of ranking, we find:

Wt.	E	A_1	$A_2 \ldots A_n$
36	1	2	6
25	2	1	5
16	3	3	4
9	4	4	3
4	5	5	2
1	6	6	1
Total	196	207	441
Reduced rank	0	0·05	1

I do not see why, in logic, such a scale is not just as valid a criterion as any of the existing simple measures, and equally dangerous to use.

May I, therefore, express appreciation of the pioneer work of the members of the symposium in putting forward many of the practical snags and limitations in the first stages of the new attack by scores, ranks and weights. Whether they can replace the, older, statisticians' experience and subjective assessment of the relationship in the figures against the human background from which they were drawn remains to be seen.

Methods suitable for ranking colour, taste or smell preferences will be most valuable to business men.

Dr. I. J. Good: I think it is possible to generalize the usual theory of paired comparisons to make it agree more closely with the manner in which people think. Given a pair of objects (A, B), it has been customary to ask for a statement of preference or of no preference. It may be better to ask for an approximate judgment of the probability, p, that A is preferable to B. If no great accuracy is required it might be possible to make such judgments at considerable speed. In order to measure the degree of agreement between the judgments of two people, the expression $(2p - 1)(2q - 1)$ can be summed over all pairs of objects, where p and q are probabilities given by the two people. (In the usual theory p and q are always equal to 0, $\frac{1}{2}$ or 1.) For greater generality different weights can be attached to the different pairs of objects. I do not know how to cope with the question of distribution. (*Note added on April 18th:* The "number of contradictions" in one person's judgments could be measured by summing $pp'p'' + (1 - p)(1 - p')(1 - p'')$ over all triads, where the notation is self-explanatory.)

Mr. S. T. David: In an attempt to find the solution of Mr. Whitfield's "Goldilocks" problem, I have had occasion to look for an explicit expression for the distribution of t in the null case. I have been unable to find anything in the literature, other than the integral given by Mr. H. G. Haden. I have, however, found out quite a simple expression which the audience may be interested to see. If instead of counting $+1$ when the numbers are in the right order and -1 when they are in the wrong order, we count $+1$ when they are in the right order and 0 when they are in the wrong order, the generating function for the frequencies of the sum of such units, say z, is:

$$(x^{n-1} + x^{n-2} + \ldots + x + 1)(x^{n-2} + \ldots + 1) \ldots (x + 1)$$

where the coefficient of x^z in the product is the frequency of z. If this expression is factorized in terms of the imaginary roots of unity, the coefficient of x^z (since the coefficients of the polynomial are symmetrical about $x^{n(n-1)/2}$) is $(-)^z\,(1^z)$, where the quantities entering into the symmetric function (1^z) are all the imaginary and negative roots of unity of all orders up to and including n.

Now $z = \frac{1}{2}\binom{n}{2}(1 + t)$.

So that the frequency of t is given by

$$(-1)^{\frac{1}{2}\binom{n}{2}(1+t)}\,(1^{\frac{1}{2}\binom{n}{2}}\,{}^{(1+t)}).$$

This function is fairly easy to calculate if expanded in terms of the monomial power sums, but I do not know if it can be handled analytically.

The CHAIRMAN: If the audience can bear with me at a somewhat late stage there is one remark I wish particularly to make, namely to refer to what I regard as the major outstanding problem in ranking theory at the present time. This is simply to find some method of specifying a population of ranks in the non-null case. If we have a ranking of n which can happen in $n!$ ways, then in order to specify such a population in general we require $n!$ parameters. This is far too large a number for any tractable mathematics to be applied to it. If any progress is to be made in such a case as that to which Mr. Stuart has referred—a very important class of case in sociological and economic investigations—it is necessary to find some way of specifying a population with a tractable number of parameters in the non-null case. I hope that this will engage the attention of some of those interested in the subject.

Mr. L. T. WILKINS wrote as follows:

I am surprised that so few references have been made to-night to the problem of unidimensionality. It seems to me that we can only expect ranked data to be meaningful and susceptible to statistical analysis if we can first show that the items ranked are in one dimension. I should have thought that practical statisticians and psychologists, not to mention sociologists, were most in need of reliable tests of dimensionality. Frankly I am not at all happy about the basic mathematical models upon which much of the theory of ranking and scaling is based.

It seems to me that psychologists frequently ask people to rank or to assess by paired comparisons something like this:

chalk (*a*)
chalk (*b*)
cheese (*x*).

I have no doubt that some rankings would result from such requests, but not reasonable ones. I wonder how much of the lack of agreement between judges may be due to the lack of "unidimensionality" in the material presented? And, what tests of this can be applied? It seems that two basic models can be used to supply some sort of an answer—the model of a series of overlapping lines, or of content areas analysed in terms of the major axes. In any three items disposed in two dimensions there are three "right" orders determined from ordinate, abscissa or a vector, and these three orders may be quite different from each other or the same, accordingly as we dispose the items.

Perhaps we should consider more carefully if the judges who provide us with the basic data should not, in most cases, be regarded as additional dimensions in each ranking. In any event, I fail to see how we may, with our present knowledge, distinguish between lack of comparability between the items presented (their unidimensionality) and lack of comparability between judges. Only in theoretical models can this problem be overcome. We really need to know what we wish to ascertain before we are able to measure it.

The type of problem which most concerns me is this. I require to predict a particular activity—a pass or a fail—or some similar feature. In the analyses which lead to this prediction I can place several assessments and some other measures; some will be dichotomies, and others classifications of variables into varying numbers of groups. I want to pool these data and to test them against a criterion. So far I can do this only with considerable loss of information, reducing the value of the results obtained.

The authors subsequently replied in writing as follows:

Mr. MORAN: On thinking the matter over, I have come to the conclusion that t and r_s do measure substantially the same aspect of the sample in nearly all cases. Since the meeting I have realized that if they are used to estimate ρ in a bivariate normal population from which the observed rankings are obtained by ranking a sample of n pairs of values, their efficiencies of estimation are asymptotically equal. From this it follows that for a given level of significance, their powers when used as tests that $\rho = 0$ are equal. Incidentally, Hotelling and Pabst have shown that the efficiency of r_s when used as an estimate of ρ in the above case is asymptotically about 91 per cent. of the efficiency of r. The choice between the two coefficients is therefore, in my view, entirely a matter of convenience which may be either computational (r_s being easier than t to find for large n) or mathematical (the distribution of t being simpler and smoother than that of r_s). I suspect that both coefficients are also equally powerful when used as tests against other alternative hypotheses (compare Dr. Daniels's paper).

I also do not agree with those speakers who regard partial ranking coefficients as resembling product moment coefficients only by accident, for they are in fact just the product moment partial correlation coefficients of the scores (either t-scores or r_s-scores). It is very difficult to obtain any results about the distribution of partial rank correlation coefficients owing primarily, I think, to the fact that their denominators are subject to sampling fluctuations. It is possible to get a little further with multiple rank correlation, and I hope to publish something on this at a later date.

Mr. WHITFIELD: Both Mr. Plackett and Mr. Babington Smith have raised the problem of what people in fact do when they rank—whether they perform a series of paired comparisons or whether they make an order arrangement more directly. There seems to be no direct answer to this. It largely depends on the person and the conditions—the variable and the nature of the items. I would agree with Mr. Babington Smith that the more difficult the differentiation between items the greater the tendency to adopt a paired comparison approach. Some of Mr. Babington Smith's comments on the weight-lifting experiment arise from the ambiguity of the paper, and a footnote has been added to correct this. His further extension to cover the random case is very similar to Dr. Daniels's construction, both being derived from a Thurstonian model. This may be a convenient model, and although I do not think it bears any real resemblance to the actual process of judgment, I cannot offer an alternative model. It is possible that if one considered the perception of the interval (or difference) between objects as more fundamental, and their scalar position as derived, one would produce a more acceptable psychological model, but whether it could be made to help the statistics I do not know. I agree with Mr. Babington Smith that if we split up a group for interview we may lose more than we gain when we come to put the groups together for a final rank order. I do not consider this to be a practical proposition, and merely recommend splitting when two (or more) variables are to be related, without any necessity for establishing a final rank order.

I am not perturbed by Mr. Wilkins's criticism of the non-unidimensionality of items presented for comparison by psychologists. It seems perfectly in order to compare "chalk" and "cheese" if one asks people "which do you want more of?" or, "which would you work harder to get?" or "which would you do more to avoid?". The unidimensionality surely lies in the question rather than in the objects, or possibly more correctly, it lies in the responses or potential responses of the individuals.

Finally, I should like to thank the various members for their comments. The problem of subjective estimation is one which needs the attention of both statisticians and psychologists, and any theoretical model must be acceptable to both if it is to be of any practical value.

Dr. DANIELS: I sympathize with Mr. Plackett's exasperation over notation, and think his suggestion of r_S and r_K is a good one. The trouble about a radical change of notation is, of course, that it makes the reading of earlier literature more tiresome than it need be. Mann's use of T seemed to me a sensible compromise, being sufficiently like t and τ for comfort, but it is clearly not ideal. Incidentally, might we not also drop the use of t for number of ties?

Mr. Plackett's query about when ρ is preferable to τ is a little difficult to answer. The worst discrepancy between them occurs when the individuals fall into two groups of about equal size, within which corresponding pairs of ranks are nearly all concordant, but between which they are nearly all discordant (or *vice versa*). Whether ρ or τ is preferable in such cases depends on the relative importance attached to inter-group and intra-group correlations. But in any case if the population is known to be of this form a simple overall measure of correlation is unlikely to satisfy the investigator. The inequality merely states the maximum penalty incurred by ignorance. When something is known about the population model, as in the simple Thurstonian case where θ is relevant, the most suitable statistic may not be any of the Γ coefficients—Mann's K, for example, is not of this form. For small θ, the determination of the coefficients λ_{ij} is practically equivalent to selecting the a_{ij}'s for maximum efficiency.

I agree with Mr. Plackett that the interpretation of an apparently linear relation in Figs. 1 and 2 requires caution, but correlation between the ζ's should not affect the test for non-linearity as outlined in §11. Genuine departure from linearity will still be revealed, though a large sample, as in Donaldson's weighing experiment, may be necessary to make the test sufficiently sensitive.

I am no psychologist and hesitate to disagree with Mr. Babington Smith, but I find it difficult to see how the process of ranking differs much from that of making paired comparisons. Does not the endeavour to avoid inconsistencies when ranking merely make the subject think harder before deciding the order of each pair? It is argued that cases are frequent for which the paired comparison method gives a valid answer where to demand a ranking is nonsensical, but if the latter is true surely a set of preference statements is equally useless without further qualification. Even Dr. Good's suggested procedure would not help matters in that case.

Mr. Stuart has raised an interesting new ranking problem, and his solution should be of value provided due allowance can be made for wisdom after the event in selecting for test large gaps in the mean ranks.

There is one further general point I wish to comment on. The validity of the postulate of a rating scale underlying ranking has been criticized, particularly by Mr. Plackett and Mr. Wilkins. I believe, however, that even in multidimensional cases where a number of factors affect judgment the ranker may rate these factors on imaginary scales, and combine them into a single rating for the purpose of ranking by a kind of mental regression technique. To take a practical example, consider the claim that wool "tops" are graded for quality purely on a visual assessment of fibre diameter, the higher quality top having the lower fibre diameter. Suppose we wish to test whether the assessor's judgment of quality is also influenced by fibre length x', which he knows from experience is positive correlated with fibre diameter x, so that an anomalously long-fibred wool sample, say, might be ranked too low in quality. It is then perhaps appropriate to envisage a two-dimensional Thurstonian model, interpreted from the regression point of view (as in my §6) with a rating scale of the form $y = \theta x + \theta' x'$, the y's being subsequently ranked. The coefficients θ, θ' would not only be inversely proportional to Thurstone's standard deviations σ and σ' but would depend also on the relative weight attached to x' by the assessor. An extension of the method of §10 could be used to estimate θ and θ', and the reality of the influence of x' tested statistically. This type of problem is quite a common one, and I do not see any way of dealing with it without setting up a simple model which one hopes is reasonably near the truth.