$$r_{n_i,x_i} = \sqrt{\frac{s-i+1}{si}}, \tag{15}$$

$$r_{n_{i+1},n_i} = \sqrt{\frac{i(s-i)}{(i+1)(s-i+1)}}. \tag{16}$$

Example 3. The cards of a deck are turned one by one until two aces have appeared. The second ace appears when the 36th card is turned. How many more cards should one expect to have to turn to find a third ace?

Solution. Here $m = 52$, $s = 4$, $i = 2$, $n_2 = 36$.

Then $\bar{n}_2 = 2\cdot\frac{53}{5}$, $\bar{x}_3 = \frac{53}{5}$, and $r_{n_2,x_3} = -\sqrt{\frac{2}{4(4-2+1)}} = -\frac{\sqrt{6}}{6}$. Also $\sigma_{x_3} = \sqrt{4d}$ and $\sigma_{n_2} = \sqrt{6d}$. Since $\frac{x_3 - \bar{x}_3}{\sigma_{x_3}} = r_{n_2,x_3}\frac{(n_2 - \bar{n}_2)}{\sigma_{n_2}}$, we have

$$x_3 = \frac{53}{5} - \frac{2}{\sqrt{6}}\cdot\frac{\sqrt{6}}{6}\left(36 - \frac{106}{5}\right) = \frac{17}{3}.$$

Of course this result could have been obtained more directly by noting that there were two aces left among the 16 remaining cards.

Conclusion. The results given in this note might be useful when it is necessary to estimate the number of items to be drawn in order to secure a desired number of a particular type, such as may be the case in obtaining a sample with previously defined characteristics. Also the note disproves such intuitive notions as the one that when looking for a desired record, one is most likely to have to search the whole pile to find it. As far as methods of sampling inspection are concerned, the one implied in this note has little to recommend it.

CARNEGIE INSTITUTE OF TECHNOLOGY,
PITTSBURGH, PA.

RANK CORRELATION WHEN THERE ARE EQUAL VARIATES[1]

BY MAX A. WOODBURY

If there is given a set of number pairs

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N), \tag{1}$$

we may assign to each variate its "rank" (i.e. one more than the number of corresponding variates in the set greater than the given variate). In this way there is obtained a set of pairs of ranks

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N). \tag{2}$$

[1] Presented at the fall meeting, Mich. section of the Math. Assn. of America, Nov. 18, 1939, Kalamazoo College.

If we assume that $X_i \neq X_j$ and $Y_i \neq Y_j$ when $i \neq j$ then it follows that each integer from 1 to N appears once and only once in the x's and the same holds for the y's. This leads at once to the formulas:

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} i = N(N+1)/2, \tag{3a}$$

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6. \tag{3b}$$

When these results are substituted in the expression for the product moment correlation coefficient we have after simplifying [1],

$$\rho = 1 - 6\sum_{i=1}^{N} D_i^2/N(N^2-1) \qquad \text{where } D_i = x_i - y_i. \tag{4}$$

If we consider the case of equal variates and follow the rule for assigning ranks given in the first paragraph, the resulting method is known as the bracket-rank method. The use of (4) in the calculation of ρ by this method is not strictly valid, because not every integer appears in the summations and so neither (3a) nor (3b) is true.

The more accurate mid-rank method assigns to each of the equal variates the average of the ranks that would be assigned if we were to give them an arbitrary order. This method preserves (3a) but not (3b). In this paper ρ_M indicates the value of ρ as calculated by (4) when the mid-rank method is used.

In a method due to DuBois [2], the equal variates are assigned the same rank so as to satisfy (3b). In this case (3a) is not satisfied.

If we assign the ranks to the equal variates in an arbitrary way, then (3a) and (3b) are of course satisfied and the use of (4) is valid. There are two disadvantages to such a method; first, the equal variates are treated differently, and second, the assignment of ranks is arbitrary. These difficulties are removed if one uses the average of the values of ρ corresponding to all possible ways of arbitrarily assigning ranks to the equal variates. Since ρ is linear in $\sum_i D_i^2$ the average value of ρ may be obtained from the average value of $\sum_i D_i^2$ and the use of (4).

Let us first consider the simple case of two equal variates in one of the variables, say X. It is clear that there are only two possible ways of assigning ranks, and that if we arrange the series by the assigned x ranks, the resulting series differ only in the y ranks corresponding to the equal X variates. If we denote the two x ranks to be assigned by m and $m+1$ and the y's corresponding for a particular arrangement by y_m and y_{m+1} we have for the average $\sum_i D_i^2$ the expression

$$\sum_{x=1}^{m-1}(x-y_x)^2 + \sum_{x=m+2}^{N}(x-y_x)^2 + \tfrac{1}{2}[(m-y_m)^2 + (m+1-y_{m+1})^2 + (m-y_{m+1})^2 + (m+1-y_m)^2]. \tag{5a}$$

By the mid-rank method the corresponding expression is

$$\text{(5b)} \qquad \sum_{x=1}^{m-1}(x-y_x)^2+\sum_{x=m+2}^{N}(x-y_x)^2+(m+\tfrac{1}{2}-y_m)^2+(m+\tfrac{1}{2}-y_{m+1})^2.$$

The correction Δ_2 to be added to the mid-rank $\sum_i D_i^2$ to get the average $\sum_i D_i^2$ is, by subtracting (5b) from (5a) and simplifying,

$$\Delta_2 = \tfrac{1}{2}. \tag{6}$$

To get Δ_K in the more general case of several equal variates, we need only consider the difference between the average value of $\sum_i D_i^2$ and that obtained by the mid-rank method. If there are K equal X variates we may assign the ranks in $K!$ ways, this results in $K!$ permutations of the y ranks for the sets arranged in order of their assigned x ranks. In $(K-1)!$ permutations y_{m+j} corresponds to the x rank of $m+i$ so that the correction to the mid-rank $\sum_{i=1}^{N} D_i^2$ is

$$\begin{aligned}\Delta_K &= \frac{(K-1)!}{K!}\Big[\sum_{j=0}^{K-1}\sum_{i=0}^{K-1}(m+i-y_{m+j})^2\Big]-\sum_{j=0}^{K-1}\left(m+\frac{K-1}{2}-y_{m+j}\right)^2 \\ &= \frac{1}{K}\sum_{j=0}^{K-1}\sum_{i=0}^{K-1}\left[(m+i-y_{m+j})^2-\left(m+\frac{K-1}{2}-y_{m+j}\right)^2\right]=\frac{K(K^2-1)}{12}.\end{aligned} \tag{7}$$

It is to be noticed that the correction is positive and depends *only* on the number of equal X variates. From this it can be concluded that for more than one group of equal variates no matter whether X's or Y's we can obtain the average $\sum_i D_i^2$ by computing a correction for each group and then adding these corrections to get the total correction to the mid-rank $\sum_i D_i^2$. Then as before noted we can by (4) calculate the average ρ (denoted as $\bar{\rho}$).

This correction to $\sum_i D_i^2$ may be converted into a correction to ρ_M. That is

if $\delta_{N,K_i} = \dfrac{6\Delta_{K_i}}{N(N^2-1)} = \dfrac{K_i(K_i^2-1)}{2N(N^2-1)}$, then

$$\bar{\rho} = \rho_M - \sum_i \delta_{N,K_i}, \tag{8}$$

where the summation extends over all groups of equal variates, and K_i is the number of equal variates in the ith group.

A table of δ_{NK} for different values of N and K is given, and also a table of Δ_K. The values Δ_K are given in the top row of the table, while the δ_{NK} are given in the rows below.

Table of Δ_K and δ_{NK}

N \ K	2	3	4	5	6	7	8	9	10	11	12	13
Δ_K	0.5000	2.000	5	10	17.5	28	42	60	82.5	110	143	182
δ_{NK}												
3	1250	—	—	—	—	—	—	—	—	—	—	—
4	0500	2000	—	—	—	—	—	—	—	—	—	—
5	0250	1000	2500	—	—	—	—	—	—	—	—	—
6	0143	0571	1429	2857	—	—	—	—	—	—	—	—
7	0089	0357	0893	1786	3125	—	—	—	—	—	—	—
8	0060	0238	0595	1190	2083	3333	—	—	—	—	—	—
9	0042	0166	0417	0833	1458	2333	3500	—	—	—	—	—
10	0030	0121	0303	0606	1061	1697	2546	3636	—	—	—	—
11	0023	0091	0227	0455	0795	1273	1909	2727	3750	—	—	—
12	0017	0070	0175	0350	0612	0979	1469	2098	2885	3846	—	—
13	0014	0055	0137	0275	0480	0769	1154	1648	2266	3022	3929	—
14	0011	0044	0110	0220	0385	0615	0923	1319	1813	2418	3143	4000
15	0009	0036	0089	0179	0313	0500	0750	1071	1473	1964	2554	3250
16	0007	0029	0074	0147	0257	0412	0618	0882	1213	1618	2103	2676
17	0006	0025	0061	0123	0214	0343	0515	0735	1011	1348	1752	2230
18	0005	0021	0052	0103	0181	0289	0433	0619	0851	1135	1476	1878
19	0004	0018	0044	0088	0154	0246	0368	0526	0724	0965	1254	1596
20	0004	0015	0038	0075	0132	0211	0316	0451	0620	0827	1075	1368
21	0003	0013	0032	0065	0114	0182	0273	0390	0536	0714	0929	1182
22	0003	0011	0028	0056	0099	0158	0237	0339	0466	0621	0807	1028
23	0002	0010	0025	0049	0086	0138	0208	0296	0408	0543	0708	0899
24	0002	0009	0022	0043	0076	0122	0183	0261	0359	0478	0622	0791
25	0002	0008	0019	0038	0067	0108	0162	0231	0317	0423	0550	0700
26	0002	0007	0017	0034	0060	0096	0144	0205	0282	0376	0489	0622
27	0002	0006	0015	0031	0053	0085	0128	0183	0252	0336	0437	0556
28	0001	0005	0014	0027	0048	0077	0115	0164	0226	0301	0391	0498
29	0001	0005	0012	0025	0043	0069	0103	0148	0203	0271	0352	0448
30	0001	0004	0011	0022	0039	0062	0093	0133	0184	0245	0318	0405
35	0001	0003	0007	0014	0025	0039	0059	0084	0116	0154	0200	0255
40	0000	0002	0005	0009	0016	0026	0039	0056	0077	0103	0134	0171
45	0000	0001	0003	0007	0012	0018	0028	0040	0054	0072	0094	0120
50	0000	0001	0002	0004	0007	0011	0016	0023	0032	0043	0055	0070
60	0000	0001	0001	0003	0005	0008	0012	0017	0023	0031	0040	0051
70	0000	0000	0001	0002	0003	0005	0007	0010	0014	0019	0025	0032
80	0000	0000	0001	0001	0002	0003	0005	0007	0010	0013	0017	0021
90	0000	0000	0000	0001	0001	0002	0003	0005	0007	0009	0012	0015
100	0000	0000	0000	0000	0001	0002	0003	0004	0005	0007	0009	0011

As an example of the use of the table we will consider the following problem, [2, p. 56], with the ranks assigned as for the mid-rank method.

Subject	I	II
A	1	2.5
B	4	10
C	4	2.5
D	4	5
E	4	7
F	4	2.5
G	7	8
H	8	2.5
I	9.5	6
J	9.5	12
K	11	11
L	13	13
M	13	9
N	13	14

For the mid-rank method we have

$$\sum_{i=1}^{14} D_i^2 = 119.5, N = 14,$$

$$\rho_M = 1 - \frac{6(119.5)}{14(196 - 1)} = 0.7374.$$

Referring to the table we find that

K_i	Δ_{K_i}	δ_{NK_i}
2	0.5	0.0011
3	2.0	0.0044
4	5.0	0.0110
5	10.0	0.0220
Total	17.5	0.0385

We know that $\bar{\rho} = 1 - \frac{6(119.5 + 17.5)}{14(196 - 1)} = 0.6989$ and in terms of δ_{NK_i}

$$\bar{\rho} = 0.7374 - 0.0385 = 0.6989$$

The value given by DuBois for his method is 0.7511.

Conclusion. A method has been developed for the treatment of rank correlation where there are groups of equal variates. The method consists of applying a generally small correction to the value as ordinarily calculated by the mid-rank method in order to find the value which would be obtained by averaging the values of the rank correlation coefficient for all possible ways of arbitrarily assigning ranks to the equal variates. Thanks are due Professor P. S. Dwyer, without whose aid and encouragement this paper would not have been written.

REFERENCES

[1] YULE AND KENDALL, *Introduction to the Theory of Statistics*, p. 248. London, 1937.
[2] P. DUBOIS, "Formulas and Tables for Rank Correlation," *Psychol. Rec.*, Vol. 3(1939) pp. 45–56.

UNIVERSITY OF MICHIGAN,
ANN ARBOR, MICHIGAN