SOME DISTRIBUTIONS ARISING IN MATCHING PROBLEMS

By P. V. Krishna Iyer

Defence Science Laboratory, New Delhi

I. Introduction

Simple matching is said to occur when elements of the same kind occur in the same order in two sets, each of n elements of k types, arranged at random, with fixed or varying probabilities. When two sets of n cards containing $n_{11}, n_{12}; n_{21}, n_{22}; n_{31}, n_{32}; \ldots, n_{k1}, n_{k2}$; cards of black, white, red, etc., colours are arranged at random in two rows and if we compare the cards of the two rows in order, we may find a number of instances where cards of the same kind occur together in the same order in both the rows. Such occurrences are usually termed ' matchings '. The idea of matchings between two sets can be extended for three or more sets. The matchings here may be between all the sets simultaneously or any two or more consecutive sets.

The distribution of the number of matchings between two or more sets of cards for varying compositions have been studied by Anderson (1943), Bartlett (1937), Battin (1942), Greenwood (1938), Greville (1941), Kaplansky and Riordon (1945), Wilks (1946) and many others. Most of this work relates to the number of matchings between cards of the same kind for ' finite sampling ' with $n_{11}, n_{12}; n_{21}, n_{22}; n_{31}, n_{32}; \ldots; n_{k1}, n_{k2}$; cards of different colours. Obviously the probability for the occurrence of the cards changes as the order changes. It is possible that under certain circumstances this probability may be fixed for the various colours. Such a scheme wherein the probability for the occurrence of any of the cards is independent of the order of the card in the set may be termed ' infinite sampling '. (This appears to be justifiable because when sampling is carried out from an infinite population the probability for any of the cards can be assumed to be fixed.) Much work does not appear to have been done on the distribution of the number of matchings for infinite sampling. The object of this paper is to consider simultaneously, both for finite and infinite sampling, the distribution of the number of matching in two or more sets of cards by certain special methods developed by the author. It is also proposed to extend these methods for the discussion of distributions of a quantitative nature arising by assigning scores, say, $\theta_1, \theta_2, \theta_3, \ldots, \theta_k$ to the cards of various colours for the different sets.

For two sets we shall deal with the distributions of $\Sigma (x_r - y_r)$, $\Sigma \mid x_r - y_r \mid$ and $\Sigma (x_r - y_r)^2$, where x_r and y_r stand for the scores of the r-th cards in the first and the second set respectively. For three sets we shall discuss the distribution of $\Sigma \mid x_r - 2y_r + z_r \mid$ and $\Sigma (\mid x_r - y_r \mid + \mid y_r - z_r \mid)$, where x_r, y_r and z_r are the r-th scores in the three sets of cards. In fact we can consider the distribution of any function $f(x_r y_r z_r)$ of the scores.

2. Simple Matchings

Suppose there are two sets of cards, D_1 and D_2, of k colours. Let there be n cards in each set arranged at random in a sequence. Assume that the probabilities for black, white, red, etc., cards in the two sets are fixed and are $p_{11}, p_{21}, \ldots, p_{k1}$ and $p_{12}, p_{22}, \ldots, p_{k2}$. We shall consider the distribution of (i) the number of matched pairs of a given colour and (ii) the total number of matched pairs for all kinds of colours. For finite sampling, Battin (1942), Wilks (1946) and others have discussed this problem by obtaining the generating functions of the distributions. We shall investigate this problem by a general approach which is applicable both for finite and infinite sampling.

(a) Number of matchings for a single colour in two sets

The probability generating function for the number of matched pairs for a given, say the r-th, colour is given by

$$[p_{r1} p_{r2} x + (1 - p_{r1} p_{r2})]^n \qquad (2.1)$$

This is obvious, because the probability for a single matching of the r-th colour is $p_{r1} p_{r2}$ and the P.G.F. for a single card is $[p_{r1} p_{r2} x + (1 - p_{r1} p_{r2})]$. If p_{r1} and p_{r2} are fixed, the probability for the s-th matching in the sequence is independent of the number of matchings noted earlier or later. Therefore the P.G.F. for n cards is (2.1).

The cumulants are the same as those for the binomial distribution, where P and Q are $p_{r1} p_{r2}$ and $(1 - p_{r1} p_{r2})$ respectively. Hence the first four cumulants are

$$\kappa_1 = nP, \ \kappa_2 = nPQ, \ \kappa_3 = nPQ\,(Q - P), \ \kappa_4 = nPQ\,(1 - 6\,PQ) \qquad (2.2)$$

The probability of s matchings for finite sampling follow from (2.1) by the following procedure: From (2.1) the probability for s matchings is

$$\binom{n}{s} \left(p_{r1} p_{r2}\right)^s \left(1 - p_{r1} p_{r2}\right)^{n-s} \qquad (2.3)$$

Expanding (2.3) it becomes

$$\binom{n}{s}\left[\left(p_{r1}p_{r2}\right)^{s}-\binom{n-s}{1}\left(p_{r1}p_{r2}\right)^{s+1}\right.$$
$$\left.+\binom{n-s}{2}\left(p_{r1}p_{r2}\right)^{s+2}+\left(-1\right)^{n-s}\left(p_{r1}p_{r2}\right)^{n}\right] \quad (2.4)$$

Substitute

$$\frac{n_{r1}^{(s)}\,n_{r2}^{(t)}}{n^{(s)}\,n^{(t)}}=p_{r1}^{s}\,p_{r2}^{t}$$

in (2.4). Then (2.4) reduces to

$$\binom{n}{s}\left[\frac{n_{r1}^{(s)}n_{r2}^{(s)}}{(n^{(s)})^{2}}-\binom{n-s}{1}\frac{n_{r1}^{(s+1)}\,n_{r2}^{(s+2)}}{(n^{(s+1)})^{2}}\right.$$
$$\left.+\binom{n-s}{2}\frac{n_{r1}^{(s+2)}\,n_{r2}^{(s+2)}}{(n^{(s+2)})^{2}}\cdots\right] \quad (2.5)$$

and is the probability for s matchings.

The cumulants of the distribution can be obtained from the first four factorial moments for infinite sampling. The factorial moments for the infinite sampling are

$$\left.\begin{aligned}&\mu'_{[1]}=n\,p_{r1}\,p_{r2}\quad;\quad\mu'_{[2]}=n^{(2)}\,p_{r1}{}^{2}\,p_{r2}{}^{2}\;;\\&\mu'_{[3]}=n^{(3)}\,p_{r1}{}^{3}\,p_{r2}{}^{3};\;\mu'_{[4]}=n^{(4)}\,p_{r1}{}^{4}\,p_{r2}{}^{4}\end{aligned}\right\} \quad (2.6)$$

Making the substitution used in (2.5), the first four factorial moments for finite samplings reduce to

$$\left.\begin{aligned}&\mu'_{[1]}=\frac{n_{r1}\,n_{r2}}{n}\quad;\quad\mu'_{[2]}=\frac{n_{r1}^{(2)}\,n_{r2}^{(2)}}{n^{(2)}}\;;\\&\mu'_{[3]}=\frac{n_{r1}^{(3)}\,n_{r2}^{(3)}}{n^{(3)}};\mu'_{[4]}=\frac{n_{r1}^{(4)}\,n_{r2}^{(4)}}{n^{(4)}}\;.\end{aligned}\right\} \quad (2.7)$$

The second moment reduces to

$$\mu_{2}=\frac{n_{r1}^{(2)}\,n_{r2}^{(2)}}{n^{(2)}}+\frac{n_{r1}\,n_{r2}}{n}-\left(\frac{n_{r1}\,n_{r2}}{n}\right)^{2} \quad (2.8)$$

It may be observed that the distributions for both finite and infinite sampling tend to the normal form when $p_{r1}\,p_{r2}$ is finite and n, n_{11}, n_{12}, ..., etc., are large. When $p_{r1}\,p_{r2}$ is very small, the distribution tends to the Poisson form.

(*b*) *Total number of matchings in two sets*

The P.G.F. for the distribution of the total number of matchings for two sets can be seen to be

$$[(\Sigma p_{r1}\, p_{r2})\, x + (1 - \Sigma p_{r1}\, p_{r2})]^n \tag{2.9}$$

This follows by arguing on the same lines as in (*a*) above. Taking P' and Q' to be

$$P' = \sum_{r=1}^{k} p_{r1}\, p_{r2}; \quad Q' = (1 - \Sigma p_{r1}\, p_{r2})$$

the cumulants are symbolically the same as given in (2.2).

The probability for s matchings from (2.9) is

$$\binom{n}{s} P'^s (1 - P')^{n-s} \tag{2.10}$$

Expanding (2.10), it reduces to

$$\binom{n}{s}\left[P'^s - \binom{n-s}{1} P'^{s+1} + \binom{n-s}{2} P'^{s+2} + \ldots (-)^{n-s} P'^n\right] \tag{2.11}$$

Making the substitution used in (2.5), (2.11) after expansion in powers of $p_{11}, p_{12}, \ldots, p_{k2}$, we get the probability for s matchings for finite sampling. The cumulants for infinite sampling are the same as shown in (2.2). For finite sampling

$$\left.\begin{aligned} \mu_1' &= \frac{\sum_{1}^{k} n_{r1}\, n_{r2}}{n}, \\ \mu_2 &= \frac{\sum n_{r1}^{(2)}\, n_{r2}^{(2)}}{n^{(2)}} + \frac{\sum n_{r1}\, n_{r2}}{n} - \left(\frac{\sum n_{r1}\, n_{r2}}{n}\right)^2 \end{aligned}\right\} \tag{2.12}$$

The distribution of the total number of matchings tends to the normal form because the cumulants can be put as a linear function of n *plus* a function $0\left(\frac{1}{n}\right)$.

(*c*) *Matchings in* m *sets*

The matchings may be between (i) all the sets, (ii) two adjacent sets and (iii) two or more adjacent sets. We shall consider the P.G.F.

for the distribution of the number of matchings between two or more adjacent sets.

The probability for a matching as defined above is

$$P'' = \sum_{l=1}^{m} \sum_{t=1}^{m} \sum_{r=1}^{k} p_{rl}\, p_{rl+1}\, p_{rl+2} \cdots p_{rl+t}$$

where p_{rl} is the probability for a card of a particular colour in the l-th set. Therefore the P.G.F. for this distribution is

$$[P''x + (1 - P'')]^n \tag{2.13}$$

(d) Matchings between cards of different kinds

In sections (*a*), (*b*) and (*c*) above we discussed matching between cards of the same kind. We shall now obtain the distribution of the number of matchings between cards of different kinds. The P.G.F. for the distribution of the number of times that a black and a white card (including *vice versa*) come together is

$$[(p_{11}p_{22} + p_{21}p_{12})\, x + (1 - p_{11}p_{22} - p_{21}p_{12})]^n \tag{2.14}$$

where $p_{11}, p_{21}, p_{31}, \ldots, p_{k1}$ and $p_{12}, p_{22}, \ldots, p_{k2}$ are the probabilities for first and second set of cards being black, white, etc.

The expected number and the variance of the distribution for infinite sampling are

$$n\,(p_{11}p_{22} + p_{21}p_{12}) \text{ and } n\,(p_{11}p_{22} + p_{21}p_{12})\,(1 - p_{11}p_{22} - p_{21}p_{12}) \tag{2.15}$$

For finite sampling the corresponding values are

$$\mu_1' = \frac{n_{11}\, n_{22} + n_{21}\, n_{12}}{n},$$

$$\mu_2 = \frac{1}{n^{(2)}} \left[n_{11}^{(2)} n_{22}^{(2)} + 2 n_{11} n_{22} n_{21} n_{12} + n_{21}^{(2)} n_{12}^{(2)} \right] + \frac{n_{11}n_{22} + n_{21}n_{12}}{n} - \left(\frac{n_{11}n_{22} + n_{21}n_{12}}{n} \right)^2 \tag{2.16}$$

3. Distributions from Matchings of Quantitative Measurements

Let the cards of the two sets be assigned the scores $\theta_1, \theta_2, \ldots, \theta_k$ for the different colours. Then the probabilities for $\theta_1, \theta_2, \ldots, \theta_k$ for

the two sets are $p_{11}, p_{21}, \ldots, p_{k1}$ and $p_{12}, p_{22}, \ldots, p_{k2}$ respectively. We shall consider the distribution of $\Sigma (x_r - y_r)$, $\Sigma \mid x_r - y_r \mid$ and $\Sigma (x_r - y_r)^2$, where x_r and y_r are the scores of r-th cards in the two sets. This aspect of the problem of matching has been mentioned by Wilks (1947) in his book. But no attempt appears to have been made so far to discuss the actual distributions.

The difference between the first and the second sets taken in order can be any of the values shown in the matrix $[(\theta_r - \theta_s)]$, where r and s take all values from 1 to k. The corresponding probabilities are given by $[p_{r1}\, p_{s2}]$. The generating function of the distribution of $\Sigma \mid x_r - y_r \mid$ is given by

$$\left[\sum_{r=1}^{k} p_{r_1} \sum_{t=1}^{k} p_{t2}\, x^{\mid \theta_r - \theta_t \mid}\right]^n \tag{3.1}$$

Expanding (3.1) and taking the coefficients of x^s we get the probability for $\Sigma \mid x_r - y_r \mid$ being equal to s.

If the number of θ's in the first set is fixed, *i.e.*, there are $n_{1_1}\ \theta_1$'s, $n_{21}\ \theta_2$'s, $\ldots$, $n_{k1}\ \theta_{k1}$'s in set 1, while the probability for the θ's in the second set is fixed, as in infinite sampling, then the probability for $\Sigma \mid x_r - y_r \mid$ to be s is the coefficient of x^s in

$$\frac{n!}{n_{11}!\, n_{21}! \ldots n_{k1}!}\, p_{11}^{n_{11}} p_{21}^{n_{21}} \ldots p_{k1}^{nk1} \prod_{r=1}^{k} \left[\sum_{t=1}^{k} p_{t2}\, x^{\mid \theta_r - \theta_t \mid}\right]^{n_{r1}} \tag{3.2}$$

This follows by expanding (3.1) and substituting

$$p_{11}^{\,r} p_{21}^{\,s} p_{31}^{\,t} \ldots = \frac{n_{11}^{(r)} n_{21}^{(s)} n_{31}^{(t)} \ldots}{n^{(r+s+t \ldots)}} \tag{3.3}$$

in the expansion. All the other terms will be zero. In fact (3.2) reduces to

$$\prod_{r=1}^{k} \left[\sum_{1}^{k} p_{t2}\, x^{\mid \theta_r - \theta_t \mid}\right]^{n_{r1}} \tag{3.4}$$

If both sets refer to finite sampling then the probability for $\Sigma \mid x_r - y_r \mid$ to be s is the coefficient of x^s in (3.4) subject to the substitution

$$p_{12}^{\,r} p_{22}^{\,s} p_{32}^{\,t} \ldots = \frac{n_{12}^{(r)} n_{22}^{(s)} n_{32}^{(t)} \ldots}{n^{(r+s+t \ldots)}} \tag{3.5}$$

Using (3.5) it will be seen that (3.4) will reduce to a fairly simple form.

Suppose now $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = 2$, $\theta_4 = 3$ and $\theta_5 = 4$ and $n_{11} = n_{21} = n_{31} = n_{41} = n_{51} = 5$. Then (3.2) for infinite sampling of the second set reduces to

$$\frac{1}{5^{25}} (1 + x + x^2 + x^3 + x^4)^{10} (x + 1 + x + x^2 + x^3)^{10}$$

$$(x^2 + x + 1 + x + x^2)^5 \tag{3.6}$$

The coefficient of x^s in the expression (3.6) is the probability for $\Sigma \mid x_r - y_r \mid = s$. The above distribution has been calculated and given in Table I.

The generating functions for $\Sigma (x_r - y_r)$ and $\Sigma (x_r - y_r)^2$ can also be obtained from (3.2) and (3.4) by substituting $(\theta_r - \theta_s)$ or $(\theta_r - \theta_s)^2$ for $\mid \theta_r - \theta_s \mid$.

We may now obtain the moments of the above distributions. They can be obtained either from the generating functions or by using certain results developed by the author for calculating factorial moments. As the latter method appears to be simpler it is proposed to obtain the first and the second moments by this method.

$$\mu_1' = E \left\{ \sum_{r,\, s=1}^{k} x_{rs} \mid \theta_r - \theta_s \mid \right\}$$

Now $\mid \theta_r - \theta_s \mid$ is fixed, while x_{rs}, the number of times that θ_r and θ_s will occur together, varies. Therefore

$$E \left\{ \sum_{r,\, s=1}^{k} x_{rs} \mid \theta_r - \theta_s \mid \right\} = \sum_{r,\, s=1}^{k} \mid \theta_r - \theta_s \mid E (x_{rs})$$

$$= n \sum_{r,\, s=1}^{k} \mid \theta_r - \theta_s \mid p_{r1}\, p_{s2} \tag{3.7}$$

The second moment can be obtained by taking the expectation of

$$E \left\{ \sum_{r,\, s=1}^{k} \delta x_{rs} \mid \theta_r - \theta_s \mid \right\}^2 \tag{3.8}$$

Expanding (3.8), we get

$$\mu_2 = \Sigma (\theta_r - \theta_s)^2 E (\delta x_{rs})^2 + 2 \Sigma \mid \theta_r - \theta_s \mid \mid \theta_r - \theta_t \mid E (\delta x_{rs} \delta x_{rt})$$

$$+ 2 \Sigma \mid \theta_r - \theta_s \mid \mid \theta_t - \theta_s \mid E (\delta x_{rs}\, \delta_{ts})$$

$$+ 2 \Sigma \mid \theta_r - \theta_s \mid \mid \theta_t - \theta_u \mid E (\delta x_{rs}\, \delta x_{ts}) \tag{3.9}$$

TABLE I

Probability distribution of $\sum_{1}^{25} | x_r - y_r |$ *in the matching of two sets of* 25 *cards, say* x *and* y, *each consisting of five groups of* 5 *cards, each group being numbered* 0, 1, 2, 3 *and* 4, *the probability for the* y's *being fixed*

$\sum_{1}^{25} \lvert x_r - y_r \rvert$	Probability
0	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 4
1	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 3 4 2
2	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 6 6 7 6
3	·0 0 0 0 0 0 0 0 0 0 0 0 0 3 5 1 6 5 0
4	·0 0 0 0 0 0 0 0 0 0 0 0 3 4 6 2 8 1 7
5	·0 0 0 0 0 0 0 0 0 0 0 2 7 1 9 9 8 2 7
6	·0 0 0 0 0 0 0 0 0 0 1 7 7 7 0 1 4 2 0
7	·0 0 0 0 0 0 0 0 0 0 9 9 4 1 4 7 6 9 1
8	·0 0 0 0 0 0 0 0 0 4 8 6 6 0 8 6 8 8 1
9	·0 0 0 0 0 0 0 0 2 1 1 8 6 4 3 4 4 5 9
10	·0 0 0 0 0 0 0 0 8 3 1 3 4 0 9 7 6 9 8
11	·0 0 0 0 0 0 0 2 9 7 1 4 2 2 1 3 7 7 8
12	·0 0 0 0 0 0 0 9 7 5 9 5 3 9 7 6 0 5 7
13	·0 0 0 0 0 0 2 9 6 7 3 5 4 5 8 9 3 3 6
14	·0 0 0 0 0 0 8 4 0 4 0 7 7 2 6 6 2 5 4
15	·0 0 0 0 0 2 2 2 8 9 6 0 9 1 3 1 0 7 5
16	·0 0 0 0 0 5 5 6 1 5 5 6 8 2 0 4 1 5 2
17	·0 0 0 0 1 3 1 0 6 7 6 8 3 5 5 6 1 9 0
18	·0 0 0 0 2 9 2 7 5 3 3 8 0 3 9 7 0 3 4
19	·0 0 0 0 6 2 1 6 2 6 8 9 8 6 8 3 5 3 8
20	·0 0 0 1 2 5 8 1 5 6 6 9 7 7 9 0 4 1 2
21	·0 0 0 2 4 3 2 9 6 5 9 4 5 4 9 1 4 4 6
22	·0 0 0 4 5 0 4 3 8 6 4 0 0 8 2 6 2 6 8
23	·0 0 0 7 9 9 8 9 6 3 4 7 6 8 0 6 5 8 4
24	·0 0 1 3 6 4 7 1 3 7 3 6 0 7 0 2 4 2 5
25	·0 0 2 2 4 0 2 2 9 3 5 6 5 5 1 4 9 2 2
26	·0 0 3 5 4 2 8 3 4 6 3 6 5 6 3 1 8 8 5
27	·0 0 5 4 0 4 0 5 7 5 3 5 8 0 2 3 5 2 1
28	·0 0 7 9 5 8 8 1 2 3 5 3 7 3 4 1 7 4 0
29	·0 1 1 3 2 7 5 2 9 2 1 5 0 1 8 3 8 0 8
30	·0 1 5 5 9 3 2 1 7 6 2 9 7 8 8 7 3 6 5
31	·0 2 0 7 7 6 1 2 4 3 0 8 0 4 7 9 0 9 6
32	·0 2 6 8 1 0 1 5 0 8 7 3 5 9 1 6 0 0 6
33	·0 3 3 5 2 6 1 4 2 3 2 2 2 2 8 0 7 2 5
34	·0 4 0 6 4 7 1 1 0 5 3 2 9 9 2 4 7 2 2
35	·0 4 7 7 9 9 1 6 4 9 8 9 8 4 1 3 1 4 8
36	·0 5 4 5 3 9 4 3 5 1 7 5 1 4 0 0 9 9 4
37	·0 6 0 3 9 8 9 9 1 7 9 0 2 0 3 2 2 6 6
38	·0 6 4 9 3 5 4 1 2 0 0 2 0 8 9 0 7 0 5
39	·0 6 7 7 8 7 0 2 6 0 1 5 2 9 5 9 1 4 4
40	·0 6 8 7 1 9 7 6 7 9 1 4 4 2 4 3 0 5 7
41	·0 6 7 6 5 8 3 4 3 7 9 0 0 7 0 0 5 7 7
42	·0 6 4 6 9 6 0 5 9 0 6 7 4 9 2 3 3 0 4
43	·0 6 0 0 8 1 5 6 2 3 4 7 8 3 8 8 3 3 2
44	·0 5 4 1 8 5 0 6 4 1 7 8 2 2 3 2 1 4 0
45	·0 4 7 4 5 0 2 7 3 0 0 0 4 8 7 7 4 1 6
46	·0 4 0 3 4 0 5 4 0 3 1 5 2 3 1 9 8 4 8

TABLE I—*Contd.*

$\sum_{1}^{25} [1x_r - y_r]$	Probability
47	·0 3 3 2 8 8 0 5 2 5 8 0 7 1 5 1 2 1 9
48	·0 2 6 6 5 3 5 1 7 4 4 4 7 2 7 9 0 4 9
49	·0 2 0 7 0 0 9 7 5 9 5 3 8 1 5 9 6 9 7
50	·0 1 5 5 8 9 1 1 9 2 7 3 8 0 5 5 8 2 4
51	·0 1 1 3 7 7 4 0 4 8 1 8 7 1 7 0 6 1 2
52	·0 0 8 0 4 3 0 7 8 9 8 3 3 3 4 3 5 1 6
53	·0 0 5 5 0 4 1 9 1 7 8 0 5 2 9 0 6 9 1
54	·0 0 3 6 4 3 8 1 9 3 8 7 7 4 4 1 2 6 2
55	·0 0 2 3 3 1 7 0 9 1 1 0 9 5 4 5 6 6 1
56	·0 0 1 4 4 1 0 0 9 4 8 9 6 6 6 0 6 9 5
57	·0 0 0 8 5 9 2 2 9 3 1 4 6 9 6 9 1 6 0
58	·0 0 0 4 9 3 7 7 0 1 6 0 2 2 3 2 2 4 3
59	·0 0 0 2 7 3 1 3 6 9 8 7 2 8 5 6 2 3 9
60	·0 0 0 1 4 5 2 3 7 1 4 4 2 7 9 1 2 1 6
61	·0 0 0 0 7 4 1 2 1 4 8 8 8 1 3 8 0 6 1
62	·0 0 0 0 3 6 2 4 2 8 7 8 2 0 2 8 8 9 9
63	·0 0 0 0 1 6 9 4 5 6 3 3 7 9 1 2 6 1 5
64	·0 0 0 0 0 7 5 5 9 2 5 2 9 2 4 5 8 1 3
65	·0 0 0 0 0 3 2 0 9 0 6 8 7 9 1 1 1 1 6
66	·0 0 0 0 0 1 2 9 2 6 7 2 7 3 4 7 2 3 0
67	·0 0 0 0 0 0 4 9 2 4 2 5 6 5 4 0 8 4 6
68	·0 0 0 0 0 0 1 7 6 6 9 5 7 8 3 2 8 4 8
69	·0 0 0 0 0 0 0 5 9 4 4 7 5 3 7 8 4 4 6
70	·0 0 0 0 0 0 0 1 8 6 4 9 5 6 0 1 2 7 2
71	·0 0 0 0 0 0 0 0 5 4 1 9 2 2 5 8 8 7 8
72	·0 0 0 0 0 0 0 0 1 4 4 6 7 2 0 9 7 8 3
73	·0 0 0 0 0 0 0 0 0 3 5 1 1 9 6 2 5 4 6
74	·0 0 0 0 0 0 0 0 0 0 7 6 5 0 4 2 3 9 2
75	·0 0 0 0 0 0 0 0 0 0 0 1 4 6 9 3 8 8 4
76	·0 0 0 0 0 0 0 0 0 0 0 0 2 4 2 7 9 9 8 7
77	·0 0 0 0 0 0 0 0 0 0 0 0 0 3 3 2 8 6 0 0
78	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 5 7 0 1 9
79	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 6 8 4 4
80	·0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 7 4
	1·0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Now

$$\left.\begin{aligned} E(\delta x_{rs})^2 &= np_{r1}\, p_{s2}\,(1 - p_{r1}\, p_{s2}) \\ E(\delta x_{rs}\, \delta x_{rt}) &= -\, np_{r1}{}^2\, p_{s2}\, p_{t2} \\ E(\delta x_{rs}\, \delta x_{ts}) &= -\, np_{r1}\, p_{t1}\, p_{s2}{}^2 \\ E(\delta x_{rs}\, \delta x_{tu}) &= -\, np_{r1}\, p_{t1}\, p_{s2}\, p_{u2} \end{aligned}\right\} \qquad (3.10)$$

Substituting the above results in (3.9), we get

$$\mu_2 = n\,[\Sigma\,(\theta_r - \theta_s)^2\, p_{r1}\, p_{s2} - \{\Sigma \mid \theta_r - \theta_s \mid p_{r1}\, p_{s2}\}^2] \qquad (3.11)$$

The formula (3.11) holds good only for infinite sampling.

When the first set refers to finite sampling and the second set to infinite sampling μ_1' and μ_2 are obtained by the substitution used in (2.5) in the moments about the origin. Thus

$$\mu_1' = \Sigma \mid \theta_r - \theta_s \mid n_{r1}\, p_{s2} \tag{3.12}$$

$$\left.\begin{aligned} E\,(\delta x_{rs})^2 &= n_{r1}\, p_{s2}\,(1 - p_{s2}) \\ E\,(\delta x_{rs}\, \delta x_{rt}) &= -\, n_{r1}\, p_{s2}\, p_{t2} \\ E\,(\delta x_{rs}\, \delta x_{ts}) &= 0 \\ E\,(\delta x_{rs}\, \delta x_{tu}) &= 0 \end{aligned}\right\} \tag{3.13}$$

Therefore

$$\mu_2 = [\Sigma\,(\theta_r - \theta_s)^2\, n_{r1}\, p_{s2}\,(1 - p_{s2}) - 2\,\Sigma \mid \theta_r - \theta_s \mid \mid \theta_r - \theta_t \mid \times n_{r1}\, p_{s2}\, p_{t2}] \tag{3.14}$$

When both the sets refer to finite sampling

$$\mu_1' = \frac{1}{n}\left[\Sigma \mid \theta_r - \theta_s \mid n_{r1}\, n_{s2}\right] \tag{3.15}$$

$$\left.\begin{aligned} E\,(\delta x_{rs})^2 &= \frac{n^{(2)} n_{r1}^{(2)} n_{s2}^{(2)}}{(n^{(2)})^2} + \frac{n_{r1}\, n_{s2}}{n} - \left(\frac{n_{r1}\, n_{s2}}{n}\right)^2 \\ &= \frac{n_{r1}\, n_{s2}}{n}\left[\frac{(n_{r1} - 1)\,(n_{s2} - 1)}{n - 1} - \frac{n_{r1}\, n_{s2}}{n} + 1\right] \\ E\,(\delta x_{rs} \delta x_{rt}) &= \frac{n^{(2)}\, n_{r1}^{(2)}\, n_{s2}\, n_{t2}}{(n^{(2)})^2} - \frac{n^2_{r1}\, n_{s2}\, n_{t2}}{n^2} \\ E\,(\delta x_{rs} \delta x_{ts}) &= \frac{n_{r1}\, n_{t1}\, n_{s2}}{n}\left[\frac{n_{s2} - 1}{n - 1} - \frac{n_{s2}}{n}\right] \\ E\,(\delta x_{rs}\, \delta x_{tu}) &= \frac{n_{r1}\, n_{t1}\, n_{s2}\, n_{u2}}{n}\left[\frac{1}{n - 1} - \frac{1}{n}\right] \end{aligned}\right\} \tag{3.16}$$

It can be seen that (3.9) reduces to

$$\begin{aligned} \mu_2 = {} & \sum \frac{n_{r1} n_{s2}}{n}\left[\frac{(n_{r1} - 1)\,(n_{s2} - 1)}{n - 1} - \frac{n_{r1}\, n_{s2}}{n} + 1\right](\theta_r - \theta_s)^2 \\ & + 2 \sum \frac{n_{r1}\, n_{s2}\, n_{t2}}{n}\left[\frac{n_{r1} - 1}{n - 1} - \frac{n_{r1}}{n}\right] \mid \theta_r - \theta_s \mid \mid \theta_r - \theta_t \mid \\ & + 2 \sum \frac{n_{r1}\, n_{t1}\, n_{s2}}{n}\left[\frac{n_{s2} - 1}{n - 1} - \frac{n_{s2}}{n}\right] \mid \theta_r - \theta_s \mid \mid \theta_t - \theta_s \mid \\ & + 2 \sum \frac{n_{r1}\, n_{s2}\, n_{t1}\, n_{u2}}{n}\left[\frac{1}{n - 1} - \frac{1}{n}\right] \mid \theta_r - \theta_s \mid \mid \theta_t - \theta_u \mid \end{aligned} \tag{3.17}$$

Using the results for $E(\delta x_{rs})^2$, $E(\delta x_{rs}\,\delta x_{rt})$, $E(\delta x_{rs}\delta x_{is})$, $E(\delta x_{rs}\delta x_{tu})$ we can obtain the expected values and the variances for $E(x_r - y_r)$ and $\Sigma(x_r - y_r)^2$ for all the three situations discussed in this paper.

Considering the distribution of $\Sigma(x_r - y_r)$, $\mu_1' = 0$, $\mu_2 =$ (3.11) or (3.14) or (3.17) with the moduli replaced by the algebraic values.

For the distribution of $\Sigma(x_r - y_r)^2$, μ_1' and μ_2 are given by the same expressions as for $\Sigma\,|\,x_r - y_r\,|$ with the change that $|\,\theta_r - \theta_s\,|$ and $(\theta_r - \theta_s)^2$ are replaced by $(\theta_r - \theta_s)^2$ and $(\theta_r - \theta_s)^4$ respectively.

4. Some Particular Cases

The first two moments of $\Sigma(x_r - y_r)$, $\Sigma\,|\,x_r - y_r\,|$ and $\Sigma(x_r - y_r)^2$ for two sets of cards, each of four, five and six groups of four, five and six cards respectively with scores 0, 1, 2, 3, 4 and 5 for the different groups are given in Table II for the various situations mentioned in this paper.

TABLE II

Mean and variance for finite and infinite sampling

Type of sampling	Distribution of									
	No. of groups	No. of cards in a group	$\Sigma(x_r - y_r)$		$\Sigma\,\lvert\,x_r - y_r\,\rvert$			$\Sigma(x_r - y_r)^2$		
			μ_1'	μ_2	μ_1'	μ_2	coef. of vari. %	μ_1'	μ_2	coef. of vari. %
Both sets infinite	4	4	0	40	20	15	19·37	40	132	28·72
	5	5	0	100	40	36	15·00	100	540	23·24
	6	6	0	210	70	$73\frac{8}{9}$	12·28	210	1673	19·48
1st set finite, 2nd set infinite	4	4	0	20	20	14	18·71	40	116	26·93
	5	5	0	50	40	33·2	14·41	100	470	21·68
	6	6	0	105	70	$67\frac{2}{3}$	11·75	210	1449	18·13
1st set finite, 2nd set finite	4	4	0	0	20	$13\frac{13}{15}$	18·62	40	$106\frac{2}{3}$	25·82
	5	5	0	0	40	$31\frac{2}{3}$	14·07	100	$416\frac{2}{3}$	20·41
	6	6	0	0	70	63·2	11·36	210	1260	16·90

In the above table the coefficient of variation is least for the distribution of $\Sigma\,|\,x_r - y_r\,|$ and therefore between the three statistics

considered above, it may be desirable to take $\Sigma \mid x_r - y_r \mid$ for any statistical test. This question is examined in greater detail in a later section.

5. Distributions Arising by Matching Three Sets of Cards

We shall discuss in this section some of the distributions arising from three sets of cards. Let the cards have the scores $\theta_1, \theta_2, \ldots, \theta_k$ in the three sets. Assume that the probabilities for their occurrence are $p_1, p_2, \ldots, p_k$ and remain the same for all the sets. (The more general case is not considered here because it is not likely to be of much use in practical applications.) As in the previous sections, distributions of various functions of the r-th score in the different sets can be considered. For the present attention is devoted to the following two linear functions:

$$\Sigma \mid x_r - 2y_r + z_r \mid \text{ and } (\Sigma \mid x_r - y_r \mid + \Sigma \mid y_r - z_r \mid)$$

where x_r, y_r and z_r are the scores of cards in the three sets at the r-th place. The generating functions of these distributions are complicated and therefore no attempt has been made in this paper to obtain them. It can be shown that all such distributions tend to the normal form as n, the number of cards in the set increases. The first and the second moments which are useful in examining the deviations from randomness of the scores of the three sets are given below:

(*a*) *Distributions of* $\Sigma \mid x_r - 2y_r + z_r \mid$

Assuming that the probabilities for the occurrence of $\theta_1, \theta_2, \ldots, \theta_k$ are fixed,

$$E\left(\sum_1^n \mid x_r - 2y_r + z_r \mid\right) = n\, E(\mid x_r - 2y_r + z_r \mid) \tag{5.1}$$

Now

$$E(\mid x_r - 2y_r + z_r \mid) = \sum_{r,s,t=1}^{k} \mid \theta_r - 2\theta_s + \theta_t \mid p_r p_s p_t \tag{5.2}$$

Therefore

$$\mu'_1 = n \sum_{r,s,t=1}^{k} \mid \theta_r - 2\theta_s + \theta_t \mid p_r p_s p_t \tag{5.3}$$

Since the probabilities, p's, are fixed and the functions considered involve only the r-th scores of the three sets, the variance for n cards (or observations) is

$n \times$ (variance of the function of the scores in a single column).

The variance of $| x_r - 2y_r + z_r |$ for a single column is

$$E(| x_r - 2y_r + z_r |^2) - \{E | x_r - 2y_r + z_r |\}^2$$

$$= \sum_{r,s,t=1}^{k} (\theta_r - 2\theta_s + \theta_t)^2 p_r p_s p_t - \{\Sigma | \theta_r - 2\theta_s + \theta_t | p_r p_s p_t\}^2 \quad (5.4)$$

Hence n (5.4) is the variance of $\Sigma | x_r - 2y_r + z_r |$ for fixed p's. This result can be used for examining the significance of the difference between three samples.

For finite sampling of $3n$ observations in which $\theta_1, \theta_2, \ldots, \theta_k$ occur $n_1, n_2, \ldots, n_k$ times respectively and are arranged in three rows, the variance can be obtained by substituting

$$\frac{n_1^{(r)} n_2^{(s)} n_3^{(t)} \ldots}{3n^{(r+s+t)}} \text{ for } p_1^r p_2^s p_3^t \ldots$$

in the moments about the origin. The results so obtained are useful in deciding whether the distribution of the $3n$ observations in three rows is random or not.

When the different rows consist of equal number of observations and the probability for the θ's vary from row to row, μ_2 can be obtained from

$$n [\Sigma (\theta_{r1} - 2\theta_{s2} + \theta_{t3})^2 p_{r1} p_{s2} p_{t3} - (\Sigma | \theta_{r1} - 2\theta_{s2} + \theta_{t3} | p_{r1} p_{s2} p_{t3})^2]$$
$$+ \{n \Sigma | \theta_{r1} - 2\theta_{s2} + \theta_{t3} | p_{r1} p_{s2} p_{t3}\}^2 -$$
$$- \left[n \sum \frac{| \theta_{r1} - 2\theta_{s2} + \theta_{t3} | n_{r1} n_{s2} n_{t3}}{n^{(3)}} \right]^2 \quad (5.5)$$

by substituting

$$\frac{n_{r1}^{(s)} n_{s2}^{(t)} n_{t3}^{(u)} \ldots}{n^{(s)} n^{(t)} n^{(u)} \ldots} \text{ for } p_{r1}^s p_{s2}^t p_{t3}^u,$$

where n_{r1}, n_{s2}, n_{t3} represent the number of θ_r's, θ_s's and θ_t's in the first, second and third rows.

(*b*) *Distribution of* $\Sigma | x_r - y_r | + \Sigma | y_r - z_r |$.

The cumulants of this distribution for fixed p's is n times the cumulants for a single column. Now for a single column

$$\mu'_1 = E(| x_r - y_r | + | y_r - z_r |) = 2 \Sigma | \theta_r - \theta_s | p_r p_s \quad (5.6),$$

The variance for a single column is given by

$$\mu_2 = E\{|x_r - y_r| + |y_r - z_r|\}^2 - \{E(|x_r - y_r| + |y_r - z_r|)\}^2$$

$$= E(|x_r - y_r|)^2 + E(|y_r - z_r|)^2 + 2E(|x_r - y_r| \, |y_r - z_r|)$$

$$- 4[\Sigma |\theta_r - \theta_s| p_r p_s]^2$$

$$= 2\{[\Sigma |\theta_r - \theta_s|^2 p_r p_s - (\Sigma |\theta_r - \theta_s| p_r p_s)^2]$$

$$+ [\Sigma |\theta_r - \theta_s| \, |\theta_s - \theta_r| p_r^2 p_s + \Sigma |\theta_s - \theta_r| \, |\theta_r - \theta_s| p_r p_s^2$$

$$+ (\Sigma |\theta_r - \theta_s| \, |\theta_s - \theta_t| + \Sigma |\theta_s - \theta_r| \, |\theta_r - \theta_t| + \Sigma |\theta_r - \theta_t| \, |\theta_t - \theta_s|$$

$$+ \Sigma |\theta_s - \theta_t| \, |\theta_t - \theta_r| + \Sigma |\theta_t - \theta_r| \, |\theta_r - \theta_s| + \Sigma |\theta_t - \theta_s| \, |\theta_s - \theta_r|)$$

$$p_r p_s p_t - (\Sigma |\theta_r - \theta_s| p_r p_s)^2]\} \qquad (5.7)$$

The results (5.6) and (5.7) are useful in deciding whether three samples belong to the same population or not.

By following the procedure described in (*a*) above, the expected value and the variance for finite sampling can be deduced.

It is obvious that the methods described in this section can be extended for four, five and more rows. The functions taken may be orthogonal linear functions of the *r*-th scores in the various sets. From Fisher and Yates Tables the orthogonal functions for four rows and five rows can be as following:

Four rows

$$\left.\begin{aligned} &\text{(i)} \; -3x_r - y_r + z_r + 3w_r \\ &\text{(ii)} \; x_r - y_r - z_r + w_r \\ &\text{(iii)} \; -x_r + 3y_r - 3z_r + w_r \end{aligned}\right\} \qquad (5.8)$$

Five rows

$$\left.\begin{aligned} &\text{(i)} \; -2x_r - y_r + w_r + 2v_r \\ &\text{(ii)} \; 2x_r - y_r - 2z_r - w_r + 2v_r \\ &\text{(iii)} \; -x_r + 2y_r - 2w_r + v_r \\ &\text{(iv)} \; x_r - 4y_r + 6z_r - 4w_r + v_r \end{aligned}\right\} \qquad (5.9)$$

The expected values and the variances of (5.8) and (5.9) can be worked out on the same lines as for three rows. Here also all the

cumulants will be linear functions of the number of observations in each row.

6. Testing of Two Samples

We have considered in this paper the distributions of three *statistics* $\Sigma (x_r - y_r)$, $\Sigma \mid x_r - y_r \mid$ and $\Sigma (x_r - y_r)^2$ for two samples. The distributions of all of them tend to the normal form as n tends to infinity. All the three statistics can, therefore, be used for testing the significance of the difference between two given samples. Before discussing the actual procedure for testing we may note the following:

(1) The *statistics* $\frac{1}{n} \Sigma (x_r - y_r)$, $\frac{1}{n} \Sigma \mid x_r - y_r \mid$ and $\frac{1}{n} \Sigma (x_r - y_r)^2$ are all consistent.

(2) The single tail confidence intervals based on the above *statistics* are consistent in the sense defined by Wald and Wolfowitz (1940), when n tends to infinity. This means, the probability of rejecting the null hypothesis, when it is false, approaches one as the sample size increases.

The consistency of the *statistics* can be easily established by using Tchebycheff's inequality. This readily follows from the fact that the variances of the different *statistics* are inversely proportional to n, the size of the samples. The consistency of the single tail regions can be proved by using the technique employed by Man and Whitney (1947).

Let H_0 be the hypothesis that the probabilities for the occurrence of $\theta_1, \theta_2, \ldots, \theta_k$ in the two samples are $p_1, p_2, \ldots, p_k$. Let the probabilities for the alternate hypothesis H_1 be $p_1 + \delta_1, p_2 + \delta_2, \ldots, p_k + \delta_k$ such that

$$\sum_1^k \delta_r = 0 \tag{6.1}$$

Taking the *statistic* $\Sigma \mid x_r - y_r \mid$, the expected value and the variances for the hypotheses H_0 and H_1 are as follows:

Hypothesis H_0

$$\mu_1' = n \Sigma \mid \theta_r - \theta_s \mid p_r p_s = n z_1 \tag{6.2}$$

$$\mu_2 = n \{\Sigma (\theta_r - \theta_s)^2 p_r p_s - (\Sigma \mid \theta_r - \theta_s \mid p_r p_s)^2\} \tag{6.3}$$

$$= n w_1$$

TABLE III. *Powers of different tests for various alternatives*

In each cell the first figure represents the power of the test $\Sigma (x_r - y_r)$, the second $\Sigma \mid x_r - y_r \mid$ and the third $\Sigma (x_r - y_r)^2$

δ_2 \ δ_1	−·175	−·150	−·125	−·100	−·075	−·050	−·025	0	·025	·050	·075	·100	·125	·150	·175
−·175	·050 ·050 ·050	·053 ·055 ·055	·056 ·062 ·066	·060 ·071 ·083	·064 ·081 ·107	·069 ·095 ·138	·074 ·113 ·179	·079 ·136 ·229	·085 ·165 ·289	·091 ·202 ·358	·099 ·247 ·435	·106 ·300 ·517	·115 ·362 ·601	·125 ·432 ·683	·136 ·508 ·759
−·150	·047 ·045 ·047	·050 ·050 ·050	·053 ·056 ·056	·057 ·063 ·068	·061 ·072 ·087	·065 ·084 ·113	·070 ·100 ·147	·075 ·121 ·192	·081 ·149 ·247	·087 ·183 ·312	·094 ·226 ·387	·102 ·278 ·469	·110 ·340 ·555	·120 ·410 ·641	·130 ·488 ·723
−·125	·044 ·042 ·048	·047 ·045 ·046	·050 ·050 ·050	·053 ·056 ·057	·057 ·063 ·070	·061 ·073 ·090	·066 ·087 ·119	·071 ·106 ·157	·077 ·131 ·206	·083 ·163 ·266	·090 ·204 ·337	·097 ·254 ·418	·105 ·315 ·505	·115 ·385 ·595	·125 ·463 ·683
−·100	·041 ·040 ·051	·044 ·042 ·047	·047 ·046 ·047	·050 ·050 ·050	·054 ·056 ·058	·058 ·064 ·072	·062 ·075 ·094	·067 ·091 ·125	·072 ·113 ·167	·078 ·142 ·221	·085 ·180 ·288	·092 ·228 ·365	·100 ·286 ·452	·109 ·356 ·545	·120 ·434 ·638
−·075	·038 ·039 ·058	·040 ·041 ·052	·043 ·043 ·047	·047 ·046 ·047	·050 ·050 ·050	·054 ·056 ·059	·058 ·065 ·074	·063 ·078 ·098	·068 ·096 ·133	·074 ·121 ·179	·080 ·155 ·239	·087 ·200 ·312	·095 ·255 ·397	·104 ·322 ·491	·114 ·400 ·588
−·050	·035 ·041 ·069	·037 ·042 ·059	·040 ·043 ·052	·043 ·044 ·048	·046 ·046 ·047	·050 ·050 ·050	·054 ·056 ·059	·059 ·066 ·076	·064 ·080 ·103	·069 ·101 ·141	·075 ·131 ·193	·082 ·171 ·260	·090 ·222 ·341	·098 ·286 ·433	·108 ·362 ·534
−·025	·032 ·045 ·083	·034 ·045 ·071	·037 ·045 ·061	·039 ·045 ·053	·043 ·046 ·048	·046 ·047 ·046	·050 ·050 ·050	·054 ·056 ·060	·059 ·067 ·079	·064 ·083 ·109	·070 ·108 ·152	·077 ·142 ·210	·084 ·188 ·284	·093 ·247 ·373	·102 ·320 ·474

0	·029 ·052 ·102	·031 ·052 ·087	·033 ·051 ·078	·036 ·049 ·062	·039 ·048 ·053	·042 ·047 ·048	·046 ·048 ·046	·050 ·050 ·050	·055 ·056 ·061	·060 ·068 ·082	·065 ·087 ·115	·072 ·115 ·164	·079 ·155 ·229	·087 ·218 ·313	·096 ·276 ·411
·025	·026 ·062 ·126	·028 ·062 ·107	·030 ·060 ·090	·033 ·057 ·076	·035 ·054 ·063	·038 ·051 ·054	·042 ·049 ·048	·046 ·048 ·046	·050 ·050 ·050	·055 ·057 ·062	·060 ·070 ·086	·066 ·091 ·123	·073 ·124 ·178	·081 ·170 ·252	·089 ·231 ·346
·050	·023 ·077 ·157	·025 ·076 ·134	·027 ·074 ·113	·029 ·070 ·094	·032 ·065 ·078	·035 ·060 ·065	·038 ·054 ·054	·041 ·050 ·048	·045 ·048 ·045	·050 ·050 ·050	·055 ·057 ·064	·061 ·072 ·090	·067 ·096 ·132	·075 ·134 ·195	·083 ·187 ·280
·075	·020 ·098 ·194	·022 ·097 ·167	·024 ·093 ·141	·026 ·088 ·121	·028 ·082 ·098	·031 ·074 ·080	·034 ·066 ·066	·037 ·058 ·055	·041 ·052 ·047	·045 ·049 ·045	·050 ·050 ·050	·055 ·058 ·065	·062 ·074 ·095	·069 ·102 ·144	·077 ·145 ·216
·100	·017 ·126 ·241	·019 ·125 ·208	·021 ·121 ·178	·023 ·114 ·150	·025 ·106 ·124	·027 ·095 ·102	·030 ·084 ·083	·033 ·073 ·067	·037 ·062 ·055	·040 ·054 ·047	·045 ·049 ·045	·050 ·050 ·050	·056 ·058 ·067	·062 ·077 ·101	·070 ·109 ·158
·125	·015 ·164 ·300	·016 ·163 ·261	·018 ·158 ·224	·019 ·149 ·190	·021 ·139 ·159	·024 ·125 ·131	·026 ·111 ·107	·029 ·096 ·086	·032 ·081 ·069	·036 ·067 ·056	·040 ·056 ·047	·045 ·050 ·044	·050 ·050 ·050	·056 ·059* ·070	·063 ·080 ·109
·150	·012 ·214 ·372	·014 ·212 ·327	·015 ·206 ·283	·016 ·196 ·242	·018 ·183 ·204	·020 ·167 ·170	·022 ·148 ·139	·025 ·128 ·113	·028 ·108 ·090	·031 ·089 ·072	·035 ·072 ·057	·039 ·059 ·047	·044 ·051 ·044	·050 ·050 ·050	·057 ·060 ·073
·175	·010 ·279 ·456	·011 ·276 ·409	·012 ·269 ·358	·014 ·257 ·310	·015 ·241 ·264	·017 ·222 ·221	·019 ·199 ·183	·021 ·174 ·149	·024 ·148 ·119	·027 ·123 ·095	·030 ·099 ·074	·034 ·078 ·059	·039 ·061 ·048	·044 ·051 ·044	·050 ·050 ·050

TABLE IV. *Region of bias and unbiassedness for different tests*

Hypothesis δ_1 (deviation from ·20)	Tests	Region of unbiassedness W_1	Region of bias W_2	Remarks
−·175 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot175$ $W_1 \leqslant -\cdot175$ & $\geqslant 0$ $W_1 \leqslant -\cdot175$ & $\geqslant -\cdot100$	$W_2 > -\cdot175$ $-\cdot175 < W_2 < 0$ $-\cdot175 < W_2 < -\cdot100$	Most powerful
−·150 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot150$ $W_1 \leqslant -\cdot150$ & $\geqslant 0$ $W_1 \leqslant -\cdot150$ & $\geqslant -\cdot075$	$W_2 > -\cdot150$ $-\cdot150 < W_2 < 0$ $-\cdot150 < W_2 < -\cdot075$	Most powerful
−·125 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot125$ $W_1 \leqslant -\cdot125$ & $\geqslant 0$ $W_1 \leqslant -\cdot125$ & $\geqslant -\cdot050$	$W_2 > -\cdot125$ $-\cdot125 < W_2 < 0$ $-\cdot125 < W_2 < -\cdot050$	Most powerful
−·100 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot100$ $W_1 \leqslant -\cdot100$ & $\geqslant \cdot025$ $W_1 \leqslant -\cdot100$ & $\geqslant -\cdot025$	$W_2 > -\cdot100$ $-\cdot100 < W_2 < \cdot025$ $-\cdot100 < W_2 < -\cdot025$	Most powerful
−·075 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot075$ $W_1 \leqslant -\cdot075$ & $\geqslant \cdot025$ $W_1 \leqslant -\cdot075$ & $\geqslant 0$	$W_2 > -\cdot075$ $-\cdot075 < W_2 < \cdot025$ $-\cdot075 < W_2 < 0$	Most powerful
−·050 ..	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot050$ $W_1 \leqslant -\cdot050$ & $\geqslant \cdot025$ $W_1 \leqslant -\cdot050$ & $\geqslant \cdot025$	$W_2 > -\cdot050$ $-\cdot050 < W_2 < \cdot025$ $-\cdot050 < W_2 < \cdot025$	Most powerful excepting a small region near $\delta_1 = -\cdot050$
−·025 .	$\Sigma(x_r-y_r)$ $\Sigma\lvert x_r-y_r\rvert$ $\Sigma(x_r-y_r)^2$	$W_1 \leqslant -\cdot025$ $W_1 \leqslant -\cdot025$ & $\geqslant \cdot050$ $W_1 \leqslant -\cdot025$ & $\geqslant \cdot050$	$W_2 > -\cdot025$ $-\cdot025 < W_2 < \cdot050$ $-\cdot025 < W_2 < \cdot050$	Most powerful for $\delta_2 > \delta_1$ do $\delta_2 < \delta_1$

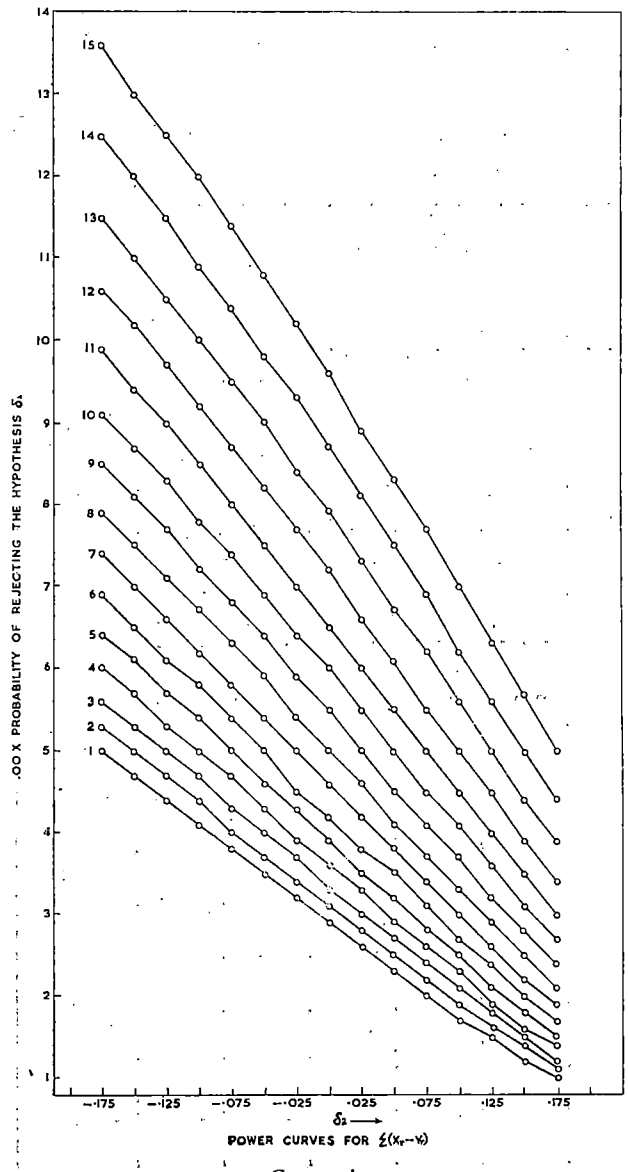
0 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant 0$	$W_2 \geqslant 0$	Most powerful for $\delta_2>\delta_1$
	$\Sigma\lvert x_r-y_r\rvert$	$W_1 \leqslant 0$ & $\geqslant \cdot 050$	$0 < W_2 < \cdot 050$	do $\delta_2<\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant 0$ & $\geqslant \cdot 075$	$0 < W_2 < \cdot 075$	
·025 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 025$	$W_2 \geqslant -\cdot 025$	
	$\Sigma\lvert x_r-y_r\rvert$	$W_1 \leqslant \cdot 025$ & $\geqslant \cdot 075$	$\cdot 025 < W_2 < \cdot 075$	Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant \cdot 025$ & $\geqslant \cdot 100$	$\cdot 025 < W_2 < \cdot 100$	do $\delta_2<\delta_1$
·050 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 050$	$W_2 \geqslant \cdot 050$	
	$\Sigma\lvert x_r-y_r\rvert$	$W_1 \leqslant \cdot 050$ & $\geqslant \cdot 100$	$\cdot 050 < W_2 < \cdot 100$	Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y)^2$	$W_1 \leqslant \cdot 050$ & $\geqslant \cdot 125$	$\cdot 050 < W_2 < \cdot 125$	do $\delta_2<\delta_1$
·075 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 075$	$W_2 \geqslant \cdot 075$	
	$\Sigma\lvert x_r-y_r\rvert$	$W_1 \leqslant \cdot 075$ & $\geqslant \cdot 125$	$\cdot 075 < W_2 < \cdot 125$	Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant \cdot 075$ & $\geqslant \cdot 150$	$\cdot 075 < W_2 < \cdot 150$	do $\delta_2<\delta_1$
·100 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 100$	$W_2 > \cdot 100$	
	$\Sigma\lvert x_r-y_r\rvert$	Unbiassed for all alternatives		Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant \cdot 100$ & $\geqslant \cdot 150$	$\cdot 100 < W_2 < \cdot 150$	do $\delta_2<\delta_1$
·125 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 125$	$W_2 > \cdot 125$	
	$\Sigma\lvert x_r-y_r\rvert$	Unbiassed for all alternatives		Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant \cdot 125$	$W_2 > \cdot 125$	do $\delta_2<\delta_1$
·150 ..	$\Sigma(x_r-y_r)$	$W_1 \leqslant \cdot 150$	$W_2 > \cdot 150$	
	$\Sigma\lvert x_r-y_r\rvert$	Unbiassed for all alternatives		Most powerful for $\delta_2>\delta_1$
	$\Sigma(x_r-y_r)^2$	$W_1 \leqslant \cdot 150$	$W_2 > \cdot 150$	do $\delta_2<\delta_1$
·175 ..	$\Sigma(x_r-y_s)$	Unbiassed for all alternatives less than ·175		Most powerful for $\delta_2>\delta_1$
	$\Sigma\lvert x_r-y_r\rvert$	do		do $\delta_2<\delta_1$
	$\Sigma(x_r-y_r)^2$	do		

Power was not actually calculated beyond the limits $\delta = -\cdot 175$ to $\delta = \cdot 175$. But the regions of bias and unbiassedness have been decided from the trend of the power curves.

Alternate hypothesis H_1

$$\left.\begin{aligned} \mu_1' &= nz_1 + n\psi_1(d_1, d_2, \ldots, d_k) = nz_2 \\ \mu_2 &= nw_1 + n\psi_2(d_1, d_2, \ldots, d_k) = nw_2 \end{aligned}\right\} \qquad (6.4)$$

where ψ_1 and ψ_2 are functions of $d_1, d_2, \ldots, d_k$.



GRAPH 1

Choose the tail region of rejection such that

$$\frac{\Sigma \mid x_r - y_r \mid}{n} - z_1 \geqslant t_n \sqrt{\frac{w_1}{n}},$$

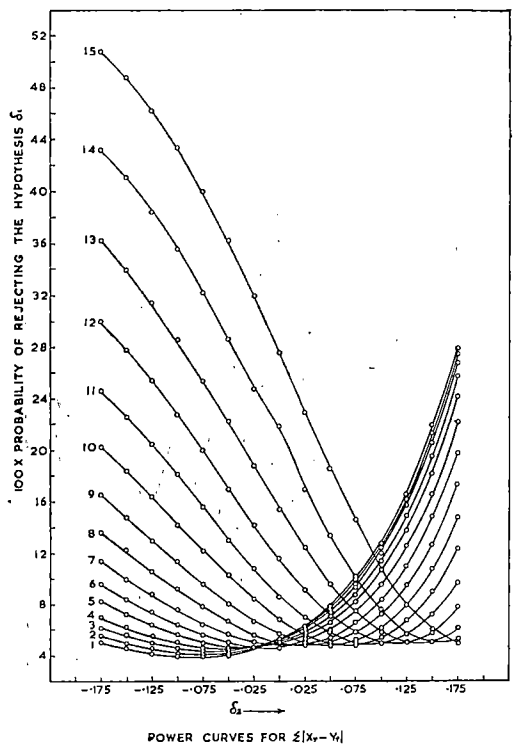
where $t_n \to t$ as n tends to infinity. Then

$$P\left[\frac{\Sigma \mid x_r - y_r \mid}{n} - z_1 \geqslant t_n \sqrt{\frac{w_1}{n}}\right] < \frac{1}{t_n^2} \tag{6.5}$$

Now the chance of accepting H_0 when H_1 is true is given by

$$P\left\{\frac{\Sigma \mid x_r - y_r \mid}{n} - z_1 < t_n \sqrt{\frac{w_1}{n}} \,\middle|\, H_1\right\}$$

$$= P\left\{\frac{\Sigma \mid x_r - y_r \mid}{n} - z_2 \geqslant k \sqrt{\frac{w_2}{n}}\right\} < \frac{1}{k^2} \tag{6.6}$$



GRAPH 2

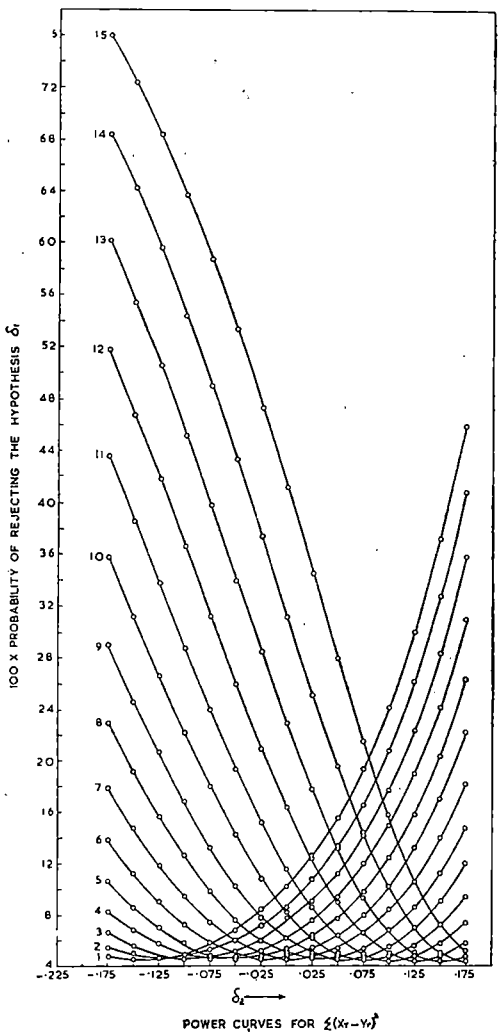
where

$$k = \left| \left[t_n \sqrt{\frac{w_1}{n}} - \psi_1 (d_1, d_2, \ldots, d_k) \right] \right| \Big/ \sqrt{\frac{w_2}{n}},$$

provided n is large and $\psi\ (d_1, d_2 \ldots d_k)$ is positive.

The probability of rejecting the null hypothesis is greater than

$$1 - \frac{w_2}{[t_n \sqrt{w_1} - \sqrt{n}.\psi_1 (d_1, d_2, \ldots, d_k)]^2}. \qquad (6.7)$$



GRAPH 3

In the above Graphs curves 1, 2, 3, ..., 15 refer to hypotheses $\delta_1 = -\cdot 175, -\cdot 150, -\cdot 125, \ldots, +\cdot 175$ respectively.

If ψ_1 is negative, we shall have to consider the other tail as the region of rejection.

We may now describe the actual procedure for testing two samples. Let two samples of sizes n be denoted by

$$\text{Samples} \quad \text{I.} \quad x_1, x_2, x_3, \ldots, x_n$$

$$\text{II.} \quad y_1, y_2, y_3, \ldots, y_n$$

Let $x_1, x_2, \ldots, x_n$; $y_1, y_2, \ldots, y_n$ assume any of the values $\theta_1, \theta_2, \ldots, \theta_k$ with probabilities $p_1, p_2, \ldots, p_k$. The p's are estimated by pooling the two samples together and finding the proportion of $\theta_1, \theta_2, \ldots, \theta_k$ in them. The expected values and variances of $\Sigma (x_r - y_r)$, $\Sigma \mid x_r - y_r \mid$ and $\Sigma (x_r - y_r)^2$ can be calculated by using the results given in the previous sections. The observed values can be computed from the data. When n is not small the standardized deviate

$$z = \frac{x - m}{\sigma},$$

where x and m are the observed and the expected values of the *statistics*, and σ the standard deviation of the distribution, enables us to decide the significance of the difference between the two given samples. The probability that $\mid z \mid \geqslant K$ is calculated on the assumption that the distribution of z is normal. As usual, if the probability that the observed $\mid z \mid$ is less than or equal to 0·05, then we consider the two samples to be different from each other.

7. Power Curves

We have already seen that of the three *statistics* considered in this paper, $\Sigma \mid x_r - y_r \mid$ has the least coefficient of variation for the special cases examined and therefore $\Sigma \mid x_r - y_r \mid$ may be preferred to the other *statistics* in actual practice. We shall examine this point in greater detail by finding the power curves of the three *statistics* for a particular value of $n = 25$ for five values of θ, *i.e.*, $\theta_1 = 0$, $\theta_2 = 1$, $\theta_3 = 2$, $\theta_4 = 3$ and $\theta_5 = 4$ with probabilities $p_1 = p_2 = p_3 = \cdot 2$ and $p_4 = \cdot 2 + \delta$ and $p_5 = \cdot 2 - \delta$ for the hypothesis, $\delta = \delta_1$ as compared to the alternate hypothesis $\delta = \delta_2$ for varying values of δ_1 and δ_2 ranging from $-0 \cdot 175$ to $\cdot 175$ for the first kind of errors 0·05. Tables III and IV, and graphs 1, 2 and 3 show the power of the different tests for various alternative hypotheses for the two equal tail regions of the normal curve for the probability 0·05. The unbiassed reigon as determined from Table III is given in Table IV. It will be seen from this table that for the hypotheses δ_1, ranging from $-\cdot 175$ to $-\cdot 050$,

$\Sigma (x_r - y_r)^2$ is the most powerful, while for $\delta_1 > -\cdot 050$, $\Sigma \mid x_r - y_r \mid$ is the most powerful of the three tests when $\delta_2 > \delta_1$. When $\delta_2 < \delta_1$, $\Sigma (x_r - y_r)^2$ is the most powerful. $\Sigma (x_r - y_r)$ is uniformly the least powerful of the three tests.

8. Summary

This paper deals with a number of distributions arising from matching of two or more decks of cards of k characters which may be qualitative or quantitative. For two samples X and Y of size n, the sequence of observations being

$$X \;\; .. \;\; x_1, x_2, x_3, \ldots, x_n$$

$$Y \;\; .. \;\; y_1, y_2, y_3, \ldots, y_n$$

the distributions of $\Sigma (x_r - y_r)$, $\Sigma \mid x_r - y_r \mid$ and $\Sigma (x_r - y_r)^2$ have been discussed for both finite and infinite sampling. All the cumulants are linear functions of the number of observations and therefore the distributions of the three *statistics* tend to the normal form as n tends to infinity. A table showing the probabilities for the distribution of $\Sigma \mid x_r - y_r \mid$ for $n = 25$, x and y taking any of the values 0, 1, 2, 3 and 4 with probabilities 0·2 has been given.

It has been shown that the three *statistics* are consistent and can be used for testing the significance of the difference between two given samples. Their powers in comparing two samples, each of 25 observations taking the values 0, 1, 2, 3 and 4 with equal probabilities, have been examined for different hypotheses and alternatives and it has been found that $\Sigma \mid x_r - y_r \mid$ is the most powerful for certain regions, while for other regions $\Sigma (x_r - y_r)^2$ is most powerful.

My sincere thanks are due to Prof. D. S. Kothari for suggesting to me to investigate the distribution of $\Sigma \mid x_r - y_r \mid$ for two sets of samples taking the values 0, 1, 2, 3 and 4 with probabilities 0·2. I also wish to express my grateful thanks to Dr. F. C. Auluck, Messrs. S. P. Aggarwal and M. N. Bhattacharyya for helping me in preparing the tables presented in this paper.

References

Anderson, T. W. .. "On Card Matching," *Ann. Math. Stat.*, 1943, **14**, 426.

Bartlett, M. S. .. "Properties of sufficiency and statistical tests," *Proc. Roy. Soc.*, 1937, **160 A**, 294.

Battin, I. L. .. "On the problem of multiple matching," *Ann. Math. Stat.*, 1942, **13**, 294.

Greenwood, J. A. .. "The first four moments of a general matching problem," *Ann, Eug.*, 1940, **10,** 290.

Greville, T. N. E. .. "The frequency distribution of a general matching problem," *Ann. Math. Stat.*, 1941, **12,** 350.

Kaplansky, I. and Riordan, J. "Multiple matching and runs by the symbolic method," *ibid.*, 1945, **16,** 272.

Mann, H. B. and Whitney, D. R. "On a test whether one of two random variables is stochastically larger than the other," *ibid.*, 1947, **18,** 50.

Wald, A. and Wolfowitz, J. "On a test whether two samples are from the same population," *ibid.*, 1940, **11,** 147.

Wilks, S. S. .. *Mathematical Statistics*, Princeton University Press, 1946.