THE THEORY OF PROBABILITY DISTRIBUTIONS OF POINTS ON A LINE*

By P. V. Krishna Iyer

Department of the Design and Analysis of Scientific Experiment, University of Oxford

Introduction and Review of Literature

This paper deals with the theory of certain probability distributions arising from points arranged on a line. The points are of k different characters, which for convenience are described as colours and will consist of m points on a line. There are two situations for consideration in such investigations. They are, to borrow a term used by Mahalanobis (1944), free and non-free sampling. In free sampling the colour of each point is determined, on the null hypothesis, independently of all the other points. The probabilities of the points belonging to different colours, say black, white, etc., are $p_1, p_2, \ldots p_k$ respectively such that $\sum_1^k p_r = 1$. In non-free sampling the number of points from the different colours is specified in advance, say $n_1, n_2, \ldots n_k$, subject to the relation $\sum_1^k n_r = m$. The arrangements of these fixed number of points are varied on the line. The distributions investigated are those for (i) the number of joins between adjacent points of same colour, (ii) the number of joins between adjacent points of two specified different colours, and (iii) the total number of joins between adjacent points of different colours which arise when the m points satisfying the conditions mentioned above are arranged on a line.

Before proceeding further, we shall have a rapid survey of the literature on this topic. Many people have considered the theory of probability distribution of points on a line under different headings, such as, the theory of runs, the random sequence of numerical observations, the distribution of the number of black-white joins, the distribution of groups in a sequence of numerical observations, the distribution of the number of patches. Mood (1940), after summarizing the earlier works by Karl Marbe (1899), Bruns (1906), Von Bortkewicz (1917), Von Mises (1921), Ising (1925), Wishart and Hirschfeld (1936), Stevens (1939), Wald and Wolfowitz (1940) and others has dealt with

* Part of a thesis approved for the Degree of Doctor of Philosophy of the Oxford University.

many aspects of the theory of runs for a multinomial population of points on a line. Mahalanobis (1944) obtained the mean and the variance for the distribution of the number of patches (black, white, black and white) when black and white cells with probabilities p and $q = 1 - p$ were arranged on a line. He has also published the results of extensive sampling experiments for the mean and the variance of the distribution of the number of patches of varying sizes for two characters. For three characters, he has discussed some results based on experimental sampling. It may be mentioned here that some of the results given by Mahalanobis follow from Mood's (1940) work. Later on, it will be shown that some of the results published by Mahalanobis are not correct.

The above brief survey shows that considerable advance has already been made on the distribution theory of runs. But the methods used in deriving the results are very complicated and no attempt has been made to obtain cumulants of orders higher than the second. Some of the second moments given by Mood, who has done extensive work on this topic, are incorrect. Further, a thorough discussion of the limiting forms of the distributions can be better done by the use of cumulants and the moment generating functions. With these objects in view, attempts have been made in this paper to obtain some of the distributions and their cumulants by simpler methods which are new. It will be seen that these investigations, besides leading to further advances, have resulted in throwing more light on the problems dealt with by other people.

1. Distribution of Black-Black Joins for Two Colours

(*a*) *Non-free sampling.* Let n_1 black and n_2 white points be arranged on a line. This can be done in ${}_mC_{n_1}$ ways, where $m = n_1 + n_2$. Out of these,

$$ {}_{n_1-1}C_{r-1} \cdot {}_{n_2+1}C_r \qquad (1.1) $$

arrangements will have $(n_1 - r)$ black-black joins. This follows from the fact that $(n_1 - r)$ black-black joins can be obtained by dividing the n_1 black points into r groups and arranging them in the $(n_2 + 1)$ places from n_2 white points on a line. Now n_1 black points can be divided into r groups in ${}_{n_1-1}C_{r-1}$ ways and the r groups can be arranged in the $(n_2 + 1)$ places in ${}_{n_2+1}C_r$ ways. Therefore the number of arrangements giving $(n_1 - r)$ black-black joins is the product of ${}_{n_1-1}C_{r-1}$ and ${}_{n_2+1}C_r$.

The first four moments with respect to the origin and the cumulants calculated from the distribution given above are

$$\mu_1' = \frac{n_1(n_1-1)}{m}, \quad (1.2)$$

$$\mu_2' = \frac{n_1(n_1-1)}{m} + \frac{n_1(n_1-1)^2(n_1-2)}{m(m-1)}, \quad (1.3)$$

$$\kappa_2 = \frac{n_1(n_1-1)\,n_2(n_2+1)}{m^2(m-1)}, \quad (1.4)$$

$$\mu_3' = \frac{n_1(n_1-1)}{m} + 3\,\frac{n_1(n_1-1)^2(n_1-2)}{m(m-1)} + \frac{n_1(n_1-1)^2(n_1-2)^2(n_1-3)}{m(m-1)(m-2)}, \quad (1.5)$$

$$\kappa_3 = \frac{n_1(n_1-1)\,n_2(n_2+1)(n_1-n_2)(n_1-n_2-2)}{m^3(m-1)(m-2)}, \quad (1.6)$$

$$\mu_4' = \frac{n_1(n_1-1)}{m} + \frac{7n_1(n_1-1)^2(n_1-2)}{m_1(m_1-1)} + \frac{6n_1(n_1-1)^2(n_1-2)^2(n_1-3)}{m(m-1)(m-2)}$$
$$+ \frac{n_1(n_1-1)^2(n_1-2)^2(n_1-3)^2(n_1-4)}{m(m-1)(m-2)(m-3)}, \quad (1.7)$$

$$\kappa_4 = \frac{n_1(n_1-1)\,n_2(n_2+1)}{m^4(m-1)^2(m-2)^2(m-3)}\{m^5 - 6m^4(2n_1-1)$$
$$+ m^3(42n_1^2 - 30n_1 - 1) - 6m^2(10n_1^3 - 7n_1^2 - 3n_1 + 1)$$
$$+ 6m(5n_1^4 + 2n_1^3 - 13n_1^2 + 6n_1) - 36n_1^2(n_1-1)^2\} \quad (1.8)$$

The limiting form of the distribution for large values of n_1 and n_2 can now be discussed. Defining $k = \frac{n_1}{m}$ and adopting the notation $0\left(\frac{1}{m^2}\right)$ for a power series in $\frac{1}{m}$ with zero coefficient for $\frac{1}{m}$,

$$\gamma_1^2 = \frac{(2k-1)^4}{k^2(1-k)^2}\frac{1}{m} + 0\left(\frac{1}{m^2}\right), \quad (1.9)$$

and $$\gamma_2 = \frac{(1 - 12k + 42k^2 - 60k^3 + 30k^4)}{k^2(1-k)^2}\frac{1}{m} + 0\left(\frac{1}{m^2}\right) \quad (1.10)$$

When $m\ (= n_1 + n_2)$ tends to infinity, γ_1 and γ_2 tend to zero. Further it will be found that γ_3, γ_4, etc., also tend to zero. Therefore the distribution of

$$y = \frac{x - \frac{n_1(n_1-1)}{m}}{\sqrt{\frac{n_1(n_1-1)\,n_2(n_2+1)}{m^2(m-1)}}}, \quad (1.11)$$

where x = the observed number of black-black joins, tends to the normal form as m tends to infinity. The distribution takes the Poisson form when n_1 is small and n_2 large. All the cumulants approach the same limit $\frac{n_1^2}{m}$.

(*b*) *Free sampling.* In general, the distribution for free sampling can be obtained from that for non-free sampling by multiplying the frequencies for the different classes by the probability of obtaining n_1 black and n_2 white points in a specified order, and adding up the results so obtained for all values of n_1 from 0 to m. Thus the probability for k black-black joins is

$$\sum_{0}^{m} {}_{n_1-1}C_{n_1-k-1} \cdot {}_{n_2+1}C_{n_1-k}\, p^{n_1} q^{n_2}. \quad (1.12)$$

The sum of the product of $\Sigma x^r f_x$ about zero for non-free sampling and the probability, $p^{n_1} q^{n_2}$, for all values of n_1 from 0 to m gives the rth moment about zero for free sampling. The first four cumulants for values of m higher than the order of the cumulant so obtained are

$$\kappa_1 = (m-1)\, p^2, \quad (1.13)$$

$$\kappa_2 = (m-1)\, p^2 + 2\,(m-2)\, p^3 - (3m-5)\, p^4, \quad (1.14)$$

$$\kappa_3 = (m-1)\, p^2 + 6\,(m-2)\, p^3 - 3\,(m+1)\, p^4 - 12\,(2m-5)\, p^5 + 4\,(5m-11)\, p^5, \quad (1.15)$$

$$\kappa_4 = (m-1)\, p^2 + 14\,(m-2)\, p^3 + (15m-73)\, p^4 - 24\,(5m-11)\, p^5 - 12\,(5m-28)\, p^6 + 24\,(15m-44)\, p^7 - 2\,(105m-279)\, p^8. \quad (1.16)$$

When p is finite and m large, γ_1 and γ_2 tend to zero, and therefore the distribution tends to the normal form. When p is very small the distribution reduces to the Poisson form, because all the cumulants tend to the limit mp^2.

2. Distribution of Black Runs or Black Patches

Stevens (1939), Wald and Wolfowitz (1940), Mood (1940) and Mahalanobis (1944) have dealt with the distribution theory of runs or patches. A run or a patch has been defined as a succession of similar points preceded and succeeded by points of different colours. But they have not given moments higher than the second. Here we shall obtain the higher cumulants and discuss the distribution on this basis.

(*a*) *Non-free sampling.* The number of arrangements having r black runs or patches from n_1 black and n_2 white points on a line is the same as that for $(n_1 - r)$ black-black joins in Section 1(*a*). This is obtained by the same arguments described there. Thus this distribution is the same as that given for black-black joins excepting for the difference that the class value is r in place of $(n_1 - r)$. This will make a difference in the first moment only which is

$$\frac{n_1\,(n_2+1)}{m}. \quad (2.1)$$

The other moments and the remarks made in connection with the distribution of black-black joins are true here also.

(*b*) *Free sampling*. The probability for r black runs is

$$\sum_{n_1=0}^{m} {}_{n_1-1}C_{r-1} \cdot {}_{n_2+1}C_r \, p^{n_1} q^{n_2}. \tag{2.2}$$

The first four cumulants are

$$\kappa_1 = (m-1)\, pq + p, \tag{2.3}$$

$$\kappa_2 = mpq - (3m-5)\, p^2q^2 - 2p^2q, \tag{2.4}$$

$$\kappa_3 = mpq - 9\,(m-1)\, p^2q^2 + 4\,(5m-11)\, p^3q^3 + 2\,(6p^3q^2 - p^2q), \tag{2.5}$$

$$\kappa_4 = mpq - 3\,(7m+6)\, p^2q^2 + 4\,(15m-51)\, p^3q^3 - 2\,(105m-279)\, p^4q^4 + p^2q\,(4m+33) + (120p^4q^3 + 36p^4q - p^3q). \tag{2.6}$$

As in Section 1, when p is fixed and m large, γ_1 and γ_2 tend to zero. Similarly when p is small, the cumulants are all equal to mpq. Hence the limiting forms of the distributions are the same as in Section 1.

3. Distribution of Black Runs of Length r for two Colours

This problem has been dealt with by many people and Mood has discussed it in great detail. But no explicit formula for the number of runs of given length r in all the arrangements with k black runs has been given. The main object here is to obtain this explicit formula. Incidentally we shall obtain some of Mood's results by simpler methods.

(*a*) *Non-free sampling*. The number of times that black runs of length r will occur in the ${}_mC_{n_1}$ arrangements by arranging n_1 black and n_2 white points on a line is

$$\frac{(n_2+1)\,(m-r-1)!}{(n_1-r)!\,(n_2-1)!}. \tag{3.1}$$

Other ways of getting a black run of length r are by inserting a block of $r+2$ points consisting of r black points preceded and succeeded by a white point in the various arrangements that are possible with the remaining $(m-r-2)$ points. These $(m-r-2)$ points can be arranged in

$$\frac{\lfloor m-r-2}{\lfloor n_1-r \; \lfloor n_2-2} \text{ ways.}$$

In each of the arrangements, the block of $(r+2)$ points can be inserted in $(m-r-1)$ ways. Hence the total number of times that black runs of length r occur in the $\dfrac{\lfloor m}{\lfloor n_1 \; \lfloor n_2}$ arrangements of n_1 black and n_2 white points is

$$\frac{2\,\lfloor m-r-1}{\lfloor n_1-r\;\lfloor n_2-1}+\frac{(m-r-1)\;\lfloor m-r-2}{\lfloor n_1-r\;\lfloor n_2-2}=\frac{(n_2+1)\;\lfloor m-r-1}{\lfloor n_1-r\;\lfloor n_2-1}$$

It may appear that the reasoning given above is not valid when $2r < n_1$. But it can be seen that if $f_1, f_2, f_3, \ldots f_s \ldots$ are the number of arrangements in which runs of length r occur once, twice, thrice and so on, then

$$\Sigma_s C_1 f_s = \frac{(n_2+1)\;\lfloor m-r-1}{\lfloor n_1-r\;\lfloor n_2-1} \tag{3.2}$$

similarly $\Sigma \dfrac{s\,(s-1)}{2} f_s$ = number of times that black runs of length r occur twice in the $\dfrac{\lfloor m}{\lfloor n_1\;\lfloor n_2}$ arrangements,

$$= \frac{n_2\,(n_2+1)\;\lfloor m-2r-2}{\lfloor n_1-2r\;\lfloor n_2-2\;\lfloor 2}, \tag{3.3}$$

$\Sigma \dfrac{s\,(s-1)\,(s-2)}{\lfloor 3}$ = number of times that black runs of lengh r occur thrice,

4. Moment Generating Function and Cumulants for the Distribution of Black-Black Joins for Free Sampling

The method described in Section 3 will give, about the origin zero, the factorial moments $\mu'_{[r]}$ for the probability distribution of black-black or white-white joins. Taking the case of black-black joins, if $E(r)$ is the expectation for r black-black joins, then

$$E(r) = \frac{1}{\lfloor r} \mu'_{[r]} \; . \tag{4.1}$$

It will be seen that $E(r)$ is the sum of the expectations of the different ways of obtaining r black-black joins by arranging

(1) $(r + 1)$ black points in one block among $(m - r - 1)$ points black and white;

(2) $(r + 2)$ black points divided into two groups, a group having at least two points for a black-black join, among $(m - r - 2)$ points black and white;

. . . .

. . . .

(r) $2r$ black points divided into r groups, a group having at least two points, among $(m - 2r)$ points black and white.

The expectation for r black-black joins from $(r + s)$ black points is

$$_{r-1}C_{s-1} \cdot {}_{m-r}C_s \cdot p^{r+s}$$

Thus

$$E(r) = \frac{1}{\lfloor r} \mu'_{[r]} = p^{r+1} \sum_{s=1}^{r} {}_{r-1}C_{s-1} \cdot {}_{m-r}C_s \, p^{s-1}. \tag{4.2}$$

Now

$$M_m = 1 + \mu'_{[1]} \frac{\theta}{1} + \mu'_{[2]} \frac{\theta^2}{\lfloor 2} + \dots \mu'_{[r]} \frac{\theta^r}{\lfloor r} \dots, \tag{4.3}$$

where θ stands for $(e^t - 1)$,

is the moment generating function in terms of the factorial moments. If M_m represents the moment generating function for m points it will be found that

$$\frac{M_m - M_{(m-1)}}{p} = \frac{p\theta\,(1 - p^{m-1}\,\theta^{m-1})}{1 - p\theta} + \sum_{r=2}^{m-1} (p\theta)^{(r-1)} \,[M_{(m-r)} - 1]$$

$$= (p\theta)^{n-1} + p\theta\,[M_{(m-2)} + p\theta M_{(m-3)} \dots + (p\theta)^{(m-3)} M_1]$$

$$= (p\theta)^{(m-1)} + p\theta\,\frac{E^{(m-2)} - (p\theta)^{m-2}}{E - p\theta}\, M_1,$$

where E is defined by $E^r M_m = M_{m+r}$. Further simplification gives the difference equation

$$M_{(m+1)} - (1 + p\theta)\, M_m + p\theta\,(1 - p)\, M_{(m-1)} = p^m \theta^{m-1}\,[M_2 - (1 + p\theta)\, M_1].$$

Since $M_0 = 0$, the above equation reduces to

$$M_2 - (1 + p\theta) M_1 = p^m\theta^{m-1} [M_2 - (1 + p\theta) M_1],$$

for $m = 1$. This will be true only if

$$M_2 - (1 + p\theta) M_1 = 0.$$

Therefore the difference equation of the moment generating function is

$$M_{m+1} - (1 + p\theta) M_m + p\theta (1 - p) M_{m-1} = 0. \tag{4.4}$$

The solution of the difference equation is given by

$$M_m = C_1 a^m + C_2 b^m, \tag{4.5}$$

where a and b are the roots of the quadratic equation

$$x^2 - (1 + p\theta) x + p\theta (1 - p) = 0 \tag{4.6}$$

and C_1 and C_2 are so determined that $M_1 = 1$ and $M_2 = 1 + p^2 q$.

$$a = \tfrac{1}{2} \{(1 + p\theta) + \sqrt{(1 - p\theta)^2 + 4p^2\theta}\} \tag{4.7}$$

$$b = \tfrac{1}{2} \{(1 + p\theta) - \sqrt{(1 - p\theta)^2 + 4p^2\theta}\} \tag{4.8}$$

$$C_1 = \frac{\sqrt{(1 - p\theta)^2 + 4p^2\theta} + 1 - p\theta + 2p^2\theta}{2a (a - b)} \tag{4.9}$$

$$C_2 = \frac{\sqrt{(1 - p\theta)^2 + 4p^2\theta} - 1 + p\theta - 2p^2\theta}{2b (a - b)} \tag{4.10}$$

The cumulants, κ_r, for the distribution of black-black joins can now be obtained by evaluating

$$\left[\frac{d^r}{dt^r} \log M_m\right]_{t=0} \tag{4.11}$$

$$\log M_m = \log C_1 + m \log a + \log \left(1 + \frac{C_2}{C_1}\frac{b^m}{a^m}\right) \tag{4.12}$$

$$\left[\frac{d^r}{dt^r} \log M_m\right]_{t=0} = \left[\frac{d^r}{dt^r} \log C_1\right]_{t=0} + m \left[\frac{d^r}{dt^r} \log a\right]_{t=0}$$

$$+ \left[\frac{d^r}{dt^r} \log \left(1 + \frac{C_2}{C_1}\frac{b^m}{a^m}\right)\right]_{t=0} \tag{4.13}$$

It can be seen that the above expression is linear in m, because the last term will always have b as a factor so long as $r < m$, and $b = 0$ when $t = 0$. Therefore the cumulants are linear functions in m.

5. Distribution of Black-Black Joins and Black Runs for k Colours

This distribution has already been worked out by Mood (1940) and it reduces to the case of two colours as the distribution will not in any way be affected, whatever be the number of colours.

6. Distribution of Black-White Joins for two Colours

For free sampling this distribution has been dealt with fully by Wishart and Hirschfeld (1936). For non-free sampling Stevens (1939) worked out the distribution; Wald and Wolfowitz (1940) derived independently the first and second moments. Mahalanobis (1944) gave the first and second moments for both methods of sampling. His second moment does not agree with that of Wald and Wolfowitz (1940). The moments have been obtained here by a new method which can be applied for two and higher dimensional lattices also.

(*a*) *Non-free sampling.* The frequency for $2r$ and $2r+1$ black-white joins when there are n_1 black and n_2 white points distributed at random may be obtained as follows: Divide n_1 black points into r and $(r+1)$ groups. This can be done in ${}_{n_1-1}C_{r-1}$ and ${}_{n_1-1}C_r$ ways respectively. Now $2r$ black-white joins can be had by arranging among n_2 white points (i) r groups of black points omitting the two ends and (ii) $(r+1)$ groups of black points with a black group at each end. Similarly $(2r+1)$ black-white joins can be obtained by arranging $(r+1)$ group of black points between n_2 white points such that there is a black patch at one end and a white at the other end. Thus the number of ways of obtaining $2r$ and $(2r+1)$ black-white joins is

$$ {}_{n_1-1}C_{r-1} \cdot {}_{n_2-1}C_r + {}_{n_1-1}C_r \cdot {}_{n_2-1}C_{r-1} \tag{6.1} $$

$$ \text{and } 2 \cdot {}_{n_1-1}C_r \cdot {}_{n_2-1}C_r, \text{ respectively.} \tag{6.2} $$

From this point onwards moments for free and non-free sampling are not considered independently. It is shown in the next section that one can be derived from the other.

(*b*) *Free sampling.* The sum of the product of the probability, $p^{n_1}q^{n_2}$, and the frequency of black-white joins for a given class value for all values of n_1 from 0 to m gives the probability distribution for free sampling.

It has already been seen that the moments for free sampling are obtained by summing the product of the probability, $p^{n_1}q^{n_2}$, and $\Sigma x^r f_x$ about zero for non-free sampling for all the possible values of n_1. This sum for the r-th moment about zero reduces to the form

$$ A_{1r}(p+q)^{m-2}pq + A_{2r}(p+q)^{m-4}p^2q^2 + A_{3r}(p+q)^{m-6}p^3q^3 + \dots + A_{rr}(p+q)^{m-2r}p^rq^r = A_{1r}pq + A_{2r}p^2q^2 + A_{3r}p^3q^3 + \dots A_{rr}p^rq^r, \tag{6.3} $$

where $A_{1r}, A_{2r}, A_{3r} \dots$ are given by the following relations:

$$ \left.\begin{aligned} S_{r(1,m-1)} &= A_{1r}, \\ S_{r(2,m-2)} &= A_{2r} + {}_{m-2}C_1 \cdot A_{1r}, \\ S_{r(3,m-3)} &= A_{3r} + {}_{m-4}C_1 \cdot A_{2r} + {}_{m-2}C_2 \cdot A_{1r}, \\ S_{r(4,m-4)} &= A_{4r} + {}_{m-6}C_1 \cdot A_{3r} + {}_{m-4}C_2 \cdot A_{2r} + {}_{m-2}C_3 \cdot A_{1r}, \\ &\dots\dots\dots\dots \end{aligned}\right\} \tag{6.4} $$

In the above equations, $S_{r(k,m-k)}$, stands for $\Sigma x^r f_x$ for non-free sampling with k black and $(m-k)$ white points.

From the above relations it follows that the four moments about zero for free sampling can be determined from the frequency distribution of black-white joins for $(1, m-1)$, $(2, m-2)$, $(3, m-3)$, and $(4, m-4)$ black and white points. It will be further seen that the first four moments for non-free sampling for values of n_1 and n_2 greater than four can be obtained by equating the coefficient of $p^{n_1}q^{n_2}$ in equation 6.3 to $S_{r(n_1, n_2)}$. Hence

$$\left.\begin{aligned}
S_{1(n_1,n_2)} &= {}_mC_{n_1}.\mu_1'(n_1, n_2) = {}_{m-2}C_{n_1-1}.A_{11},\\
S_{2(n_1,n_2)} &= {}_mC_{n_1}.\mu_2'(n_1, n_2) = {}_{m-2}C_{n_1-1}.A_{12}+{}_{m-4}C_{n_1-2}.A_{22},\\
S_{3(n_1,n_2)} &= {}_mC_{n_1}.\mu_3'(n_1, n_2) = {}_{m-2}C_{n_1-1}A_{13}+{}_{m-4}C_{n_1-2}.A_{23}\\
&\qquad + {}_{m-6}C_{n_1-3}.A_{33},\\
S_{4(n_1,n_2)} &= {}_mC_{n_1}.\mu_4'(n_1, n_2) = {}_{m-2}C_{n_1-1}.A_{14}+{}_{m-4}C_{n_1-2}A_{24}\\
&\qquad + {}_{m-6}C_{n_1-3}.A_{34}+{}_{m-8}C_{n_1-4}A_{44},
\end{aligned}\right\} \quad (6.5)$$

where $\mu'_{r(n_1,n_2)}$ stands for the rth moment about zero for the non-free distribution with n_1 black and n_2 white points.

The first four moments for the distribution of black-white joins for both free and non-free sampling can now be obtained by finding the frequency distribution of black-white joins for $(1, m-1)$, $(2, m-2)$, $(3, m-3)$ and $(4, m-4)$ black and white points.

Frequency distributions of black-white joins for $(1, m-1)$, $(2, m-2)$, $(3, m-3)$ *and* $(4, m-4)$ *black and white points arranged on a line.*

No. of black-white joins	No. of points			
	black 1 white $(m-1)$	black 2 white $(m-2)$	black 3 white $(m-3)$	black 4 white $(m-4)$
1	2	2	2	2
2	$(m-2)$	$(m-2)$	$(m-2)$	$(m-2)$
3	..	$2(m-3)$	$4(m-4)$	$6(m-5)$
4	..	${}_{(m-3)}C_2$	$(m-4)^2$	$3{}_{(m-4)}C_2$
5	..	..	$2{}_{(m-4)}C_2$	$6{}_{(m-5)}C_2$
6	..	..	${}_{(m-4)}C_3$	$(m-5){}_{(m-5)}C_2$
7	..	..	..	$2{}_{(m-5)}C_3$
8	..	..	..	${}_{(m-5)}C_4$
Total ..	m	${}_mC_2$	${}_mC_3$	${}_mC_4$

The distributions shown above give

$$\begin{aligned}
A_{11} &= 2(m-1),\\
A_{12} &= 2(2m-3),\\
A_{13} &= 2(4m-7),\\
A_{14} &= 2(8m-15),
\end{aligned}$$

$$A_{22} = 4\,(m-2)\,(m-3),$$
$$A_{23} = 12\,(2m-5)\,(m-3),$$
$$A_{24} = 16\,(7m-20)\,(m-3),$$
$$A_{33} = 8\,(m-3)\,(m-4)\,(m-5),$$
$$A_{34} = 48\,(2m-7)\,(m-4)\,(m-5),$$
$$A_{44} = 16\,(m-4)\,(m-5)\,(m-6)\,(m-7).$$

The first four moments for free sampling can now be written with the help of equations 6.3. They are

$$\mu_1' = 2\,(m-1)\,pq, \qquad (6.6)$$
$$\mu_2' = 2\,(2m-3)\,pq + 4\,(m-2)\,(m-3)\,p^2q^2, \qquad (6.7)$$
$$\mu_3' = 2\,(4m-7)\,pq + 12\,(2m-5)\,(m-3)\,p^2q^2 + 8\,(m-3)\,(m-4)\,(m-5)\,p^3q^3, \qquad (6.8)$$
$$\mu_4' = 2\,(8m-15)\,pq + 16\,(7m-20)\,(m-3)\,p^2q^2 + 48\,(2m-7)\,(m-4)\,(m-5)\,p^3q^3 + 16\,(m-4)\,(m-5)\,(m-6)\,(m-7)\,p^4q^4. \qquad (6.9)$$

The respective cumulants for free sampling reduce to

$$\kappa_1 = 2\,(m-1)\,pq, \qquad (6.10)$$
$$\kappa_2 = 2\,(2m-3)\,pq - 4\,(3m-5)\,p^2q^2, \qquad (6.11)$$
$$\kappa_3 = 2\,(4m-7)\,pq - 72\,(m-2)\,p^2q^2 + 32\,(5m-11)\,p^3q^3, \qquad (6.12)$$
$$\kappa_4 = 2\,(8m-15)\,pq - 4\,(84m-185)\,p^2q^2 + 96\,(20m-49)\,p^3q^3 - 32\,(105m-279)\,p^4q^4. \qquad (6.13)$$

The value of κ_2 given by Mahalanobis (1944) reduces to

$4\,(m-1)\,pq - 4\,(3m-5)\,p^2q^2$; this is not correct.

It may be mentioned here that the factorial moments about the origin zero for the free sampling probability distribution of black-white joins can also be obtained like those for the black-black joins described in Section 4 by finding the expectation for r black-white joins. Wishart and Hirschfeld (1936) have dealt with this question in detail. The difference equation of the moment generating function of this distribution is

$$M_{m+1} - M_m - pq\theta\,(\theta + 2)\,M_{m-1} = 0. \qquad (6.14)$$

Proceeding in a manner similar to that indicated in 4 it can be seen that the cumulants for this distribution are also linear functions in m.

The equations 6.5 give the following values for the first four moments about zero for non-free sampling. The usual notation ${}_{n_1}P_r$ has been used to denote the number of permutations of r things taken from n_1 of them.

$$\mu_1'{}_{(n_1,n_2)} = \frac{2n_1n_2}{m}, \tag{6.15}$$

$$\mu_2'{}_{(n_1,n_2)} = \frac{2(2m-3)n_1n_2}{m(m-1)} + \frac{4\,{}_{n_1}P_2\cdot{}_{n_2}P_2}{m(m-1)} \tag{6.16}$$

$$\mu_3'{}_{(n_1,n_2)} = \frac{2(4m-7)n_1n_2}{m(m-1)} + \frac{12(2m-5)\,{}_{n_1}P_2\cdot{}_{n_2}P_2}{m(m-1)(m-2)} + \frac{8\,{}_{n_1}P_3\cdot{}_{n_2}P_3}{m(m-1)(m-2)}, \tag{6.17}$$

$$\mu_4'{}_{(n_1,n_2)} = \frac{2(8m-15)n_1n_2}{m(m-1)} + \frac{16(7m-20)\,{}_{n_1}P_2\cdot{}_{n_2}P_2}{m(m-1)(m-2)} + \frac{48(2m-7)\,{}_{n_1}P_3\cdot{}_{n_2}P_3}{m(m-1)(m-2)(m-3)} + \frac{16\,{}_{n_1}P_4\cdot{}_{n_2}P_4}{m(m-1)(m-2)(m-3)} \tag{6.18}$$

The corresponding cumulants for non-free sampling are

$$\kappa_{1(n_1,n_2)} = \frac{2n_1n_2}{m}, \tag{6.19}$$

$$\kappa_{2(n_1,n_2)} = \frac{2n_1n_2(2n_1n_2-m)}{m^2(m-1)}, \tag{6.20}$$

$$\kappa_{3(n_1,n_2)} = \frac{6n_1n_2}{(m-1)(m-2)} - \frac{8(m+3)n_1^2n_2^2}{m^2(m-1)(m-2)} + \frac{32n_1^3n_2^3}{m^3(m-1)(m-2)}, \tag{2.21}$$

$$\kappa_{4(n_1,n_2)} = \frac{-2(7m^2+13m-6)n_1n_2}{m(m-1)(m-2)(m-3)} + \frac{4(4m^3+45m^2-37m-18)n_1^2n_2^2}{m^2(m-1)^2(m-2)(m-3)} - \frac{96(2m^2+3m-6)n_1^3n_2^3}{m^3(m-1)^2(m-2)(m-3)} + \frac{96(5m-6)n_1^4n_2^4}{m^4(m-1)^2(m-2)(m-3)} \tag{6.22}$$

Here also Mahalanobis's value of

$\kappa_2 = \dfrac{4mpq}{m-1}$ is not correct.

The limiting form of the distribution for free sampling has already been shown to be asymptotically normal by Wishart and Hirschfeld (1936). They have also shown that when $p=q$, the moments are the same as those for a binomial distribution of index $(m-1)$.

The distribution for non-free sampling also tends to the normal form when m is large. This can be seen from the fact that, if $e_r = \frac{n_r}{m}$ is finite, the cumulants for non-free sampling involve m in the first and lower powers only and therefore γ_1 and γ_2 are of the order $\frac{1}{\sqrt{m}}$ and $\frac{1}{m}$ respectively.

7. Distribution of Black-White Joins for k Colours

(*a*) *Non-free sampling.* In the first instance I shall deal with the distribution for three colours. Let there be n_1 black, n_2 white and n_3 red points. The distribution for the number of black patches or runs when there are n_1 black and n_3 red points can be obtained by using the results given in Section 1(*a*). Let the frequency for r black runs be f_r for n_1 black and n_3 red points. Black-white joins can be obtained by introducing a white point or a white run on either side of a black patch or between black patches. A white patch on either side of a black patch gives one black-white join. One between black patches gives two joins. If there are r black patches, there are $2r$ places giving one black-white join, $(n_1 - r)$ places giving two joins and $(n_3 - r + 1)$ places between the red patches will not give any black-white join. It is possible now to obtain the distribution for black-white joins as follows: Divide the n_2 white points into k groups and arrange them in $n_1 + n_3 + 1$ places from n_1 black and n_3 red points. The number of ways of dividing n_2 white points into k groups is ${}_{n_2-1}C_{k-1}$. Distribute the k groups such that there are s, t and u white patches in the $2r$, $n_1 - r$ and $n_3 - r + 1$ places described before. This can be done in

$${}_{2r}C_s \cdot {}_{n_1-r}C_t \cdot {}_{n_3-r+1}C_u \text{ ways.}$$

Hence the total number of ways of distributing the k patches is

$${}_{n_2-1}C_{k-1} \cdot {}_{2r}C_s \cdot {}_{n_1-r}C_t \cdot {}_{n_3-r+1}C_u. \tag{7.1}$$

The number of black-white joins obtained in such an arrangement is $s + 2t$. Taking $s + 2t = a$, the frequency of a black-white joins from j_r $(= {}_{n_1-1}C_{r-1} \cdot {}_{n_3+1}C_r)$ black patches is

$${}_{n_1-1}C_{r-1} \cdot {}_{n_3+1}C_r \sum_{k=1}^{n_2} \sum_{s=0}^{a} {}_{n_2-1}C_{k-1} \cdot {}_{2r}C_s \cdot {}_{n_1-r}C_{\frac{a-s}{2}} \cdot {}_{n_3-r+1}C_{k-\frac{a+s}{2}}$$

Each of the f_r's will give a similar expression for a black-white joins and r takes values from 1 to n_1. Hence the frequency for a black-white joins when there are three colours is

$$\sum_{r=1}^{n_1} {}_{n_1-1}C_{r-1} \cdot {}_{n_3+1}C_r \sum_{k=1}^{n_2} \sum_{s=0}^{a} {}_{n_2-1}C_{k-1} \cdot {}_{2r}C_s \cdot {}_{n_1-r}C_{\frac{a-s}{2}} \cdot {}_{n_3-r+1}C_{k-\frac{a+s}{2}}$$

The first moment $= \frac{2n_1 n_2}{m}$, is obtained by finding the sum of

$$(s + 2t)\ {}_{2r}C_s \cdot {}_{n_1-r}C_t \cdot {}_{n_3-r+1}C_u \cdot {}_{n_2-1}C_{s+t+u-1} \cdot {}_{n_1-1}C_{r-1} \cdot {}_{n_3+1}C_r$$

for all possible values of s, t, u and r and dividing it by

$$\frac{|(n_1 + n_2 + n_3}{|n_1\ |n_2\ |n_3}.$$

Similarly the second moment works out to be

$$\frac{4n_1 n_2 (m - n_1)(m - n_2)}{m^2 (m - 1)} - \frac{2n_1 n_2 (m - n_1 - n_2 + 1)}{m (m - 1)}, \quad (7.2)$$

where $m = n_1 + n_2 + n_3$.

The formulæ for the first and the second moments show that the moments for the distribution of black-white joins depend only on the total number of points (m) and the number of black (n_1) and white (n_2) points. Hence these formulæ hold good for any number of colours also.

(*b*) *Free sampling.* As in the case of two colours, the distribution can be obtained by multiplying the frequency for a given class value for n_1 white, n_2 black and n_3 red points by $p_1^{n_1} p_2^{n_2} p_3^{n_3}$ and adding them up for all values of n_1, n_2 and n_3, where p_1, p_2 and p_3 are the probabilities for the different colours. The moments can be got by multiplying the values for $\Sigma x^r f_x$ for non-free sampling by $p_1^{n_1} p_2^{n_2} p_3^{n_3}$ and adding them up for all possible values of n_1, n_2 and n_3. The first two moments so calculated are

$$2 (m - 1) p_1 p_2, \quad (7.3)$$

and

$$2 (m - 1) p_1 p_2 + 2 (m - 2) p_1 p_2 (p_1 + p_2) - 4 (3m - 5) p_1^2 p_2^2 \quad (7.4)$$

It will be seen that these two moments involve only m, p_1 and p_2. Hence they are true for k colours also.

As in the previous investigations in Section 6, it can be shown that the power of m in the third and fourth cumulants is unity and therefore the value of γ_1 and γ_2 tend to the limit zero when m is large. Thus the limiting form of the distribution is normal for large values of m. When p_1 and p_2 are very small and m is large, all the cumulants reduce to $2mp_1p_2$. It follows, therefore, that the distribution tends to Poisson's form.

8. Distribution of the Total Number of Joins between Points of Different Colours for Three Colours

(*a*) *Non-free sampling.* Let there be n_1 black, n_2 white and n_3 red points. Let f_r* be the frequency of r patches for n_1 black and

* This distribution is the same as that for black-white joins with the classes increased by unity.

n_2 white points. One red patch introduced between a black and a white patch increases the total number of patches by unity. If on the other hand, a red patch is introduced between two black or white points, *i.e.*, inside any of the black or white patches, it results in producing two extra patches. There are $(r + 1)$ places giving one extra patch and $(n_1 + n_2 - r)$ places giving two extra patches for each red patch introduced in the arrangement of n_1 black and n_2 white points with r patches. Divide the n_3 colours into k patches and introduce them between the $(r + 1)$ and $(n_1 + n_2 - r)$ places described above. The number of ways of arranging k red patches such that s of them are in the $(r + 1)$ places and $(k - s)$ in the $(n_1 + n_2 - r)$ places is

$$f_r \cdot {}_{n_3-1}C_{k-1} \cdot {}_{r+1}C_s \cdot {}_{n_1+n_2-r}C_{k-s}. \qquad (8.1)$$

The number of patches for each of the above arrangements is $r + s + 2(k - s)$. Taking $r + s + 2(k - s) = a$, the frequency for a patches in the whole distribution can be obtained by summing 8.1 over all values of r, s and k subject to the conditions

$$r + s + 2(k - s) = a, \; r \leqslant n_1 + n_3 \text{ and } k \leqslant n_3.$$

The first and the second moments can be determined by summing

$$[r + s + 2(k - s)] \times (8.1) \text{ and } [r + s + 2(k - s)]^2 \times (8.1)$$

for all possible values of r, s and k. This gives

$$\mu_1' = \frac{2}{m} \dagger(\Sigma n_r n_s) + 1, \qquad (8.2)$$

$$\mu_2 = -2\frac{\Sigma n_r n_s}{m(m-1)} + \frac{2n_1 n_2 n_3}{m(m-1)} + 4\frac{\Sigma n_r^2 n_s^2}{m^2(m-1)}. \qquad (8.3)$$

The average for the total number of joins between points of different colours is

$$\mu_1' - 1 = \frac{2}{m}(\Sigma n_r n_s). \qquad (8.4)$$

The second moment is the same whether the distributions considered are for patches or for joins.

(*b*) *Free sampling*. The distribution and the moments for free sampling can be obtained by following the same procedure as in Section 7(*b*). The first and the second moments in this case are

$$\mu_1' = 2(m - 1)\Sigma p_r p_s, \qquad (8.5)$$

$$\mu_2 = 2(2m - 3)\Sigma p_r p_s - 2(9m - 14)p_1 p_2 p_3 - 4(3m - 5)\Sigma p_r^2 p_s^2. \qquad (8.6)$$

---

† Throughout this paper the summations for the μ's have been taken for all values of r, s, t and u from 1 to k (k being the number of colours) subject to the condition $r < s < t < u$.

I shall now indicate how the second moment can be obtained by using an extension of the method described in Section 6(*b*). For two colours it was shown that the second moment can be calculated by finding the frequency distributions of black-white joins for $(1, m-1)$ and $(2, m-2)$ black and white points. When there are three colours the second moment can be worked out from the frequency distributions of joins between points of different colours for the following combination of points: 1 black, $m-1$ white; 2 black, $m-2$ white; and 1 black, 1 red, $m-2$ white. It can be seen that the second moment about zero will reduce to

$$A_{12}(p_1+p_2+p_3)^{m-2}\Sigma p_r p_s + A_{112}(p_1+p_2+p_3)^{m-3}p_1p_2p_3 + A_{22}(p_1+p_2+p_3)^{m-4}\Sigma p_r^2 p_s^2 = A_{12}\Sigma p_r p_s + A_{112}p_1p_2p_3 + A_{22}\Sigma p_r^2 p_s^2;$$

in which

$A_{12} = \Sigma x^2 f_x$ about zero for the distribution with 1 black and $(m-1)$ white points;

$A_{112} = \Sigma x^2 f_x$ for the distribution with 1 black, 1 red and $(m-2)$ white points $- A_{12} \times$ coefficient of terms like $p_1^{m-2}p_2p_3$ in the expansion of $(p_1+p_2+p_3)^{m-2}\Sigma p_r p_s$; and

$A_{22} = \Sigma x^2 f_x$ for the distribution with 2 black and $(m-2)$ white points $- A_{12} \times$ coefficient of terms like $p_1^{m-2}p_2^2$ in $(p_1+p_2+p_3)^{m-2}\Sigma p_r p_s$.

The values of A_{12} and A_{22} are respectively the same as the coefficients of $\Sigma p_r p_s$ and $\Sigma p_r^2 p_s^2$ for two colours. To find A_{112}, form the frequency distribution for one black, one red and $(m-2)$ white points.

Frequency distribution of joins between points of different colours for 1 black, 1 red and (m − 2) *white points*

No. of joins	Frequency
2	6
3	$6(m-3)$
4	$(m-3)(m-4)$

$$\Sigma x^2 f_x = 16m^2 - 58m + 54$$

$$A_{112} = 16m^2 - 58m + 54 - 2(2m-3)^2 = 2(4m-9)(m-2).$$

Therefore, the second moment for the probability distribution of the total number of joins for three colours is

$$\mu_2' = 2(2m-3)\Sigma p_r p_s + 2(4m-9)(m-2)p_1p_2p_3 + 4(m-2)(m-3)\Sigma p_r^2 p_s^2, \quad (8.7)$$

$$\mu_2 = 2(2m-3)\Sigma p_r p_s - 2(9m-14)p_1p_2p_3 - 4(3m-5)\Sigma p_r^2 p_s^2 \quad (8.8)$$

9. First and Second Moments for Four and More Colours for the Total Number of Joins between Points of Different Colours

It can be seen that the expectations of the total number of joins for free and non-free sampling are

$$2\,(m-1)\,\Sigma p_r\,p_s, \tag{9.1}$$

and

$$\frac{2}{m}\,\Sigma n_r n_s \text{ respectively.} \tag{9.2}$$

The second moment about zero will be of the form

$$A_{12}\,\Sigma p_r\,p_s + A_{112}\,\Sigma p_r\,p_s\,p_t + A_{22}\,\Sigma p_r^2 p_s^2 + A_{1112}\,\Sigma p_r\,p_s\,p_t\,p_u,$$

in which A_{12}, A_{112} and A_{22} are the same as those for three colours. A_{1112} can be obtained by getting the distribution for the total number of joins for 1 black, 1 white, 1 red and $(m-3)$ green points.

Frequency distribution for total number of joins for 1, 1, 1, $(m-3)$ *black, white, red and green points*

No. of joins	Frequency
3	24
4	$36\,(m-4)$
5	$12\,(m-4)\,(m-5)$
6	$(m-4)\,(m-5)\,(m-6)$

$\Sigma x^2 f_x$ about zero $= 36m^3 - 240m^2 + 450m - 408$.

$$A_{1112} = \Sigma x^2 f_x - A_{12} \times \text{coefficient of } p_1^{m-3}\,p_2 p_3 p_4 \text{ in } (p_1+p_2+p_3+p_4)^m \;\; 2$$
$$\Sigma p_r\,p_s - A_{112} \times \text{coefficient of } p_1^{m-3}\,p_2 p_3 p_4 \text{ in } (p_1+p_2+p_3+p_4)^m \;\; 3$$
$$\Sigma p_r\,p_s\,p_t,$$

$$\begin{aligned} i.e.,\ A_{1112} &= (36m^3 - 240m^2 + 540m - 408) - 6\,(2m-3)\,(m-2)^2 \\ &\quad - 2\,(4m^2 - 17m + 18)\,(3m-8) \\ &= -8\,(m-2)\,(m-3) \end{aligned}$$

$$\begin{aligned} \mu_2' &= 2\,(2m-3)\,\Sigma p_r\,p_s + 2\,(4m-9)\,(m-2)\,\Sigma p_r\,p_s\,p_t \\ &\quad + 4\,(m-2)\,(m-3)\,\Sigma p_r^2 p_s^2 - 8\,(m-2)\,(m-3) \\ &\quad \Sigma p_r\,p_s\,p_t\,p_u, \end{aligned} \tag{9.3}$$

$$\begin{aligned} \mu_2 &= 2\,(2m-3)\,\Sigma p_r\,p_s - 2\,(9m-14)\,\Sigma p_r\,p_s\,p_t \\ &\quad - 4\,(3m-5)\,\Sigma p_r^2 p_s^2 + 8\,(3m-5)\,\Sigma p_r\,p_s\,p_t\,p_u. \end{aligned} \tag{9.4}$$

$$\begin{aligned} \mu_2'{}_{(n_1, n_2 \ldots n_k)} &= \frac{2\,(2m-3)\,\Sigma n_r n_s}{m\,(m-1)} + \frac{4\Sigma n_r\,(n_r-1)\,(n_s)\,(n_s-1)}{m\,(m-1)} \\ &\quad + \frac{2\,(4m-9)\,\Sigma n_r n_s n_t}{m\,(m-1)} - \frac{8\,\Sigma\, n_r n_s n_t n_u}{m\,(m-1)}, \end{aligned} \tag{9.5}$$

$$\mu_{2(n_1,n_2\cdots n_k)} = -\frac{2}{m(m-1)}\Sigma n_r n_s + \frac{2}{m(m-1)}\Sigma n_r n_s n_t + \frac{4}{m^2(m-1)}\Sigma n_r^2 n_s^2 - \frac{8}{m^2(m-1)}\Sigma n_r n_s n_t n_u \qquad (9.6)$$

It can be shown that the power of m in all the cumulants for free sampling is unity and therefore the values of γ_1 and γ_2 will tend to zero when m is large, *i.e.*, the distribution for the total number of joins tends to the normal form as m increases. This will be true for non-free sampling also, for the reasons mentioned in Section 1 (*a*).

10. First and Second Moments of a Special Distribution Useful in Examining the Departure from Randomness of the Spread of Disease in a Given Interval in Plants on a Line

Suppose there are m plants on a line, n_1 of which are diseased at a particular time. Let n_2 new plants get affected by the disease a month later. It is required to test whether the spread of the disease between the first and the second observation can be considered to be random or not.

This can be done by finding the standardized deviate for the difference between the expected and the observed values for the number of times that (i) two newly diseased plants and (ii) a newly diseased and a healthy plant occur together in the observed configuration of plants, provided the first and the second moments for these distributions are known.

It can be shown by the methods developed in this paper that the expected values, $E(P)$ and $E(Q)$, and the corresponding variances, $V(P)$ and $V(Q)$, for these distributions are as given below:—

(i) *Number of times that two newly diseased plants occur together.*

$$E(P) = \frac{T_1 n_2 (n_2 - 1)}{(m - n_1)(m - n_1 - 1)}, \qquad (10.1)$$

$$V(P) = \frac{T_1 n_2 (n_2 - 1)}{(m - n_1)(m - n_1 - 1)} + \frac{2T_2 n_2 (n_2 - 1)(n_2 - 2)}{(m - n_1)(m - n_1 - 1)(m - n_1 - 2)} + \frac{\{T_1(T_1 - 1) - 2T_2\} n_2 (n_2 - 1)(n_2 - 2)(n_2 - 3)}{(m - n_1)(m - n_1 - 1)(m - n_1 - 2)(m - n_1 - 3)} - [E(P)]^2. \qquad (10.2)$$

(ii) *Number of times that a newly diseased and a healthy plants occur together.*

$$E(Q) = \frac{2T_1 n_2 (m - n_1 - n_2)}{(m - n_1)(m - n_1 - 1)}, \qquad (10.3)$$

$$V(Q) = \frac{2(T_1+T_2)\,n_2\,(m-n_1-n_2)}{(m-n_1)(m-n_2-1)}$$

$$+ \frac{4[T_1(T_1-1)-2T_2]\,n_2(n_2-1)(m-n_1-n_2)(m-n_1-n_2-2)}{(m-n_1)(m-n_1-1)(m-n_1-2)(m-n_1-3)}$$

$$- [E(Q)]^2, \qquad (10.4)$$

where $T_1 = m - n_1 - k$ and $T_2 = m - n_1 - 2k$; k being the number of runs of healthy plants at the time of the first observation.

11. Applications

The following are some of the situations in which the results given in this paper can be used for testing the departure from randomness of an observed distribution:—

(a) distribution of diseased plants observed in plants on a line;

(b) spread of disease in a given interval;

(c) a given sequence of observations.

The method will be clear from the examples that follow.

(a) *Distribution of diseased plants on a line.* Suppose that there are forty plants on a line, of which eleven diseased are distributed in the manner shown in Fig. 1.

OOOOO—XX—OOOOOO—XXX—OO—X—OOO—X—OOOO—XX—OOOO—XX—OOOOO

'O' denotes a healthy plant

'X' denotes a diseased plant

Fig. 1.

The number of joins between healthy and diseased plants in the above distribution is twelve. The expected number

$$E_m(R) = \frac{2n_1n_2}{m} = \frac{2 \times 11 \times 29}{m} = 15 \cdot 95.$$

The variance

$$V_m(R) = \frac{2n_1n_2(2n_1n_2-m)}{m^2(m-1)} = \frac{2\times 11\times 29\,(2\times 11\times 29-40)}{40 \times 40 \times 39}$$

$$= 6 \cdot 114,$$

and therefore

$$x = \frac{R - E_m(R)}{\sqrt{V_m(R)}} = \frac{12 - 15 \cdot 95}{\sqrt{6 \cdot 114}} = -1 \cdot 60.$$

Since x does not exceed $1 \cdot 96$, the 5% significance level for a normal deviate, there is no serious reason to doubt the null hypothesis that the diseased plants are distributed at random. This method has also been used by Vander Plank (1946).

(*b*) *Spread of disease in a given interval.* Suppose that a later count on the plants of Fig. 1 shows ten more plants to have contracted the disease, and that these are distributed as shown in Fig. 2.

OOOOOXX++++O+XXXOOXOOOXO+OOXXO++OXXOO++O

'O' denotes a healthy plant
'X' ,, diseased plant at the time of first counting
'+' ,, a newly diseased plant

FIG. 2.

In the above distribution $T_1 = 22$ and $T_2 = 15$. From (10.1) and (10.2)

$$E(P) = \frac{22 \times 10 \times 9}{29 \times 28} = 2 \cdot 438;$$

$$V(P) = 2 \cdot 438 + \frac{2 \times 15 \times 10 \times 9 \times 8}{29 \times 28 \times 27}$$

$$+ \frac{(22 \times 21 - 2 \times 15)\, 10 \times 9 \times 8 \times 7}{29 \times 28 \times 27 \times 26} - (2 \cdot 438)^2$$

$$= 1 \cdot 30.$$

The number of joins between two newly diseased plants adjacent to each other is 5. Hence

$$x = \frac{5 - 2 \cdot 438}{\sqrt{1 \cdot 30}} = 2 \cdot 25,$$

which, at least by the normal approximation, is a highly significant deviation. The distribution of the newly diseased plants therefore cannot be considered random.

(*c*) *A given sequence of observations.* Suppose the order of occurrence of nine types of events *A, B, C, D, E, F, G, H,* and *J* is as given below:—

D H F D D H A A G H C G D D F D F D E J J G D J A J A
H H A H J H D A G C D D D C D D F F E A E G F B H B
C G B A E E E D E B D A G C B D C B C H G

The randomness of the above sequence can be tested by comparing the observed number of times that two unlike events occur together with its expected value. Now two unlike events occur together 62 times in the given sequence. The expected value and the variance of this number are given in (9.2) and (9.6) in which $n_1, n_2, \cdots n_9$ are the number of times that the events A, B, C, D, E, F, G, H, and J occur in the sequence. For the above sequence $n_1, n_2, \cdots n_9$ are 9, 6, 7, 16, 7, 6, 8, 9, and 5 respectively and m is 73. The expected value and the variance are 63·73 and 6·93. They have been calculated by using the relations

$$2\,\Sigma n_r n_s = (\Sigma n_r)^2 - \Sigma n_r^2$$

$$3\,!\;\Sigma n_r n_s n_t = (\Sigma n_r)^3 - 3\,(\Sigma n_r^2)\,(\Sigma n_r) + 2\Sigma n_r^3,$$

$$4\,!\;\Sigma n_r n_s n_t n_u = (\Sigma n_r^4) + 8\,(\Sigma n_r)\,(\Sigma n_r^3) + 3\,(\Sigma n_r^2)^2 - 6\,\Sigma n_r^4 - 6\,(\Sigma n_r)^2\,(\Sigma n_r^2)$$

$$2\,\Sigma n_r^2 n_s^2 = (\Sigma n_r^2)^2 - \Sigma n_r^4.$$

The standarized deviate

$$x = \frac{62 - 63\cdot 73}{\sqrt{6\cdot 93}} = -\,0\cdot 67$$

is clearly not significant and gives little cause to doubt the randomness of the sequence.

12. Summary

The paper discusses certain probability distributions that arise from m points arranged on a line such that each of the points may possess any one of k characters or colours subject to the following conditions:—

(1) the chance of a point taking the rth colour is p_r such that $\sum_1^k p_r = 1$;

(2) a fixed number of points, say $n_1, n_2, \ldots n_k$, such that $\sum_1^k n_r =$ the total number of points on the line, are respectively black, white, red, etc.

The first of these is termed 'free sampling' and the second 'non-free sampling'.

The distributions considered are mainly the following:—

(1) the number of joins between adjacent points of the same colour;

(2) the number of joins between adjacent points of two specified colours; and

(3) the total number of joins between adjacent points belonging to different colours.

All these distributions tend to the normal form when p_r and n_r are fixed, and m tends to infinity. The distributions for (1) and (2) tend to the Poisson form when some of the p_r's and n_r's are small.

The results obtained in this paper have been used for testing the departure from randomness in the following cases:—

(*a*) distribution of diseased plants observed on a line;

(*b*) spread of disease in a given interval; and

(*c*) a given sequence of observations.

My thanks are due to Dr. D. J. Finney for guidance and for facilities in carrying out the investigations described in this paper.

References

Bortkiewicz, L. V. .. *Die Iteration*, Berlin, 1917.

Bruns, H. .. *Wahrscheinlichkeitsrechnung und Kollektivmasslehre*, Leipzig, 1906, 216.

Ising, E. .. "Beitrag zur theorie des Firromagnestimus," *Zeit. f. physik.*, 1925, **31**, 253.

Mahalanobis, P. C. .. "On large-scale sampling surveys," *Phil. Trans. Roy. Soc.*, London, 1944, **231**, 329.

Marbe, K. .. *Naturphilosophische Untersuchnungen zur Wahrschenlichkeitslehre*, Leipzig, 1899.

———— .. *Mathematische Bemerkungen*, Munchen, 1916.

Mises, R. V. .. "Das Problem der Iteration," *Ziet. f. angew. Math. u. Mech.*, 1921, **1**, 298.

Mood, A. M. .. "The distribution theory of runs," *Ann. Math. Stat.*, 1940, **11**, 367.

Stevens, W. L. .. "Distribution of groups in a sequence of alternatives," *Ann. Eug.*, 1939, **9**, 10.

Wald, A., and Wolfowitz, F. "On a test whether two samples are from the same population," *Ann. Math. Stat.*, 1940, **11**, 147.

Wishart, J., and Hirschfeld, H. O. "A theorem concerning the distribution of joins between line segments," *J. Math. Soc.*, London, 1936, **11**, 227.

Van der Plank, J. E. .. "A method for estimating the number of random groups of adjaeent diseased plants in a homogeneous field," *Trans. Roy. Soc.*, South Africa, 1946, **31**, 269.