A REVIEW OF ANALYSIS OF KARYOGRAPHS OF THE HUMAN CELL IN MITOSIS

D. E. BARTON*
UNIVERSITY COLLEGE, LONDON
F. N. DAVID**
UNIVERSITY COLLEGE, LONDON
and
UNIVERSITY OF CALIFORNIA, BERKELEY
EVELYN FIX**
UNIVERSITY OF CALIFORNIA, BERKELEY
and MAXINE MERRINGTON
UNIVERSITY COLLEGE, LONDON

1. Introduction

Since the publication of the first clear human karyograph by Tjio and Levan [13], much research, both theoretical and experimental, investigating the behavior of the chromosomes in the human cell during mitosis, has been carried out. Encouraged by L. S. Penrose and using karyographs made by him and by scientists working under him at the Galton Laboratory and, more recently, at the Kennedy-Galton Centre for Mental Retardation Research, Harperbury, we have carried out an intensive statistical study of the positions of the chromosomes as indicated by their centromeres. A typical karyograph is illustrated in figure 1.

It is the purpose of this present paper to present this study as a connected whole—it has previously been reported piecemeal as results were obtained—and to give such further work as has been done. Further, since such tests as we have devised may all be considered as variants of randomization tests in the plane, we advance here a way in which the randomization set might be weighted (or distorted), which leads to an approximate power function for the tests.

The experimental difficulties in the production and labeling of the karyograph are considerable. We may note that owing to the method of preparation used it appeared probable that any pattern in the chromosome centromeres would be largely destroyed and we ourselves were told at the beginning of our investigation that we might expect a completely random arrangement. We showed that

* Now at the Institute of Computer Sciences and Queen Mary College, University of London.
** With the partial support of the National Institutes of Health, USPHS Grant GM-10525.

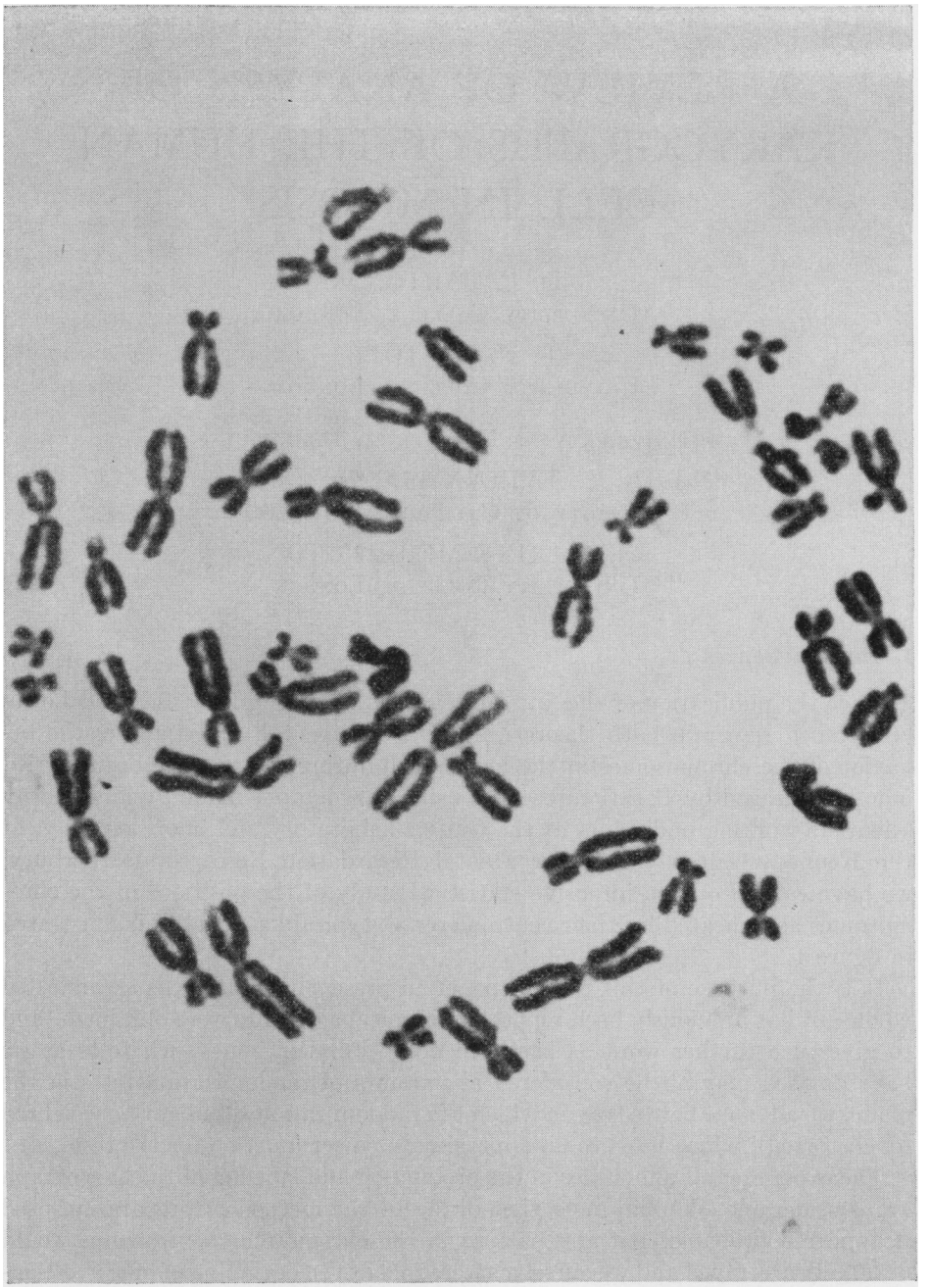

FIGURE 1

Karyograph of a normal female cell.

this random arrangement was not present in karyographs from normal women and men. We may also note the difficulty of distinguishing and numbering the different chromosome pairs. The cell is three dimensional while a karyograph is necessarily flat. It is probably true that in some cases the chromosomes lie nearly in the metaphase plate which is conceptually flat but randomly oriented to the photographic plate. For this and other reasons, the effect of projection may be considerable.

At the beginning of the development of the karyograph technique, cytogeneticists believed that it was possible to be entirely consistent in the assigning of the appropriate number to a chromosome. During recent months there has been a falling away from this position and now the pendulum has perhaps swung too far, and it is customary only to consider the chromosomes in certain rather arbitrary groups. Finally, from a statistical point of view we were faced with undoubted selection in the material chosen for photography. The cytologist, entirely reasonably, did not photograph those preparations which he could see under the microscope were not spread out with the chromosomes in a state where they had a chance of being identifiable. It was principally for these reasons, coupled with the fact that the cell wall is not visible in more than 50 per cent of our photographs, that we turned to the development of randomization tests in order to test the hypothesis of randomness and to describe the contiguity (or otherwise) of the various chromosome pairs. For the purposes of setting up the alternative hypothesis, it is necessary to describe the experimental techniques of preparation of the karyograph in more detail. This is done in section 6 where some discussion of the general cytobiological background is also given.

2. Statement of problem studied

The normal human cell has twenty three pairs of chromosomes: consisting of the autosomes labeled 1 up to 22, with 1 being the largest and 22 the smallest, and two sex chromosomes (a pair of X's for the female and an X and a Y for the male), that is, 46 chromosomes in all. The X chromosome is approximately the size of the number 6 and the Y (roughly) the size of the number 21 or 22. The abnormal cell often has more or less than this number. Thus, for example, mongols (that is, those with Down's syndrome) are known to have an extra chromosome of number 21 in addition to the normal complement, while Klinefelter's syndrome in males is associated with an extra X chromosome (that is, an XXY constitution).

Studying first normal cells, we began by investigating whether the distances between chromosomes of like number (homologous pairs) were such that they could be regarded as randomly placed each to the other. Sex chromosomes are treated as homologous even when, as in males, they differ.

We assume N positions in the plane (x_{it}, y_{it}), $i = 1, \cdots, r_t$, $t = 1, 2, \cdots, n$, with $\sum_{t=1}^{n} r_t = N$, which are almost certainly not randomly situated each to

the other. That is, they are not uniformly distributed; nor do they have any assignable distribution law. The null hypothesis to be tested is that the chromosome *numbers* are randomly placed over these space points and the alternative hypothesis that if they are not randomly placed then like numbers tend to be placed close together. Departure from the null hypothesis in the opposite sense is less likely but not ruled out.

Treating the ${}^{N}C_2$ distances between points as a randomization set, we first considered the criterion

$$D = \frac{\sum_{t=1}^{n} \frac{1}{r_t} \sum_{i=1}^{r_t} \sum_{j=1}^{r_t} [(x_{ti} - x_{tj})^2 + (y_{ti} - y_{tj})^2]}{\frac{1}{N} \sum_{t=1}^{n} \sum_{\ell=1}^{n} \sum_{i=1}^{r_t} \sum_{j=1}^{r_\ell} [(x_{ti} - x_{\ell j})^2 + (y_{ti} - y_{\ell j})^2]}. \tag{2.1}$$

If $\bar{x}$, $\bar{y}$ are the means over all N points and if we write

$$\begin{aligned} m_{ab} &= \frac{1}{N} \sum_{t=1}^{n} \sum_{j=1}^{r_t} (x_{tj} - \bar{x})^a (y_{tj} - \bar{y})^b \\ k_{1t} &= \frac{1}{r_t} \sum_{i=1}^{r_t} (x_{ti} - \bar{x}), \qquad k'_{1t} = \frac{1}{r_t} \sum_{i=1}^{r_t} (y_{ti} - \bar{y}), \end{aligned} \tag{2.2}$$

then D reduces to

$$D = 1 - \frac{1}{N(m_{20} + m_{02})} \sum_{t=1}^{n} r_t (k_{1t}^2 + k_{1t}'^2). \tag{2.3}$$

It is thus immaterial whether we consider D or D^* where

$$D^* = \frac{1}{N(m_{20} + m_{02})} \sum_{t=1}^{n} r_t (k_{1t}^2 + k_{1t}'^2). \tag{2.4}$$

The first form of D, when the r_t are all equal (or without the factor $1/r_t$, in general) is evidently a bivariate intraclass correlation coefficient. It was in this form that Professor Penrose introduced the statistic to our attention.

It will be noticed that both D and D^* are invariant under a rotation and translation of the axes of reference, an important point when it is remembered the axes of reference are arbitrarily chosen. We have

$$\begin{aligned} E(D) &= \frac{N - n}{N - 1}, \qquad 0 < D < 1, \\ \text{Var } (D) &= \frac{1}{N^{(4)}} \left\{ 2(n-1)(N-n) \left[(N-2)c_1 - 2c_2 + \frac{2}{(N-1)} \right] \right. \\ &\qquad \left. - R_1 N^{(2)} \left[1 + 2c_1 - c_2 \frac{N+1}{N-1} \right] \right\}, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} c_1 &= \frac{m_{20}^2 + 2m_{11}^2 + m_{02}^2}{(m_{20} + m_{02})^2}, \qquad c_2 = \frac{m_{40} + 2m_{22} + m_{04}}{(m_{20} + m_{02})^2}, \\ R_1 &= \sum_{t=1}^{n} \left(\frac{1}{r_t} - \frac{n}{N} \right), \end{aligned} \tag{2.6}$$

which latter is zero for the normal cell, since $r_t = 2$ for $t = 1, \cdots, n$.

We calculated the third and fourth moments of D using bivariate symmetric function tables constructed by David and Fix ([10], pp. 144–177) and also constructed the exact distribution of D for some small values of N. On the basis of these investigations we concluded (Barton and David [3]) that the distribution of D could be accurately described by a beta distribution (a Pearson Type I curve), but that the assumption of a normal distribution was unlikely to lead to serious error.

3. Technique employed

The effect of projection may be neutralized by rotation to principal axes and then scaling by the dispersions in these directions. This is merely a transformation of the randomization set; any tendency of pairs to be too close or too widely separated will be only marginally affected. The D statistic based on these circularized points when written in terms of the actual measured coordinates becomes (say)

$$(3.1)\quad T^* = \frac{N-1}{N-n}\left[1 - \frac{1}{2N(m_{20}m_{02} - m_{11}^2)}\sum_{i=1}^{n} r_i(m_{02}k_{1i}^2 - 2m_{11}k_{1i}k'_{1i} + m_{20}k_{1i}'^2)\right]$$

when scaled so that $E(T^*) = 1$. We have, further,

$$(3.2)\quad \operatorname{Var} T^* = \frac{(N-1)^2}{N^{(4)}(N-n)^2}\left[\frac{b}{4}N(N+1)R_1 - 2(n-1)(N-n) - 2N^{(2)}R_1 + \frac{(n-1)}{N-1}(N-n)(N^2-3N+4)\right]$$

where

$$(3.3)\quad b = (m_{40}m_{02}^2 - 4m_{31}m_{11}m_{02} + 4m_{22}m_{11}^2 + 2m_{22}m_{20}m_{02} - 4m_{13}m_{20}m_{11} + m_{04}m_{20}^2)/(m_{20}m_{02} - m_{11}^2)^2.$$

Multiplying by the appropriate factor so that the reduced, standardized statistic has a mean zero and a standard deviation of unity we obtained table I. The number of cells in the category is k.

The expressions $m_{20} + m_{02}$, $m_{40} + 2m_{22} + m_{04}$, $m_{20}^2 + 2m_{11}^2 + m_{02}^2$, and so forth (which are invariant to rotation of axes) were systematically studied by us and were given in the appendix to Barton and David [2].

A more statistical way of regarding this circularization process is to say that the axes of reference have been rotated so that the correlation between x and y is zero and these have then been scaled to new variables according to the standard deviations along the new axes of reference. Thus, we write

$$(3.4)\quad X_i = \frac{x_i - \bar{x}}{(m_{20})^{1/2}}, \qquad Y_i = \frac{m_{20}(y_i - \bar{y}) - m_{11}(x_i - \bar{x})}{[m_{20}(m_{20}m_{02} - m_{11}^2)]^{1/2}}.$$

The D criterion as function of the coordinates referred to the new axes of reference is

TABLE I

$(T^* - 1)\ (\text{Var}\ T^*)^{-1/2}$ FOR 238 NORMAL AND ABNORMAL KARYOGRAPHS

77 Normal Female Cells				74 Normal Male Cells				47 Abnormal Female Cells		40 Abnormal Male Cells	
Code		Code		Code		Code		Code		Code	
1	−1.9618	158	−1.8756	10	+0.4860	160	+0.2745	2	−1.4267	4	−0.9075
11	+0.3192	159	−0.4458	12	−0.2690	161	−0.2025	3	+0.7454	8	+0.4258
17	−0.8653	162	−0.3932	13	+0.5367	178	+1.7953	5	+0.4993	9	−0.4488
20	−2.0843	163	−2.3190	14	−0.0707	179	+1.1997	6	+0.1597	24	+0.5257
21	−1.5926	164	+0.1073	15	−0.0175	203	−0.0183	7	+0.2618	38	+2.1290
22	+0.3005	165	−1.0141	16	−0.9441	204	−1.2817	19	+1.0039	39	−1.4463
23	−0.4693	166	+0.0908	18	+0.1408	205	−1.6522	25	−1.2076	40	−0.5558
28	−1.2893	167	+0.5929	27	−2.7519	213	−0.0505	26	−0.9692	44	−0.9171
29	−1.8433	173	−1.4622	30	+1.3893	214	−1.6439	43	−1.6018	54	−1.3083
33	−1.0153	174	−0.4338	31	+0.7701	215	−1.3729	49	+0.9594	56	+1.4474
34	−0.2036	175	+1.3328	32	−1.4735	216	−2.0246	50	−0.4290	57	−0.7628
36	−1.5826	200	−0.9937	35	−0.2849	219	−2.2444	66	−0.8881	62	−0.1602
37	−0.7569	201	−1.4573	41	+0.9177	220	+0.5184	67	−0.0245	63	−1.1432
42	+0.2832	206	+1.6363	45	−1.6109	224	−0.0233	68	+1.1768	76	−0.2881
46	−0.5272	207	+0.1150	47	+1.4827	225	+0.8671	69	+0.1046	78	−0.9766
59	−0.0159	208	+2.2509	48	+0.4623	226	−1.0047	70	−0.7767	79	−0.3280
61	+0.3963	209	−1.0725	51	−0.6643	227	+1.1671	71	−0.8768	80	−0.6171
64	−2.1510	210	−0.1741	52	−0.8230	232	−0.1749	75	−1.1973	81	+0.1816
73	+1.1966	211	−1.8112	53	+0.4587	233	+1.0080	82	+1.6016	87	−0.1976
94	+1.4268	212	+0.5708	55	−0.0307	234	−2.2729	83	−0.7330	88	−0.5275
97	−1.5100	217	+2.3809	58	−1.5104	235	−1.5638	84	+1.7131	90	−1.1616
102	−0.6161	218	−0.3081	60	−1.1861	236	−1.7982	85	−0.7294	92	+0.5446
103	−2.2603	239	−2.1236	65	+0.2038	237	−0.2235	95	−1.6512	93	−0.1444
104	+0.2117	243	−1.1103	72	−0.4653	238	−0.6077	96	+1.3457	113	+0.8991
106	+1.6426	244	−0.7194	74	−0.8584	240	+1.0691	98	+0.2082	114	−0.7246
109	−0.4351	245	+0.1211	77	+1.2567	241	+2.0613	112	−1.0957	131	+1.8490
117	+0.3338	246	−0.3715	86	−2.4942	251	+0.0701	120	−2.0692	132	+0.2664
118	−0.1687	254	−0.1509	89	−1.5583	252	+0.3616	126	−0.6806	147	+0.6489
119	+0.2528	255	−1.0557	99	−0.1287	253	−0.9538	127	−1.7727	149	−0.3573
121	+0.2177	256	+1.7822	100	−0.9882	257	+1.9427	128	−0.1088	153	+0.6197
122	−2.5744	264	+0.1252	101	−2.2747	258	+0.0529	129	+0.2290	176	−1.0539
123	−0.3903	265	+0.1741	105	−0.5800	259	+0.2076	134	−0.0982	177	−0.2137
124	−1.9367	266	+0.5990	110	−0.4025			141	−1.6879	202	−0.2429
125	−0.8860			115	+0.0318			146	+0.9219	221	−0.4357
130	+1.4864			116	−1.4897			154	−0.4832	222	+1.7252
137	−0.9367			133	−1.2209			155	+0.2192	223	+0.0952
138	−1.6273			135	+0.5886			156	+0.4315	228	−0.8655
139	−2.0620			136	−1.1947			168	+0.2758	229	−0.5500
140	+0.9059			143	−0.1624			172	−0.5745	230	−0.5440
142	+0.5695			144	+1.3886			247	−0.0670	231	+0.6395
145	−0.5740			150	+0.7547			248	−0.2820		
148	+0.5595			157	−1.6019			249	−0.3220		
151	−0.6418							250	−0.9368		
152	+0.6121							260	−2.0542		
								261	−2.3054		
								262	+0.4941		
								263	−0.0210		
$\Sigma(T^* - 1)(k\ \text{Var}\ T^*)^{-1/2} = -3.3819$				−2.6396				−2.1471		−0.7718	

$$T = \frac{1}{4N} \sum_{i=1}^{n} r_i^{-1} \sum_{i=1}^{r_i} \sum_{j=1}^{r_i} [(X_{ti} - X_{tj})^2 + (Y_{ti} - Y_{tj})^2] \tag{3.5}$$

$$\equiv \frac{N - n}{N - 1} T^*$$

with

$$E(T) = \frac{N - n}{N - 1}$$

$$\text{(3.6)}\qquad \operatorname{Var} T = \frac{1}{N^{(4)}}\left\{(M_{40} + 2M_{22} + M_{04})[N(N+1)R_1 - 2(n-1)(N-n)] - 2N^{(2)}R_1 + \frac{(n-1)(N-n)(N^2-3N+4)}{N-1}\right\}$$

where

$$\text{(3.7)}\qquad M_{ab} = \frac{1}{N}\sum_{t,i}(X_{ti} - \bar{X})^a(Y_{ti} - \bar{Y})^b.$$

Further details of the application of these techniques are given in [4] and [5].

Merrington and Penrose [11] used this technique to test the validity of chromosome associations which had been asserted to exist, but which had not been verified statistically. It has been asserted that: (1) the six chromosomes of 13, 14, 15 tend to lie close together; (2) the four chromosomes 21 and 22 tend to lie close together; (3) the pair of chromosomes 1, 2, and 3, each pair taken separately tend to lie close together; (4) there is an association between the pair of the 1 chromosomes and the four chromosomes 21 and 22. The sum of all the squared distances, $c_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2$, between pairs (x_i, y_i), (x_j, y_j) of members of a given group of r chromosomes were averaged over the ${}^rC_2$ pairs of members of the group and the statistic

$$\text{(3.8)}\qquad \bar{c} = \frac{1}{{}^rC_2}\sum_{i<j} c_{ij}$$

was computed. This was done for 62 cells, a mixed sample of normals and abnormals of both sexes. Since $E(\bar{c}) = 1$ under the null hypothesis, $\bar{c} - 1$ was compared with its empirical standard error $s/(62)^{1/2}$ where s^2 is the observed value of the variance of $\bar{c}$ among 62 cells. These results are reported in table II.

TABLE II

MEAN STANDARDIZED SQUARED DISTANCES BETWEEN CENTROMERES AVERAGED OVER SIXTY TWO CELLS

Alternative Hypotheses	Chromosomes Examined	Number of Cells	Number of (Distances)^2	Average of Cell Means	S.D. s	Difference from Expected Value	S.E. of Difference	Difference / S.E.
1	13, 14, 15	62	15	0.852	0.310	−0.148	0.039	−3.8
2	21, 22	62	6 or 10	0.866	0.477	−0.134	0.061	−2.2
3(i)	1	62	1	0.925	0.881	−0.075	0.112	−0.7
3(ii)	2	62	1	1.083	0.891	+0.083	0.113	+0.7
3(iii)	3	62	1	0.954	0.737	−0.046	0.094	−0.5
4	1 and 21 or 22	62	8 or 10	0.910	0.430	−0.090	0.055	−1.6
4	All homologous pairs or groups	62	—	0.939	0.145	−0.061	0.018	−3.3

Merrington and Penrose concluded that there was strong evidence that the large acrocentrics (numbers 13, 14, 15) lie closer together than would be expected on the null hypothesis and that the same is true for small acrocentrics (21, 22).

4. Possibility of outlying chromosomes

Following the work of Merrington and Penrose, further suggestions appeared in the scientific press (for example, Mittwoch [12]) to the effect that some types of chromosome tended to lie to the outside of the karyograph. We investigated this [6] by calculating the radial distance of each chromosome (after having first circularized as above) and then ranking the N distances in order of magnitude, the largest distance being given the lowest rank of one. Again, we should emphasize that the radial distances will (almost certainly) not form a random set, but we are interested only in the randomness of the chromosome numbers attached to the radial points. Let there be a group of m chromosomes of the type we are interested in, supposedly indistinguishable each from the other. Then the joint distribution of the rth and sth of these (say z_r, z_s, with $r < s$) will be

$$P\{\text{rank } z_r = u, \text{ rank } z_s = v; r < s, u < v\} = {}^{u-1}C_{r-1}\,{}^{v-u-1}C_{s-r-1}\,{}^{N-v}C_{m-s}\,{}^{N}C_{m}. \tag{4.1}$$

The mean values in repeated samples, the variances and covariances, may be obtained from the general formula

$$E[(u+k)^{(k+1)}(N-v+\ell)^{(\ell)}] = \frac{(m-s+\ell)^{(\ell)}(r+k)^{(k+1)}(N+k+\ell+1)^{(k+\ell+1)}}{(m+k+\ell+1)^{(k+\ell+1)}}. \tag{4.2}$$

Thus,

$$\begin{aligned} E(u) &= \frac{r(N+1)}{m+1}, \qquad E(v) = \frac{s(N+1)}{m+1}, \\ \operatorname{Var} u &= \frac{r(N+1)(N-m)(m-r+1)}{(m+1)^2(m+2)}, \\ \operatorname{Var} v &= \frac{s(N+1)(N-m)(m-s+1)}{(m+1)^2(m+2)}, \\ \operatorname{Corr}(u, v) &= \left[\frac{r(m-s+1)}{s(m-r+1)}\right]^{1/2}. \end{aligned} \tag{4.3}$$

The independence of N exhibited by the correlation is noteworthy. It suggests what is in fact the case, namely, that there is an extremely close continuous variable approximation to the distribution of u and v. Thus, when N is large (as here) and when m is not too big (also, as here), then u/N, v/N are, to an extremely high order, approximated by the rth and sth members of an ordered sample of m rectangular variables. Thus, if $x = u/N$ and $y = v/N$, we have the approximating distribution with p.d.f.

$$p(x, y) = \frac{m!}{(r-1)!(s-r-1)!(m-s)!}\, x^{r-1}(y-x)^{s-r-1}(1-y)^{m-s}, \qquad 0 < x < y < 1. \tag{4.4}$$

This is indeed the exact limiting distribution when $N \to \infty$.

This radial analysis was carried out [7] for 114 karyographs of six types

TABLE III

TABLE OF MEAN RANKS (OVER k DIFFERENT KARYOGRAPHS) OF THE INNER AND OUTER CHROMOSOMES OF EACH PAIR, FOR SIX DIFFERENT CATEGORIES [7]

Chromosome		Category I	II	III	IV	V	VI
Outer	1	17.2	16.4	15.6	15.4	16.8	13.3
Inner	1	36.0	33.8	29.0	29.3	33.0	31.0
Outer	2	13.6	17.1	17.4	18.8	12.8	15.4
Inner	2	26.6	28.6	29.3	36.0	26.4	30.6
Outer	3	14.8	13.9	13.2	16.4	8.4	13.5
Inner	3	31.1	29.1	28.0	31.2	30.2	30.9
Outer	4	16.3	14.3	10.8	13.3	15.2	15.0
Inner	4	28.3	27.6	25.2	28.8	39.2	28.4
Outer	5	18.7	18.4	21.4	15.3	12.8	16.5
Inner	5	32.3	32.6	31.2	30.7	26.0	30.7
Outer	6	12.0	14.4	12.0	17.6	10.4	13.3
Inner	6	31.4	25.7	25.1	27.6	29.0	30.0
Outer	7	14.1	11.0	12.7	18.8	7.6	14.6
Inner	7	30.0	22.4	30.5	32.4	19.2	27.8
Outer	8	15.6	13.1	18.4	14.2	17.6	14.7
Inner	8	31.4	33.8	30.4	26.9	37.0	36.5
Outer	9	13.2	11.8	12.5	12.2	8.2	12.2
Inner	9	32.0	29.1	30.3	28.8	18.2	28.4
Outer	10	13.1	16.8	17.7	15.7	18.2	14.0
Inner	10	29.5	31.3	34.1	31.2	35.6	31.8
Outer	11	14.8	14.0	18.3	16.9	17.4	17.8
Inner	11	28.4	32.5	32.2	32.7	33.0	34.3
Outer	12	17.1	19.6	16.1	15.4	22.0	20.5
Inner	12	32.5	35.3	30.2	36.2	41.6	37.1
Outer	13	16.6	13.6	19.6	19.3	25.4	15.1
Inner	13	31.4	31.9	35.0	34.4	34.8	33.5
Outer	14	16.9	17.2	17.8	17.0	25.0	13.9
Inner	14	29.8	33.9	29.5	30.1	35.8	27.7
Outer	15	19.1	22.3	19.4	16.7	21.6	15.8
Inner	15	33.5	37.3	35.1	31.8	38.2	35.3
Outer	16	17.3	16.7	16.2	14.9	15.4	21.4
Inner	16	32.8	33.4	30.5	27.7	29.2	35.3
Outer	17	17.0	18.5	14.0	14.5	10.0	13.6
Inner	17	28.8	32.6	31.3	33.7	31.4	29.7
Outer	18	14.8	14.5	14.7	13.7	14.6	19.8
Inner	18	30.3	32.4	33.8	31.3	28.8	35.6
Outer	19	13.1	10.4	14.6	13.6	12.0	20.0
Inner	19	33.6	27.0	28.5	28.8	27.0	33.6
Outer	20	19.7	22.2	20.0	16.2	15.8	16.3
Inner	20	32.6	34.3	34.0	33.6	33.8	32.7
Outer	21	16.5	18.8	17.8	16.4	22.6	11.9
Central	21	—	—	—	—	38.0	24.0
Inner	21	31.5	31.0	33.7	32.8	41.4	34.7
Outer	22	21.1	18.6	19.8	18.1	13.4	21.1
Inner	22	33.7	31.8	32.2	33.7	36.2	33.5
Outer	X	11.3	X17.6	X13.2	X20.6	X13.4	X20.3
Inner	Y	29.9	Y22.3	X29.9	Y20.8	X28.4	Y25.5
	k	24	19	20	26	5	20

(categories), for a variety of chromosome pairs or groups. The values of u and v were averaged over the karyographs in each group. Denoting the number in each category by k, we found the figures given in table III. In the table $m = 2$ and $r = 1$, $s = 2$ and so we refer to u and v as "inner" and "outer" ranks.

Assessing the significance of the pairs of mean ranks for any chromosome number is best effected from the percentage points of the joint distribution of $(\bar{x}, \bar{y})$, where $(\bar{x}, \bar{y})$ are the means of k independent variables, each distributed as (x, y) above. This is a hitherto unsolved problem in distribution theory, and because of the extensive mathematical derivation required, we have reserved further discussion of this for another occasion. For present purposes, it is sufficient to use the normal approximation for the joint distribution of $(\bar{x}, \bar{y})$, which is very good for k of the order of 20, as in five out of the six categories here.

This method of analysis treats the different chromosomes of a given number distinctly. It thus permits observation of any differential behavior which may occur, either because one of the chromosomes has been wrongly labeled or because for convenience, different chromosomes (as the X and Y or the large acrocentrics) have been grouped together.

A collective test is provided by simply averaging the radial ranks of a given group. Thus, if the ranks are $u_1, \cdots, u_m$ and $\bar{u} = (u_1 + \cdots + u_m)/m$, we have a simple Wilcoxon statistic. The statistics $\bar{u}$ for different chromosome groups were averaged over the k cells in each category and the reduced standardized deviate tabled. This summary table is given in table IV.

TABLE IV

SUMMARY OF REDUCED STANDARDIZED DEVIATES

Group	Chromosomes in Group	No. m	I	II	III	IV	V	VI
A	1	2	+1.616	+0.766	−0.590	−0.644	+0.212	−0.872
	2	2	−1.792	−0.297	−0.096	+2.144	−1.037	−0.483
	3	2	−0.275	−0.927	−1.409	+0.158	−1.108	−0.860
B	4, 5	4	+0.286	−0.196	−0.959	−1.162	−0.239	−0.912
C	6, X	4(3)	−1.822	−2.495	−2.423	−1.064	−1.262	−1.635
	7, 8, 11	6	−1.079	−2.020	+0.182	+0.135	−0.870	+0.228
	9, 10, 12	6	−0.559	+0.404	−0.029	−0.326	−0.014	−0.021
D	13, 14, 15	6	+1.012	+2.170	+2.225	+1.382	+2.623	−0.399
E	16	2	+0.803	+0.717	−0.096	−1.204	−0.401	+2.051
	17, 18	4	−0.581	+0.689	−0.052	−0.161	−0.955	+0.452
F	19, 20	4	+0.947	−0.018	+0.532	−0.367	−0.631	+1.108
G	21, 22	4(5)	+1.663	+1.064	+1.647	+1.399	+2.438	+0.795
	Y	1		−0.389		−1.019		+0.495
k, number of cells			24	19	20	26	5	20

In the circumstances envisaged here where m is small but N large, it is clear that $\bar{u}/N$ behaves to a high degree of approximation, like the mean of m independent variates each distributed uniformly in (0, 1). The statistic used (that

is, the mean of k independent values of $\bar{u}$) is thus one whose distribution is given by Barton and David [1]. For values of k at all large the normal approximation is quite adequate.

5. Like pairs of chromosomes

There are many other types of nonrandomness which might be tested, but on the whole it has seemed most profitable to confine ourselves to confirming (or otherwise) statements made by the cytologists. Enough statistical evidence is gradually becoming available to enable by an informed guess, possibly, to see more of the way in which the chromosomes were arranged before the pattern was almost destroyed by the method of karyograph preparation, but it is probably better to wait for the cytological theory before developing too wide a variety of tests. There is, however, another simple test for position which is quickly carried out. It has been suggested that like pairs of chromosomes lie in the same sectors of the karyograph.

Confining ourselves to cells with either pairs or isolated chromosomes and assuming that it is possible for the chromosomes to have been accurately labeled, let us choose an arbitrary axis of reference through the centroid and assign ranks to the chromosomes according to the angle which is made (by the radius from the chromosome centromeres) with the arbitrary axis of reference. If R_i is the rank of the first chromosome of the ith pair and R_i' that of the second, let

$$t_i = \min \{R_i' - R_i, N - R_i' + R_i\}. \tag{5.1}$$

Then we have

$$E(t_i) = \frac{N+1}{4} + \frac{1}{4(N-1)}, \qquad \operatorname{Var} t_i = \frac{N^2 - 2N + 3}{48} - \frac{1}{16(N-1)^2}, \quad N \text{ even.} \tag{5.2}$$

$$E(t_i) = \frac{N+1}{4}, \qquad \operatorname{Var} t_i = \frac{N^2 - 2N - 3}{48}, \quad N \text{ odd.}$$

Putting $\bar{t} = (1/n) \sum_{i=1}^{n} t_i$, a frequency table of $\bar{t}$ for ten normal male and ten normal female karyographs is given in table VI. The X and Y chromosomes in the male are counted as a pair. In table V are given values of $\bar{t}_i$, $i = 1, \cdots, 23$ for the ten males and ten females where $\bar{t}_i$ denotes the average value of t_i over a series of cells.

Restricting consideration to the mathematically simplest case, where there are n pairs of like chromosomes, so that $N = 2n$ (and $n = 23$ for the normal cell), two sorts of test function suggest themselves analogous to the two tests of the previous section.

The first is the overall test $\bar{t}$, for heterogeneity of angular ranks. We have

$$E(\bar{t}) = \frac{N^2}{4(N-1)}, \qquad \operatorname{Var} \bar{t} = \frac{1}{12n} \frac{N(N-2)^2(N^2 - 2N + 4)}{(N-1)^2(N-3)}, \tag{5.3}$$

TABLE V

TABLE OF l_i FOR TEN NORMAL FEMALES AND TEN NORMAL MALES WITH REDUCED DEVIATE $[l_i - E(t_i)]/[\text{Var } t_i/10]^{1/2}$

$$E(t_i) = \frac{N^2}{4(N-1)} = 11.7556, \text{ Var } t_i = \frac{N(N-2)(N^2-2N+4)}{48(N-1)^2} = 42.2291$$

Chromosome Pair	Female Karyographs		Male Karyographs	
	l_i	Reduced Deviate	l_i	Reduced Deviate
1	10.0	−0.854	10.3	−0.708
2	7.9	−1.876	9.3	−1.195
3	10.7	−0.514	10.5	−0.611
4	10.3	−0.708	13.0	+0.606
5	9.7	−1.000	11.3	−0.222
6	8.7	−1.487	12.8	+0.508
7	11.6	−0.076	8.1	−1.779
8	7.6	−2.022	12.5	+0.362
9	13.4	+0.800	12.7	+0.460
10	10.6	−0.562	10.3	−0.708
11	13.8	+0.995	9.9	−0.903
12	12.4	+0.314	8.3	−1.682
13	9.4	−1.146	9.7	−1.000
14	10.0	−0.854	6.5	−2.557
15	8.2	−1.730	9.9	−0.903
16	8.8	−1.438	9.3	−1.195
17	8.3	−1.682	8.3	−1.682
18	9.5	−1.098	13.1	+0.654
19	10.2	−0.757	13.1	+0.654
20	10.0	−0.854	10.7	−0.514
21	11.9	+0.070	11.0	−0.368
22	11.2	−0.270	11.5	−0.124
XX or *XY*	14.2	+1.190	8.2	−1.730

TABLE VI

TABLE OF $\bar{l}$, $w = \min t_i$, $w' = \max t_i$ FOR TEN NORMAL FEMALES AND TEN NORMAL MALES

Code No. of Cell	Normal Females			Code No. of Cell	Normal Males		
	$\bar{l}$	w'	w		$\bar{l}$	w'	w
20	9.1739	20	1	14	12.0435	23	1
21	9.2609	20	1	18	12.0870	22	1
33	10.1304	21	1	52	10.4783	22	1
64	9.4348	23	1	72	11.0870	22	1
106	13.6957	23	5	99	11.0870	22	1
117	11.3478	22	1	101	8.2174	22	1
122	8.3043	22	1	157	9.1739	21	1
164	11.6087	22	1	205	9.5217	22	2
210	11.4348	23	2	216	9.0870	23	1
124	9.2609	22	1	224	1.2609	23	2

since, as may be seen from elementary combinatorial arguments

$$\text{Corr}\,(t_i, t_j) = 2/(N-2)(N-3). \tag{5.4}$$

It seems safe, on general heuristic grounds, to use a normal approximation for the distribution of $\bar{l}$. (Indeed it is clear that $\{t_i/n\}$ behave very closely as a set of independent uniform variables.) Enumeration of the probabilities for small N confirm this; for example, $N = 8$ gives for the p.d.f. $p(\bar{l})$ the values

$\bar{l}$	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3	$3\frac{1}{2}$	4
$105\ p(\bar{l})$	2	16	32	36	14	4	1

To test whether any particular pair are significantly close in angular measure (or significantly far apart), the statistics $w = \min\{t_i\}$ (or $w' = \max\{t_i\}$) suggest themselves.

Looking first at w, the heuristic approximation given by the smallest of n uniform variables suggests that its p.d.f. will be approximately $p(w) \cong e^{-w}(e-1)$, $w = 1, 2, \cdots$, and this in turn suggests that it will not be possible to test for small values of w at any of the usual levels of significance. The exact distribution for $N = 8$ is

w	1	2	3	4
$105\ p(w)$	74	24	6	1

and confirms the suggestion of a large probability, about e^{-1}, that $w = 1$ is its minimum value.

The precise value for $P\{w = 1\}$ follows from elementary combinatorial argument; it is given by

$$1 - p(1) = \sum_{j=0}^{n} {}^{n}C_j(-1)^j 2^j/(N-1)^{(j)} \tag{5.5}$$

which has integral representation

$$\frac{1}{(N-1)!}\int_0^\infty x^{n-1}(x-2)^n e^{-x}\,dx. \tag{5.6}$$

This tends to e^{-1} as $n \to \infty$, thus verifying the suggested form. Indeed, we have the asymptotic formula

$$p(1) = 1 - e^{-1}\left\{1 - \frac{3}{N} - \frac{3}{2N^2} + O\left(\frac{1}{N^3}\right)\right\}. \tag{5.7}$$

To assess significance in the lower tail of w, we study the statistic g, the number of the $\{t_i\}$ which take the value of 1. We have

$$p(g) = {}^{n}C_g \sum_{j=0}^{n-g} {}^{n-g}C_j(-1)^j 2^{j+g}/(N-1)^{(j+g)}, \qquad g = 0, 1, \cdots, n \tag{5.8}$$

$$\sim e^{-1}/g!, \qquad n \text{ large},$$

so that we may compare g with the upper percentage point of a Poisson variable with expectation unity.

Similar considerations apply to the distribution of w' which has a chance of attaining its maximum possible value of n which tends to $1 - e^{-1/2}$ for large N.

This section presents the results of a preliminary study of statistics arising from the joint distribution of the circular rank differences contingent on the $N!/2^n n!$ arrangements of n pairs on a circle. We hope to pursue this new combinatorial problem in a later paper. Tables V and VI give some numerical results.

6. A contraction model for the alternative hypothesis

Before setting up the model it is useful to pursue the experimental and cytological background to the karyograph preparation in more detail. The picture of this given here is as it appears to the statistician: we merely describe, tentatively, the biological framework for the statistical formulation of the alternative hypothesis based on our understanding of the situation as presented to us by our medical geneticist colleagues. The picture is not intended to be definitive, or complete, from the cytogenetic standpoints.

The preparation of the karyograph may be described as follows. A culture of skin or blood cells is first treated with colchicine. This is thought to inhibit spindle formation, with its consequent arraying of chromosomes in the metaphase plate (recently this step has been omitted by some cytologists without apparently affecting the issue, though it had previously been deemed essential). Second, the cells are "blown up," osmotically, to about twice their previous size. Third, the cells are killed by acid treatment. This acts by precipitating their protein, converting the fluid protoplasm to a firm gel. Fourth, the jellied culture is squashed flat, either by mechanical pressure or, more usually nowadays, by desiccation. Fifth, the resulting single layer of flattened cells is stained and the small proportion of cells found to have been in the metaphase state of mitosis are microphotographed. The resulting photograph is the karyograph of about 15 cm diameter.

The nuclear membrane appears less sharp in prophase and is thought to disappear by metaphase but the nucleus retains its ellipsoidal shape. The physical squashing or desiccation is much more distorting, particularly in that proportion of cases where the cell wall is actually ruptured by the process, and in many cases it may be the major cause of "projection." It is surprising that any residue of pattern could be thought to survive this process of preparation. However, there are certain cells showing the phenomenon of "endoreduplication" which have 92 chromosomes (two chemically identical chromosomes for every one in the normal cell) and in the karyographs of these the pairs of "sister" chromosomes lie side by side and indeed closely parallel. For example, the value of T^* on a typical such karyograph, treating it as a cell with 46 pairs of sisters, and not assorting the homologues, was 0.0047. This is minute compared with

a null hypothesis mean of 1 and a standard error of 0.1. A karyograph with endoreduplication is shown in figure 2.

Some further blurring of the picture is provided by misidentification of members of the homologous pairs. This misidentification has been possibly overemphasized recently. The degree to which it persists clearly depends on the quality of the karyograph preparation and the care with which the identifications are made. It seems probable that the karyographs with which we have dealt have been particularly well favored in respect of these factors and that this has colored our view of the risk of misidentification. Certainly, it is necessary only to use statistics which are robust in the face of this possibility. Equally, it may well be that more information is obtained from the greater number of karyograph measurements obtainable for the same effect by use of less careful mass production techniques.

Studies are at present in progress to check up on the validity of identification procedures, and it is to be hoped that these will clarify matters considerably.

Two general lines of biological reasoning bear on the type of alternative hypothesis envisaged. The first is the pattern of behavior in species other than men. In the Diptera (for example, *Drosophila* spp), there is "somatic pairing" and homologous chromosomes almost invariably behave like the sister chromosomes in endoreduplicated cells. In the Urodeles the chromosomes lie at the circumference of the metaphase plate. In the reptiles the large chromosomes lie to the outside of the metaphase plate with the smaller ones in the middle. This is said to be a general characteristic when there is considerable diversity in chromosome size (see White [14]). All these effects tend to group the homologous pairs relatively nearer together (if anything). A contrary effect could be associated with the phenomenon known as "affinity" which has been observed in certain inbred strains of mice where in anaphase the paternal chromosomes tend to separate toward one centrosome and the maternal to the other.

In men it is thought that the *sex chromatin bodies* may be associated with a slightly out of phase behavior of the X chromosome (or chromosomes). Thus, there are analogues which suggest models of general clustering of homologues, of groupwise clustering, of clustering of particular chromosomes, and the opposite of clustering. This justifies the variety of tests proposed. The model we discuss below is appropriate to general clustering and its opposite (that is, dissociation of homologues).

The second set of ideas is specifically human and they are contingent on what is believed about the human ovum. It is thought that her ovaries are laid down in a woman whilst she herself is an embryo (fetus) and that these consist of cells which have undergone the first meiotic division, but that the second is suspended. When she is adult, it is thought that each month one cell undergoes the second stage of meiosis and that of the two daughter gametes, one descends to the uterus as the ovum and the other degenerates. An important class of congenital defects (such as the Down's, Turner's, and Klinefelter's syndromes discussed earlier) are caused by the ovum having an irregular complement of

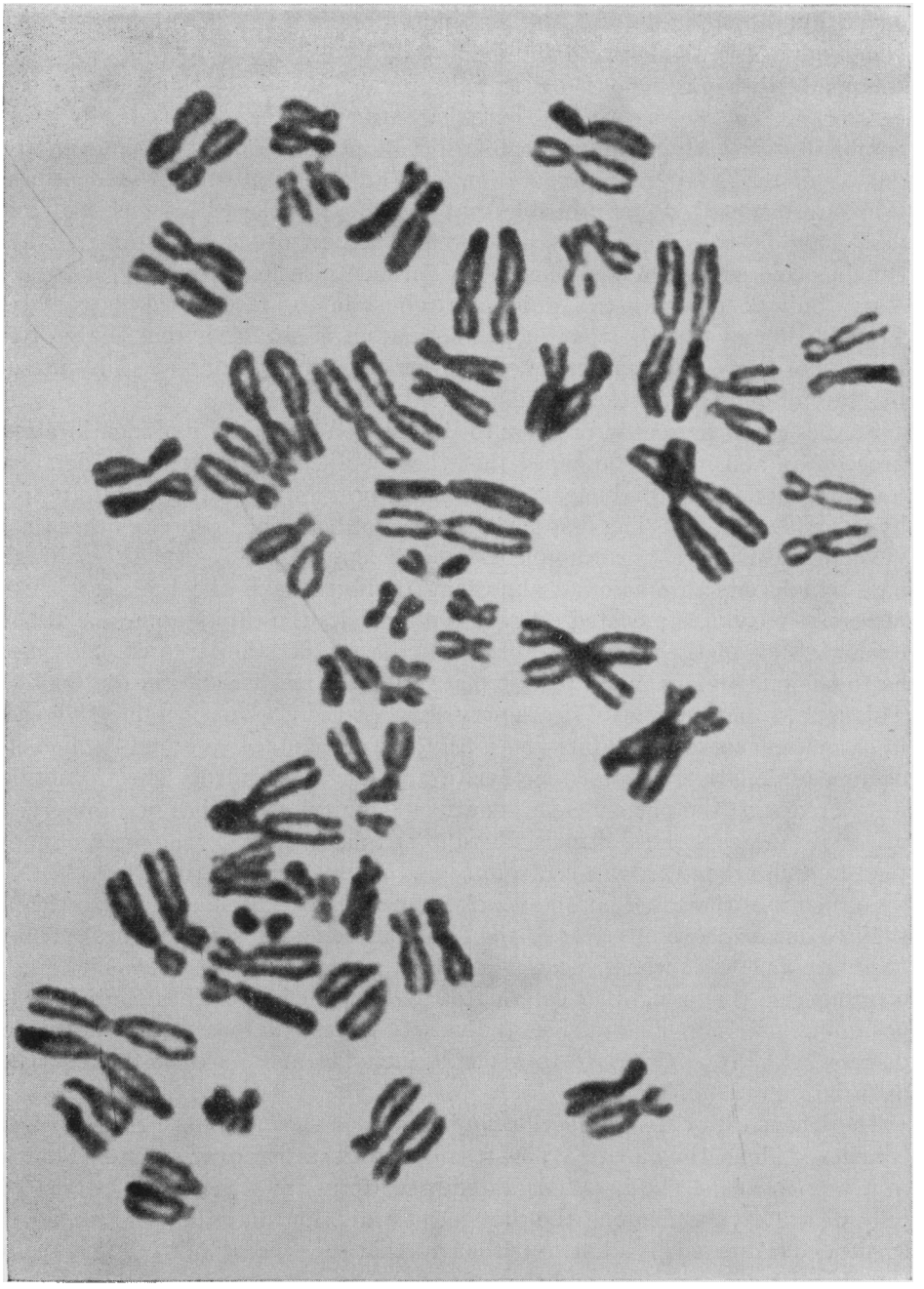

FIGURE 2
Karyograph showing endoreduplication

chromosomes, particularly an extra one or more. This may originate in an irregular meiosis at either stage. There is evidence that a tendency to such irregularity may run in families and it is conceivable that such tendencies could be reflected in a tendency for homologous pairs to associate to some degree in mitotic divisions of abnormal relatives of defectives or that a tendency to such association would be found in the mitotic division of cells of defectives. This is part of the background to the medical interest in human mitosis and helps to determine the sort of effect it would be important to detect.

The model we propose here supposes a random arrangement of chromosomes in metaphase followed by a "contraction" of homologous chromosomes toward their centroid which reduces the distance of these from their centroid by a uniform proportion ρ. For $\rho < 1$, this is a contraction but we also conceive that ρ may exceed 1 giving a "negative contraction," that is, an expansion or dissociation. This is not to be thought of as a physical model, though it could be, but as an empirical description of a graduatory nature. It provides a guide to the number of karyographs it is necessary to examine in order to give a reasonable power for testing for a given degree of contraction.

The model is a very simple one; it has the effect on T^* of multiplying its numerator by ρ^2 and multiplying its denominator by a lesser factor. In the case where there are $n = N/2$ homologous pairs, this factor is, on average, $(1 + \rho^2)/2$ and will differ only marginally from this in the abnormal cells of the types considered. Since the null hypothesis variance is very closely $1/50$ (when $N = 46$), we have under the alternative hypothesis

$$E(T^*) \cong \frac{2\rho^2}{1 + \rho^2}. \tag{6.1}$$

Similarly, we have to this degree of approximation

$$\text{Var } T^* \cong \frac{1}{50} \frac{\rho^4}{[(1 + \rho^2)/2]^4}. \tag{6.2}$$

At this level of approximation we may assess the effect of testing significance using the mean of m independent values of T^*, by a normal approximation to the distribution of such a mean. For example, if $\rho = 0.95$ and we test at the five per cent level of significance, we should need to average about 110 values of T^* to get a 0.975 power. These figures are fairly typical of the kind of power to be aimed at and the degree of contraction one might reasonably aim to detect. The inference is that it is necessary to compute T^* for the order of one hundred karyographs and average the results in order to have a reasonable chance of detecting a relative contraction so small as 5 per cent.

A more precise expression of the effects of the model is to say that under the alternative hypothesis

$$T^* = \frac{\rho^2 U}{1 - \frac{N - n}{N - 1}(1 - \rho^2)U}, \tag{6.3}$$

where U is distributed as T^* under the null hypothesis. We may thus employ the technique of David [9] to get a more accurate approximation to the power of a single value of T^*. Thus we require, for some number a,

$$(6.4) \qquad P\{T^* \leqq a\} \equiv P\left\{U \leqq a \Big/ \left[\rho^2 + a(1-\rho^2)\left(\frac{N-n}{N-1}\right)\right]\right\}.$$

For example $N = 46$, $n = 23$, and Var $U = 1/50$, (it may be remarked that the reciprocals of the variances of U in the cases of the first ten normal females in table I are 49.5, 49.3, 49.0, 48.8, 48.8, 48.5, 49.3, 49.0, 49.3, 48.8), then using the 5 per cent lower level of significance of T^* there is a 95 per cent chance of detecting a contraction of $\rho = 0.615$.

REFERENCES

[1] D. E. Barton and F. N. David, "Some notes on ordered random intervals," *J. Roy. Statist. Soc. Ser. B*, Vol. 18 (1956), pp. 79–94.

[2] ———, "Randomization bases for statistical tests. I. The bivariate case; randomization on N points in the plane," *Bull. Inst. Internat. Statist.*, Vol. 39 (1962), pp. 188–189; pp. 455–467.

[3] ———, "The analysis of chromosome patterns in the normal cell," *Ann. Hum. Genet.*, Vol. 25 (1962), pp. 323–329.

[4] ———, "The analysis of chromosome patterns in the abnormal cell," *Ann. Hum. Genet.*, Vol. 26 (1963), pp. 347–348.

[5] D. E. Barton, F. N. David, and M. Merrington, "Numerical analysis of chromosome patterns," *Ann. Hum. Genet.*, Vol. 26 (1963), pp. 349–353.

[6] ———, "The positions of the sex chromosomes of the human cell in mitosis," *Ann. Hum. Genet.*, Vol. 28 (1964), pp. 123–128.

[7] ———, "The relative positions of the chromosomes in the human cell in mitosis," *Ann. Hum. Genet.*, Vol. 29 (1965), pp. 139–146.

[8] Ciba Foundation, "London conference on the normal human karyotype," *Cytogenetics*, Vol. 2 (1963), pp. 264–268.

[9] F. N. David, "Sensitivity of analysis of variance tests for random variation between groups," *Trab. Estadist.*, Vol. 2 (1951), pp. 179–188.

[10] F. N. David, M. G. Kendall, and D. E. Barton, *Symmetric Function and Allied Tables*, Cambridge, Cambridge University Press, 1966.

[11] M. Merrington and L. S. Penrose, "Distances which involve satellited chromosomes in metaphase preparations," *Ann. Hum. Genet.*, Vol. 27 (1964), pp. 257–259.

[12] U. Mittwoch, "Sex differences in cells," *Sci. Amer.*, Vol. 209 (1963), pp. 54–62.

[13] J. H. Tjio and A. Levan, "The chromosome number of man," *Hereditas*, Vol. 42 (1956), pp. 1–6.

[14] M. J. D. White, *The Chromosomes*, London, Methuen; New York, Wiley, 1961 (5th ed.).