Biometrika (1975), 62, 2, *p.* 467
With 1 text-figure
Printed in Great Britain

A model for clustering

By DAVID J. STRAUSS
Department of Statistics, University of California, Riverside

Summary

A number of tests are available for the null hypothesis that a set of points in a region are scattered randomly, but relatively little is known about forms for the alternative. It is shown that under certain assumptions, the most severe of which is of Markov type, the probability density of the points must be of a simple explicit form depending on a single clustering parameter. The estimation of the parameter is studied and illustrated with an example.

Some key words: Clustering; Hammersley–Clifford theorem; Markov random field; Multidimensional point process; Persistence; Poisson process.

1. Introduction

In recent years the problem of generalizing the discrete Markov chain to higher dimensions has attracted considerable attention. For the two-state nearest-neighbour system on a Euclidean lattice, the so-called Markov random field, it has been proved by Spitzer (1971) and others that the probability density is necessarily of the explicit form of the Gibbs random field. The distribution, under the null hypothesis of randomness, of certain statistics measuring clustering or diversity has been studied by David (1970, 1971), who considers both the multistate and the multiple occupancy generalizations. A very general theorem, specifying the probability structure under an alternative hypothesis of clustering, has been given by Grimmett (1973).

Suppose that, given a set of points in Euclidean space, we wish to test for clustering and measure its intensity if it exists. This might be appropriate if, for example, the points represent incidence of a disease on a map of a city. An alternative to partitioning the space by a lattice framework is to seek a continuous model for the probability density of the points themselves. This leads to the idea of defining the point density to be some function of all the interpoint distances. Such a density will be invariant under translation and rotation of the data points, a property which will in many practical situations be a natural and desirable one. However, the general model of this type seems to be intractable.

This paper is concerned with the special case when the distance between two points is replaced by an indicator variable showing whether or not the points are 'close'. In addition we will require that the joint density is a symmetrical function of the data points. Otherwise our probability density would be inconsistent, in the sense of being dependent on the labelling of the data points. It is then an easy consequence of Euler's theorem on homogeneous functions that our joint density must be a function of the sum of the indicator variables. That is, the density depends only on the number of pairs of points which are close. Provided our space is large enough, the number of close pairs, Y, may range from 0 to $\frac{1}{2}n(n-1)$; the density may be assigned arbitrary values at each Y value, subject to their sum being unity.

Clearly some further restriction will be necessary if we want a reasonable model for clustering. The approach adopted here is to make an additional assumption of Markov type, and prove that it holds if and only if the joint density is of a simple geometric form, depending on a single clustering parameter. The form is similar to that of the discrete Gibbs random field, and the equivalence lemma may be regarded as a continuous analogue of Spitzer's theorem.

2. A CHARACTERIZATION LEMMA

Let D be a bounded, not necessarily connected, subset of ν-dimensional Euclidean space. The data are represented by n points $x_1, \ldots, x_n$ in D. Let r be a fixed positive number, and Y_n be the number of pairs of points whose Euclidean distance apart is less than r. Thus

$$Y_n = Y_n(x_1, \ldots, x_n) = \sum_{i<j} \alpha_{ij}, \tag{1}$$

where

$$\alpha_{ij} = \begin{cases} 1 & (|x_i - x_j| < r), \\ 0 & \text{otherwise.} \end{cases}$$

Denote by T_n the number of points in the set $\{x_1, \ldots, x_{n-1}\}$ which are within distance r of x_n; that is,

$$T_n = T_n(x_1, \ldots, x_n) = Y_n - Y_{n-1}. \tag{2}$$

We take, without loss of generality, the hypervolume of D to be unity. Consider the two following assumptions.

(a) The joint density of $(x_1, \ldots, x_n)$ is a function of y_n alone; that is, $f(x_1, \ldots, x_n) = \phi_n(y_n)$ for all $(x_1, \ldots, x_n)$ in D^n. This assumption was suggested in § 1; it holds under randomness, with $\phi_n(y) = 1$ for all possible y values.

(b) The density of X_n conditional on $x_1, \ldots, x_{n-1}$ is a function of t_n alone;

$$f(x_n | x_1, \ldots, x_{n-1}) = g_n(t_n), \tag{3}$$

say. A rough interpretation of (b) is that x_n is only affected by what happens within a range r of it, and further that only the number of points within the range is relevant. This restricted range of influence is analogous to the nearest-neighbour condition for a discrete Markov random field; see, for example, Spitzer (1971).

LEMMA. *For assumptions* (a) *and* (b) *to hold, it is necessary and sufficient that the joint density be of the form*

$$f(x_1, \ldots, x_n) = e^{vy}/M_{Y_n}(v). \tag{4}$$

Here $M_{Y_n}(.)$ is the moment generating function of Y_n under the randomness hypothesis. The parameter v measures the clustering tendency, and is independent of n.

Proof. Assume (4). Then (a) is true, and the left-hand side of (3) becomes $e^{v(y_n - y_{n-1})}$ times a function of n and v. Equation (3) now follows from (2).

Next, assume (a) and (b); (3) gives that, for all $(x_1, \ldots, x_n)$ in D^n, $\phi_n(y_n)/\phi_{n-1}(y_{n-1}) = g_n(t_n)$. Set $t_n = 0$. It will always be possible to pick a set $(x_1, \ldots, x_n)$ such that $t_n = 0$, except in the trivial case when D is a subset of every hypersphere of radius r with centre in D. Then for

$0 \leqslant y_{n-1} \leqslant \frac{1}{2}(n-1)(n-2)$, $\phi_n(y_{n-1})/\phi_{n-1}(y_{n-1}) = g_n(0)$. The lower bound for y_{n-1} will be larger if D is 'crowded'. Hence

$$\frac{\phi_n(y_{n-1}+t_n)}{\phi_n(y_{n-1})} = \frac{g_n(t_n)}{g_n(0)},$$

for all possible values of y_{n-1} and t_n.

If now we set $v = \log\{g_n(1)/g_n(0)\}$, then

$$\frac{\phi_n(y_{n-1}+t)}{\phi_n(y_{n-1})} = \frac{\phi_n(y_{n-1}+1)}{\phi_n(y_{n-1})} \cdots \frac{\phi_n(y_{n-1}+t)}{\phi_n(y_{n-1}+t-1)} = e^{vt}.$$

Thus $\phi_n(y) \propto e^{vy}$, for all possible y values. The constant of proportionality in (4) is obtained by noting that $f(x_1, \ldots, x_n)$ is a probability density. Finally, v must be independent of n. For if $v_n \neq v_{n-1}$ then (4) will not reduce (3) to a function of t_n alone.

There follow some remarks on the result.

(i) This paper deals with the case when the region of influence is a hypersphere of radius r, i.e. an interval, circle or sphere. The lemma holds for other shapes, such as a hyperinterval, with trivial modifications.

(ii) A physical interpretation of the parameter v is that given a set of data with two isolated points, the likelihood is increased by a factor e^v if one of the points is moved to within a distance r of the other. Negative values of v correspond to repulsion between the points, and $v = 0$ gives randomness.

(iii) According to (4), the density of Y for a given value of v, $f_v(y)$ is related to the density of y under randomness by

$$f_v(y) = f_0(y)\, e^{vy}/M_Y(v).$$

Thus if, for example, the null density of Y is approximately $N(\mu, \sigma^2)$, then the nonnull density is approximately $N(\mu + v\sigma^2, \sigma^2)$.

(iv) The lemma shows how, with assumptions (a) and (b), the problem of estimating the degree of clustering reduces to the problem of the distribution of Y under randomness. Because v is independent of n, it is possible to compare the clustering of different sets of data, particularly if r is the same in each case.

3. Some general properties

Let $K(.)$ be the cumulant generating function under the randomness hypothesis, and let κ_s be the sth null cumulant of Y. Let $\kappa_s(v)$ be the sth cumulant under the alternative hypothesis, when the clustering parameter is v, and let $K_v(t)$ be the corresponding cumulant generating function, with argument t. The $\kappa_s(v)$ are related to the null cumulants κ_s in the usual way in the exponential family, namely

$$\kappa_s(v) = \frac{\partial^s}{\partial v^s}\{K(v)\}. \tag{5}$$

Thus,

$$E(Y|v) = \kappa_1 + v\kappa_2 + \frac{v^2}{2!}\kappa_3 + \ldots, \tag{6}$$

and so on.

Both maximum likelihood and the method of moments lead to

$$\frac{\partial}{\partial v}\{K(\hat{v})\} = y \tag{7}$$

for the estimator $\hat{v}$. Equation (7) involves all the null cumulants, but it does not seem

possible to compute them all. To get an approximation for $\hat{v}$, two possibilities are to truncate (7) after, say, the fourth cumulant, which would give a cubic equation in $\hat{v}$, or to approximate the cumulants by those of a convenient fitted distribution, such as a linearly transformed χ^2.

For a χ^2 approximation, set $Y = a+bX$, where X is distributed as χ^2 on s degrees of freedom. The quantities a, b and s can be estimated from the cumulants of Y by

$$a = \kappa_1 - 2\kappa_2^2/\kappa_3, \quad b = \tfrac{1}{4}\kappa_3/\kappa_2, \quad s = 8\kappa_2^3/\kappa_3^2.$$

The cumulant generating function of Y is easily obtained, and (7) reduces to

$$\hat{v} = \frac{1}{2b} - \frac{s}{2(y-a)}. \tag{8}$$

Thus, for example, $\hat{v} = 0$ if $y = \kappa_1$, as expected.

Large-sample significance tests and confidence intervals for v are easily obtained from the asymptotic variance of v. The information statistic, from (4), is $K_y''(v)$. The chi-squared approximation for Y leads to

$$\text{var}\,(v) = (1-2bv)^2/(2b^2s).$$

Alternatively the fourth cumulant truncation of K can be used, giving

$$\text{var}\,(v) = (k_2 + k_3 v + \tfrac{1}{2}k_4 v^2)^{-1}.$$

The question arises of the adequacy of the χ^2 approximation. Of course an exact answer would require knowledge of all the null cumulants of Y, which seems unobtainable. Instead, one might have reasonable confidence if the χ^2 approximation and truncation of (7) at the fourth cumulant gave closely similar results. For the former assumes all the null cumulants of Y to be positive whilst the latter assumes all cumulants beyond the fourth to be zero. Unfortunately it is scarcely possible to compare the two approximations systematically since there are evidently too many variables to consider, namely the number of dimensions, the number of data points, the value of r, and the observed value of Y. However, the following comments may be useful.

(i) If the observed y is sufficiently close to its null mean κ_1, both approximations are satisfactory in the sense that the difference between the true $\hat{v}$ and the computed approximation tends to zero as $\hat{v}$ tends to zero. This is clear from (7) and (8). Indeed if $y = \kappa_1$ and so $\hat{v} = 0$, the right answer would be obtained either from (8), or by truncation of (7) even after the second cumulant, a normal approximation to the distribution of Y.

(ii) As a rule of thumb, it seems from examples that even the normal approximation will suffice if y is so close to κ_1 that we are primarily interested in a test of the hypothesis of no clustering, $v = 0$. However, for practical problems showing a moderate degree of clustering truncation after only two or three cumulants is not recommended.

(iii) Another rough rule of thumb is that when $|Y-\kappa_1|$ is less than about ten standard deviations the approximations from χ^2 and from fourth cumulant truncation are 'tolerably close'; it seems that they usually differ by less than 5 %. An example is given in § 5. The approximations breakdown for very extreme cases of clustering or segregation, and there is then no satisfactory estimator for $\hat{v}$. In any event, the assumptions of the present model would scarcely apply to, for example, the pattern of fruit trees in a commercial orchard.

4. The model in one, two and three dimensions

4·1. *General*

This section gives some formulae and results necessary for application of the model. The derivations are elementary, but in many cases rather lengthy; we give a few details below, in order to indicate the method. We use the notation of (1). Factorial powers and moments are denoted by round and square brackets respectively.

4·2. *One dimension*

The domain D here is the unit interval, on which are the n data points $x_1, \ldots, x_n$. The 'sphere of influence' of the lemma becomes, for each n, the interval $\{(x-r, x+r) \cap (0, 1)\}$. We give the first four cumulants of the statistic Y for the case when the distance r does not exceed $\frac{1}{2}$. This case is probably the most useful in practice. Results for r values greater than $\frac{1}{2}$ can be obtained similarly. We have

$$\mu = E(Y) = \binom{n}{2} E(\alpha_{12}) = \binom{n}{2}\left\{(1-2r)(2r) + 2\int_0^r (x+r)\,dx\right\}$$
$$= \binom{n}{2} r(2-r).$$

For higher cumulants it is simplest to use factorial moments. Thus

$$\mu_{[2]} = E(Y^{(2)}) = \tfrac{1}{4} n^{(4)} E(\alpha_{12}\alpha_{34}) + n^{(3)} E(\alpha_{12}\alpha_{13})$$
$$= \tfrac{1}{4} r^2 (2-r)^2 n^{(4)} + \tfrac{2}{3} r^2 n^{(3)} (6-5r).$$

The third factorial moment involves five types of term, obtained by the Vandermonde expansion of $Y^{(3)}$; see Table 1. Here $\mu_{[3]}$ is the sum of products from the last two columns.

For the fourth factorial moment, Table 2 gives eleven distinct cases. For example, $E(\alpha_{12}\alpha_{23}\alpha_{34}\alpha_{41})$ is best obtained by first fixing x_1, and then integrating over x_3 the square of the probability of obtaining a point within a distance r of both x_1 and x_3. This requires separate calculations according to whether x_1 lies within r of an end-point, between r and $2r$, or neither.

The third and fourth cumulants may be obtained from Tables 1 and 2. It can be shown that, for each value of r, as $n \to \infty$

$$\kappa_4/\kappa_2^2 = O(n^{-1}), \quad \kappa_3^2/\kappa_2^3 = O(n^{-1}),$$

suggesting a limiting normal form for Y. It is not apparent how the exact distribution of Y might be obtained.

Table 1. *Third factorial moment of Y*

Typical term in expansion	Number of terms	Expectation of term
$\alpha_{12}\alpha_{34}\alpha_{56}$	$\frac{1}{6}n^{(6)}$	$r^3(2-r)^3$
$\alpha_{12}\alpha_{13}\alpha_{45}$	$\frac{1}{2}n^{(5)}$	$\frac{2}{3}r^3(2-r)(6-5r)$
$\alpha_{12}\alpha_{13}\alpha_{14}$	$n^{(4)}$	$r^3(8-\frac{17}{2}r)$
$\alpha_{12}\alpha_{13}\alpha_{34}$	$3n^{(4)}$	$r^3(8-\frac{52}{6}r)$
$\alpha_{12}\alpha_{23}\alpha_{31}$	$n^{(3)}$	$r^3(3-2r)$
Total number of terms, $\binom{n}{2}^{(3)}$		

Table 2. *Fourth factorial moment of Y*

	Typical term	Number of terms	Expectation
(i)	$\alpha_{12}\alpha_{34}\alpha_{56}\alpha_{78}$	$\frac{1}{105}n^{(8)}$	$r^4(2-r)^4$
(ii)	$\alpha_{12}\alpha_{13}\alpha_{45}\alpha_{67}$	$\frac{2}{3}n^{(7)}$	$\frac{2}{3}r^4(2-r)^2(6-5r)$
(iii)	$\alpha_{12}\alpha_{13}\alpha_{14}\alpha_{56}$	$2n^{(6)}$	$r^4(2-r)(8-\frac{17}{3}r)$
(iv)	$\alpha_{12}\alpha_{23}\alpha_{34}\alpha_{56}$	$6n^{(6)}$	$r^4(2-r)(8-\frac{23}{6}r)$
(v)	$\alpha_{12}\alpha_{14}\alpha_{25}\alpha_{26}$	$3n^{(5)}$	$\frac{4}{9}r^4(6-5r)^2$
(vi)	$\alpha_{12}\alpha_{23}\alpha_{34}\alpha_{45}$	$12n^{(5)}$	$r^4(16-21{\cdot}4r)$
(vii)	$\alpha_{12}\alpha_{13}\alpha_{14}\alpha_{15}$	$n^{(5)}$	$r^4(16-19{\cdot}6r)$
(viii)	$\alpha_{12}\alpha_{23}\alpha_{31}\alpha_{45}$	$2n^{(5)}$	$r^3(2-r)(3-2r)$
(ix)	$\alpha_{12}\alpha_{13}\alpha_{14}\alpha_{45}$	$12n^{(5)}$	$r^4(16-\frac{92}{3}r)$
(x)	$\alpha_{12}\alpha_{23}\alpha_{34}\alpha_{41}$	$3n^{(4)}$	$\frac{16}{3}r^3(1-r)$
(xi)	$\alpha_{12}\alpha_{13}\alpha_{23}\alpha_{14}$	$12n^{(4)}$	$r^3(6-\frac{17}{3}r)$

Total number of terms, $\binom{n}{2}^{(4)}$

4·3. *Two dimensions*

The region of influence of a point x becomes a circle of radius r and centre x. The domain D, which may represent a forest or a city, may have any shape, and has unit area. We write $A = \pi r^2$. The case when an elliptical shape for the region is regarded as preferable may be handled simply by scaling it along one dimension. Other shapes for the region, such as a rectangle, would require separate calculation. In practice, to compute the cumulants we require an additional assumption that A is small enough for boundary effects to be neglected. Alternatively, the results below may be regarded as exact for the case when D is a closed surface, such as that of a sphere.

Clearly

$$\kappa_1 = E(Y) = \binom{n}{2}A, \quad \kappa_2 = \mathrm{var}\,(Y) = \binom{n}{2}A(1-A).$$

It can be shown that

$$\kappa_3 = \binom{n}{2}A(1-A)(1-2A) + n^{(3)}\left(1-\frac{3\sqrt{3}}{4\pi}-A\right)A^2.$$

In Table 2 for the fourth factorial moment, all expectations are A^4 except for (viii) and (xi), each of which is $A^3(1-\frac{3}{4}\sqrt{3}/\pi)$ and (x), which simplifies to $A^3\{1-16/(3\pi^2)\}$.

This last formula, for example, is obtained by noting that the distance between x_1 and x_3 has probability density $2\pi d$, and that for each d the required probability is the square of the area common to two circles of radius r with centres x_1 and x_3. The fourth cumulant reduces to

$$\kappa_4 = \binom{n}{2}A(1-A)\{1-6A(1-A)\} + n^{(3)}A^2\left\{A\left(1-\frac{16}{3\pi^2}-A\right)+6\left(1-\frac{3\sqrt{3}}{4\pi}-A\right)(1-2A)\right\}.$$

For the limiting distribution, we have that

$$\kappa_3^2/\kappa_2^3 = \beta_1 = (1-\tfrac{3}{4}\sqrt{(3)}/\pi - A)^2 A/(1-A) + O(n^{-1}),$$
$$3+\kappa_4/\kappa_2^2 = \beta_2 = 3 + A\{16/\pi^2 - 3\sqrt{(3)}/\pi + 7(1-A)\}/(1-A)^2 + O(n^{-1}),$$

which do not converge to the normal form as $n \to \infty$ unless the fixed value of A can be regarded as negligible.

4·4. *Three dimensions*

As in § 4·3, we need to assume that V, the volume of a sphere of radius r, is small compared with the unit volume of D. Calculations are comparable to those of § 4·3; we again find

$$\kappa_1 = \binom{n}{2}V, \quad \kappa_2 = \binom{n}{2}V(1-V),$$

whilst

$$\kappa_3 = \binom{n}{2} V(1-V)(1-2V) + n^{(3)}(15/32 - V)\, V^2.$$

For the fourth cumulant, (viii) and (xi) of Table 2 have expectation $15V^3/32$, and (x) has expectation $1654V^3/105$. Hence, after simplification,

$$\kappa_4 = \binom{n}{2} V(1-V)\{1-6V(1-V)\} + n^{(3)}V^2\{V(\tfrac{1654}{105} - V) + 6(\tfrac{15}{32} - V)(1-2V)\}.$$

For large values of n

$$\beta_1 = \frac{8V^3(\frac{15}{32} - V)}{(1-V)^3} + O(n^{-1}),$$

whilst $\beta_2 = 3 + O(n^{-1})$. Thus in three dimensions the kurtosis always converges to zero, whilst the skewness does so only if V can be regarded as negligible here.

5. The choice of r

The range of influence, r, may set at any value which seems physically reasonable for the given problem. Usually it will be sensible to try more than one value of r. It might be argued loosely that the r giving the greatest absolute value for $\hat{v}$ is the most sensitive. Two further considerations are as follows.

(*a*) To test the hypothesis $v = 0$ against the alternative $v > 0$ the uniformly most powerful critical region is given by $\{y: y > c\}$. If r is to be selected so that the test based on it is optimal, one is led to seek the value of r that yields a test with maximal asymptotic efficiency. It follows from (3) that the appropriate value of r is that which maximizes the null variance k_2. However, it turns out that this is obtained when $r = \frac{1}{2}$, $A = \frac{1}{2}$ and $V = \frac{1}{2}$ in one, two and three dimensions, and such values will normally be regarded as unacceptably large.

(*b*) Provided that it is physically realistic, a good choice of r might be one that makes the fourth cumulant of the fitted χ^2 distribution close to the true fourth cumulant of Y. This will depend on the value of n. It does not seem worthwhile to construct charts of the fourth cumulants. Here we only remark that the standardized fourth cumulants converge to zero in one and three dimensions. In two dimensions the limit is of the order of A, assumed negligible. Thus the closeness to the fitted χ^2 fourth cumulant should become unimportant as n gets large. Also it seems that for n less than 100 and r less than 0·15 in one, two or three dimensions, the fitted χ^2 cumulant almost always lies between $\frac{2}{3}$ and $\frac{4}{3}$ of the true fourth cumulant.

6. An example

Figure 1 shows the distribution of 199 redwood seedlings found on a square experimental plot. It was felt that the seedlings would be scattered fairly randomly, except that a number of tight clusters would form around some of the redwood tree stumps present in the plot. A discontinuity in the soil, very roughly demarked by the diagonal line in the figure, was expected to cause a difference in clustering behaviour between regions I and II. Moreover, almost all the redwood stumps were situated in region II.

We compare the two regions by fitting the model. Possibly the Markov assumption (3) is reasonably appropriate here. Naturally neither assumption of § 2 will be strictly valid,

perhaps the most likely difficulty being variations in fertility within each region. The difference in mean density of seedlings between the two regions is no problem because the parameter v is independent of the number of points. As will usually be the case, the choice of r has to be somewhat arbitrary. Initially the value r_1, indicated in the figure, was selected. It corresponds to about six feet on the ground, which was thought to be very roughly the range at which a pair of seedlings could 'interact'.

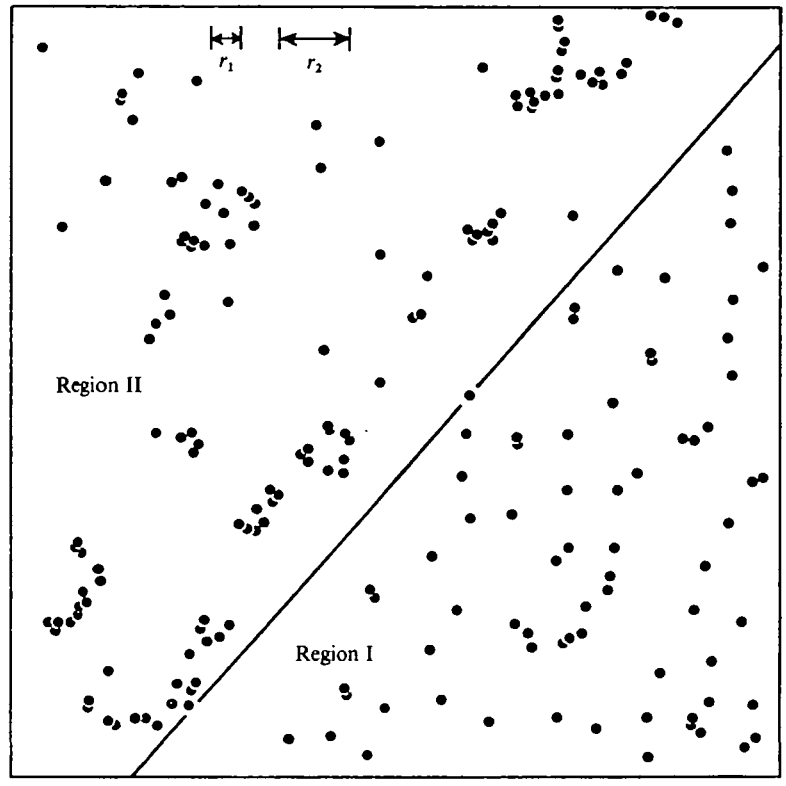

Fig. 1. 199 redwood seedlings in an experimental plot.

Using the notation and formulae of § 3, we have for region I, the lower right portion in the figure, $y = 37$, $n = 77$, $A = 0{\cdot}012223$. The cumulants are $\kappa_1 = 35{\cdot}7644$, $\kappa_2 = 35{\cdot}3273$, $\kappa_3 = 72{\cdot}1205$, $\kappa_4 = 253{\cdot}572$. The closeness of y to κ_1 confirms the visual impression that there can be little clustering in region I. We obtain $\hat{v} = 0{\cdot}034$, with a standard error of $0{\cdot}16$, both with the χ^2 approximation to $K_Y(v)$ and with the truncation of K at the fourth cumulant. Obviously region II shows clustering. Here $y = 275$, whilst $\kappa_1 = 60{\cdot}919$, etc. Fourth cumulant truncation yields $\hat{v} = 1{\cdot}03$, with a standard error of $0{\cdot}048$. For such strong clustering the χ^2 approximation is unsatisfactory. Indeed, according to (8), the maximum value possible, when $y = \infty$, is $\hat{v} = 0{\cdot}94$. Naturally the approximation to $\hat{v}$ obtained from truncation at κ_4 is also open to question, and unfortunately there seems to be no way of getting the exact value. As a very crude and *ad hoc* check, it turns out that if we set $\kappa_r = 5\kappa_{r-1}$ for $r \geqslant 5$, quite arbitrarily, but hopefully fairly generously, we still obtain $\hat{v} = 0{\cdot}93$; whilst if some of the higher cumulants, or $\hat{v}$, are negative the error in truncation should be greatly reduced.

Although testing the hypothesis of no clustering, $v = 0$, is not our main concern here, a remark may be in order. Instead of the standard normal test based on $(y-\kappa_1)/\sqrt{\kappa_2}$, the fitted χ^2 test should be more accurate. By ignoring the skewness in the distribution of Y, the normal test in each situation considered in this example underestimates the critical point. However, for a two-tailed test at the 5% level, the discrepancy in the two critical points is in each case only about 1%, and is probably of little practical importance.

It might well be that still larger values of v could be obtained with different choices of r. For comparison we give the analysis based on r_2; this distance is twice r_1, and probably too large to be appropriate. For region I, $y = 142$, $\kappa_1 = 143{\cdot}052$, etc. Either method gives $\hat{v} = -0{\cdot}008$ with standard error of 0·09, confirming the impression of no clustering.

For region II, fourth cumulant truncation gives $\hat{v} = 0{\cdot}359$. The greatly reduced value of $\hat{v}$ obtained with r_2 instead of r_1 would be misleading if viewed by itself. Finally, even here y is about 20 standard deviations above its null mean, see § 3, and the χ^2 fit gives the rather different value of 0·287.

I am grateful to a referee for comments on an earlier draft of this paper.

References

David, F. N. (1970) Measurement of diversity I. *Proc. 6th Berkeley Symp.* **1**, 631–48.
David, F. N. (1971). Measurement of diversity II. *Proc. 6th Berkeley Symp.* **4**, 109–36.
Grimmett, G. R. (1973). A theorem about random fields. *Bull. Lond. Math. Soc.* **5**, 81–4.
Spitzer, F. (1971). Markov random fields and Gibbs ensembles. *Am. Math. Mon.* **78**, 142–54.

[*Received November* 1973. *Revised December* 1974]