Acta Cryst. (1967). **22**, 329

The Statistics of the Distribution of Crystalline Substances Among the Space Groups

By A. L. Mackay

*Department of Crystallography, Birkbeck College, Malet Street, London, W.C.*1, *England*

(*Received* 29 *November* 1965 *and in revised form* 5 *August* 1966)

The summary tables of *Crystal Data* (first edition), which give the space groups to which 3782 crystal species belong, have been used in examining the probability distribution of space groups. If M_t is the number of different space groups which occur (in the tables) at least t times, then $1/M_t$ is found to vary almost linearly with t. This empirical distribution enables an experimental estimate to be made of the total number of space groups, although some have not been encountered. Such an estimate agrees reasonably with the accepted number, thus validating the technique.

The statistical relationship between the number and the size of genera in a classification system – the species problem – has been examined several times (*e.g.* Willis, 1924; Yule, 1924; Zipf, 1932; Good, 1953) but is still not entirely solved.

Since we know from very well founded theoretical investigations the number of space group categories into which different crystal species can be classified (*International Tables for X-ray Crystallography*, Vol. 1), we are in a uniquely favourable position to test the statistical analysis of the species problem. The problem is to estimate, given a sample of a heterogenous population in which N genera occur, how many more genera there are likely to be in the general population from which the sample is drawn.

There are 230 genera of space groups, but we are concerned only with the number of genera which are in practice distinguishable. Although by anomalous scattering it is possible to distinguish enantiomorphs, this is rarely done and the 11 enantiomorphous groups are excluded from our theoretical total. The pairs $I222$ and $I2_12_12_1$ and also $I23$ and $I2_13$ are not distinguishable by systematic extinctions but solved structures can be allocated without error to the correct space group. The number of genera is therefore 219.

The statistical tables compiled by Nowacki (1954), which list the space groups to which each of 3782 crystal species belong, enable us to examine the distribution by genera. In this very large sample only 178 of the 219 genera in fact occur.

We note first the wide variation in frequencies among the different space groups: one group occurs 355 times in the sample, while 33 occur only once each, 17 only twice each, and so on. Let N_t be the number of different groups each of which occurs exactly t times in this sample of 3782 species, and let M_t be the number each of which occurs at least t times. 43 groups occur 21 or more times each.

The values of t, N_t and M_t are listed in Table 1.

The species problem is the estimation, from the statistical data, without *a priori* knowledge of the answer, of how many space groups would be found if an indefinitely large sample of materials were available. This is equivalent to finding the value of M_0 by extrapolation.

Fig. 1 shows the monotonic variation of M_t with t. It was found that the data could be replotted as a sensibly straight line showing $1/M_t$ to be an almost linear function of t. This was extrapolated linearly to give the intercept $1/M_0$ which leads to a value of about 222 for M_0 (Fig. 2). The line was drawn through the final point $M_1=178$, since this represents the fullest data available. Extrapolation was also carried out by computer fitting a cubic by least-squares criteria and led to a value of 216. The weight of each point was

Table 1. *Distribution of* 3782 *crystal species among the space groups*

N_t is the number of space groups which each occur t times. M_t is the number of space groups each of which occurs at least t times. M_t calc. is given by $1/M_t=1/M_0+t/K$ where $M_0=219$ and $K=951$. The first two point of M_t calc. are thus fitted to the data. The sample is the list of 3782 crystal species from *Crystal Data* (1st. ed.).

t	N_t	M_t obs.	$\frac{1000}{M_t}$	M_t calc.
0	(41)	(219)	(4·566)	(219)
1	33	178	5·618	(178)
2	17	145	6·897	149·9
3	17	128	7·813	129·5
4	12	111	9·009	114·0
5	12	99	10·01	101·9
6	4	87	11·49	91·9
7	5	83	12·05	83·8
8	3	78	12·82	77·0
9	6	75	13·33	71·3
10	3	69	14·49	66·3
11	6	66	15·15	62·0
12	1	60	16·67	58·2
13	4	59	16·95	54·8
14	2	55	18·18	51·8
15	6	53	18·87	49·2
16	1	47	21·28	46·7
17	0	46	21·74	44·5
18	1	46	21·74	42·6
19	1	45	22·22	40·7
20	1	44	22·73	39·0
21 and		43		
above	43	..		

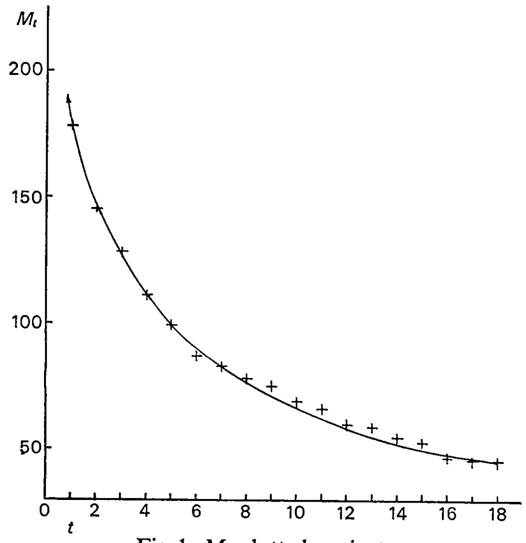

Fig. 1. M_t plotted against t.

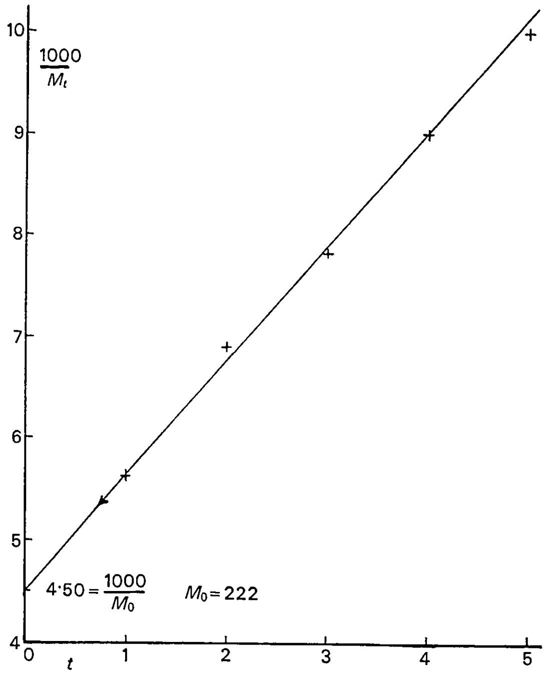

Fig. 2. $10^3/M_t$ plotted against t and extrapolated to give $1/M_0$.

made proportional to M_t. Judging approximately from the fit of the curve, the accuracy of the process is about 2% so that the estimate agrees reasonably with the accepted value of M_0 of 219.

The empirically derived function representing the straight line is thus: $1/M_t = 1/M_0 + t/K$, where K is a constant. Inserting the theoretical value of M_0 (219) and fitting the single point $M_1 = 178$, gives the value for K of 951. The values of M_t calculated from this equation are listed in the last column of Table 1, where it can be seen that they agree well with the observed values of M_t. This fit justifies the extrapolation procedure.

When the above analysis is applied separately to organic and inorganic crystals, in both cases a low result for the value of M_0 is obtained (185 and 209 respectively). This implies that the space groups to which organic and inorganic crystals belong form different populations.

We conclude that, as the procedure for estimating the number of space groups not yet encountered does indeed give a value within about 10% of the correct value (41), the application of this extrapolation technique in other fields (Mackay, 1966) justifies increased confidence. We might also deduce that, as the number of groups not yet observed agrees with statistical estimates, their absence is due to chance and not to the intrinsic impossibility for physical reasons of the occurrence of certain space groups.

References

Good, I. J. (1953). *Biometrika*, **40**, 237.

Mackay, A. L. (1966). *Statistical Methods in Linguistics*, No. 4, p. 15.

Nowacki, W. (1954). In Donnay, J. D. H. & Nowacki, W. *Crystal data.* 1st ed. Geol. Soc. Amer. Memoir 60.

Williams, C. B. (1947). *J. Ecology*, **34**, 17.

Willis, J. C. (1924). *Nature, Lond.* **112**, 178.

Yule, G. U. (1924). *Phil. Trans. B.* **213**, 21.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort.* Cambridge, U.S.A.