IS GENE DISPERSION NORMAL ?

A. J. BATEMAN

John Innes Horticultural Institution, Bayfordbury, Hertford, Herts.

Received 22.x.49

INTRODUCTION

IN the early theoretical studies of population genetics (Wright, 1931) the simplest cases naturally were considered. Populations were taken to be internally and externally panmictic. That is mating was completely at random within each population and finite migration occurred at random between otherwise isolated populations. Such is, of course, an ideal situation never realised in natural populations, though examples of true *internal* panmixis exist in crop plants in which all the seed from a population is thoroughly mixed before sowing the next generation, and in animals such as fish in ponds, where the radius of the occupied space is much smaller than the mean distance travelled per generation.

Apart from such exceptions there will always be a correlation within populations between the proximity of individuals to each other and their genetic relationship. There will be a similar correlation between the proximity of populations to each other and their genetic relationship. And since the greater the proximity the greater the chances of mating, there will always be a certain amount of in-breeding relative to the population as a whole or the group of populations, whatever may be the size and number of the populations.

In later years (Wright, 1943, 1946) allowance was made for this by considering the structure of continuous populations. The symbol N , originally the size of a panmictic population, was carried forward into the new situation in a new role. It now represented the sizes of sub-populations into which a continuous population was theoretically divisible. In order to avoid implying any particular breeding behaviour these sub-populations were referred to as "neighbourhood units." The magnitude of the new N was such as to give, with random dispersion within the neighbourhood, an average gene dispersion equivalent to that obtaining in the non-panmictic continuous population.

It may be noted in passing that fixation and drift are much less dependent on the breeding behaviour itself than on the gene dispersion. Hence the introduction of the neighbourhood concept to cover everything from true panmixis to populations of self-fertilising plants where the genes are always dispersed as zygotes.

N is given by the number of individuals within a circle of radius 2σ where σ^2 is the variance of gene dispersion within the species. Malécot

(1948), dealing with the same problem, relates the variation within a continuous population directly to σ^2 without the introduction of the neighbourhood concept. From the key position that gene dispersion holds, its estimation under natural conditions for various organisms is clearly of importance to population genetics. Such an estimation has been made for *Drosophila pseudoobscura* by Dobzhansky and Wright (1943 and 1947).

There are, however, other important parameters of continuous populations. Such are the mean degree of inbreeding and the frequency distribution of degrees of inbreeding. In dioecious organisms and self-incompatible plants the mean degree of inbreeding is adequately given by $1/N$, N being calculated from the gene dispersion. In organisms with varying degrees of self-fertilisation this is not so. In the extreme case of complete self-fertilisation, inbreeding is unity whatever the value of N . Under no circumstances can the frequency distribution of degrees of inbreeding be expressed merely in terms of $1/N$. For this we must know not only the degree of dispersion (the variance) but the nature of the dispersion; whether it is skew or has abnormal kurtosis (the fourth degree statistic).

In the absence of evidence to the contrary it is natural to assume that gene dispersion should follow the normal distribution curve. This assumption has been made, for example, by Wright (1943, 1946) and by Haldane (1948) though in no instance was normality fundamental to the argument. It is desirable, however, whenever the data make it possible, to test the nature of gene dispersion. Skewness may be expected to be frequent. It will arise whenever there is a tendency for migration in one direction; say downwind.

The standard method for the estimation of kurtosis involves fourth degree statistics. If k_2 is the second moment (variance) and k_4 the fourth moment, the kurtosis of a distribution is given by the expression $k_4/k_2^2 - 3$. For a normal distribution $k_4/k_2^2 = 3$ and kurtosis is zero. With a platykurtic distribution in which the "shoulders" of the frequency curve have increased at the expense of the peak and "tails," $k_4/k_2^2 < 3$ and kurtosis is negative. With a leptokurtic distribution peak and tails are exaggerated, $k_4/k_2^2 > 3$ and kurtosis is positive. The three kinds of kurtosis are illustrated in fig. 1. Though this method of estimation of kurtosis is mathematically most directly related to the properties of normal distributions, a practical difficulty involved in its application is that k_4 is extremely sensitive to incompleteness of sampling area. Values farthest from the mean make disproportionately large contributions to k_4 and so with a finite sampling area there is always a possibility that there may be a loss of extreme values (a truncation of the tails of the distribution), a reduction in the ratio k_4/k_2^2 and a negative bias in the estimation of kurtosis. This is not merely a theoretical possibility. Dobzhansky and Wright (1947) found that high positive kurtosis on the first day, when the sampling area was most likely to extend beyond the limits

of the distribution of the flies, steadily diminished with succeeding days, when it became almost certain that the limits of distribution exceeded the sampling area. In some cases the original positive kurtosis was thereby transformed into a negative one. An alternative, if empirical, method of studying kurtosis has now been developed (Bateman, 1947). It is limited in application to the detection of marked kurtosis, but it is not dependent on sampling covering the complete distribution. This is the method of regression.

FIG. 1.—Half the curve for normal dispersion in one dimension is represented by the curved broken line. The Y axis therefore marks the centre of dispersion. The entire curve has positive kurtosis (leptokurtic). It shows abnormally high values for both short and long extreme distances. The rectangle represents an extreme form of negative kurtosis (platykurtic). It also represents Wright's neighbourhood unit, used as a model for gene dispersal in a continuous population. The length of its base is 2σ , its height is the density of the population and its area is $\frac{1}{2}N$.

The normal distribution can be expressed empirically as $y = ge^{-bx^2}$, where x is the deviation from the mean and g and b are constants. Logarithmically this becomes $\log y = a - bx^2$. Any normal distribution can therefore be adequately represented by the regression of $\log y$ on x^2 . Kurtosis will modify the regression line by increasing or decreasing the curvature (see fig. 2). Leptokurtosis will straighten the curve and even, if extreme, produce a concave instead of a convex curve. Such a change of curvature is equivalent to a change in the power of x from 2. If the power were 1 the curve would be straight,

if $\frac{1}{2}$, concave. If a distribution is leptokurtic the best fit for the regression of $\log y$ on x^n will be given when n is significantly less than 2.

FIG. 2.—Semilogarithmic curves are given for the frequency distribution of flight lengths of pollinating insects on two crops. This corresponds to the half of a distribution curve shown in fig. 1. The observed frequencies are shown as polygons. Also shown for comparison are the regression lines for $\log Y$ on X^2 (convex broken line) X (straight entire line) and $\sqrt{X}$ (concave broken line). The ideal curve would appear to be somewhere between the X and $\sqrt{X}$ regressions.

DISPERSAL OF ANIMALS

Examples of this method applied to the flight of pollinating insects have already been given (*ibid.*). The regressions of $\log y$ on x and $x^{\frac{1}{2}}$ were definitely superior to those on x^2 . The $x^{\frac{1}{2}}$ regression was slightly but not significantly better than the x regression. Further data have now been obtained and can be analysed in the same way. Some refinements have been introduced. Firstly, only the projection of

each flight on to a predetermined line was measured. Thus the data were immediately in the form of one-dimensional distribution and did not require conversion from a two-dimensional distribution as previously. Secondly, flights longer than three feet have been excluded from the regression since it was felt that errors of observation rapidly increase beyond this distance. Thirdly, to counteract the high error variance of $\log y$ when y is small (rising to infinity when $y = 0$) a system of weighting was introduced. The information (invariance) about $\log y$ is related to the information about y in the following way :—

$$I_{\log y} = \frac{dy}{d \log y} I_y$$

$$\frac{dy}{d \log y} = y \log_e 10$$

Therefore $I_{\log y} = y^2 \log_e^2 10 I_y$

V_{y_1} is given by the usual formula for binomial distributions adapted to the present circumstances ($p \equiv y_1$, $q \equiv (\Sigma y - y_1)$ and $n \equiv \Sigma y$) i.e. $\frac{y_1(\Sigma y - y_1)}{\Sigma y}$

Hence $I_{y_1} = \frac{\Sigma y}{y_1(\Sigma y - y_1)}$

and $I_{\log y_1} = \log_e^2 10 \frac{\Sigma y \cdot y_1}{\Sigma y - y_1}$

$\log_e^2 10$ is a common factor and can be ignored when weighting. Each observed $\log y_1$ is then weighted by multiplying by

$$\frac{\Sigma y \cdot y_1}{\Sigma y - y_1}$$

Fig. 2 shows the observed values of $\log y$ compared with the regression lines corresponding to the weighted regressions on x^2 , x and $x^{\frac{1}{2}}$. The goodness of fit of these regressions can be judged from table 1 by the comparison of their remainder sums of squares. The inadequacy of the x^2 regression (corresponding to a normal distribution) is clearly demonstrated, both graphically and statistically. There seems little to choose between the x and $x^{\frac{1}{2}}$ regressions. In fact the ideal regression would appear to be one in which the index of x was about $\frac{3}{4}$. This is stronger evidence than hitherto presented of the leptokurtic nature of the dispersion of foraging bees.

An independent set of data obtained by quite different methods is furnished by Dobzhansky and Wright (1947) on the dispersal of mutant flies of *Drosophila pseudoobscura* from their point of liberation. Semilogarithmic curves for some of the days are given in fig. 3. The flies were caught in traps placed in a straight line at intervals of 20 metres. Kurtosis as measured by k_4/k_2^2 was strongly positive on

TABLE 1

A comparison of the goodness of fit obtained by the regressions of $\log y$ on x^2 , x and $x^{\frac{1}{2}}$, where x is the flight length of hive bees and y is their frequency.

Data from	Regression on	Regression sum of squares	Remainder sum of squares	Degrees of freedom	Variance Ratio	P
Swede strip	x^2	129.6305	82.9529	5	22.50 < 0.01	3.06 < 0.2
	x	201.2890	11.2944			
	$x^{\frac{1}{2}}$	208.8961	3.6873			
Swede block	x^2	162.9438	136.3764	5	13.41 < 0.01	1.33 high
	x	285.7909	13.5294			
	$x^{\frac{1}{2}}$	289.1509	10.1693			
Radish block	x^2	450.6367	231.7158	5	19.40 < 0.01	7.09 < 0.05
	x	670.4069	11.9456			
	$x^{\frac{1}{2}}$	597.6855	84.6670

FIG. 3.—Semilogarithmic curves for the distribution of mutant flies one, three and six days after liberation at distance 0 (Dobzhansky and Wright's data). Also shown (continuous curve) is the expected first day's distribution on the basis of normal distribution. It gives expectations which are too low at the point of liberation and over 200 metres, but too high for intermediate distances. Casual inspection indicates that good fits for the data on all three dates would be obtained by a linear regression of $\log$ frequency on distance.

the first day and diminished steadily to the sixth day. The main difference between this fly data and the above bee data is that while the latter were based on single flights, the shortest interval in the former was a whole day during which there would be numerous flights per fly. The positive kurtosis of the flies is therefore even more unexpected. The authors recognised that the diminution of kurtosis with succeeding days might be due to increasing incompleteness of sampling (discussed above). In a private communication, however, Wright has held that the main factor reducing kurtosis was the normalising effect of subsequent flights. He assumes that consecutive flights of *Drosophila* are uncorrelated. If that were so it would be very surprising for the distribution of flies after the first whole day's flights still to be strongly leptokurtic in spite of the normalising effect to be expected from lack of correlation between the many successive flights during that day.

The regression method has been used on the above data to see whether any further light could be thrown on the question. The frequencies were again weighted before calculating the regression, this time by dividing every frequency of mutant flies by the total number of wild type flies (the original population) at that site. This allowed for the very noticeable variation in attractiveness of the sites and gave much smoother frequency curves. Only traps near to the point of liberation gave high enough counts of mutant flies to be used in the regression method.

The results are given in table 2. The superiority of the x regression over x^2 again demonstrates the leptokurtosis. This in spite of the

TABLE 2

Day	Regression on	Regression sum of squares	Remainder sum of squares	Degrees of freedom	Variance ratio	P
1	x^2	3.682968	1.359011	9	2.40	0.1
	x	4.476328	0.565651			
2	x^2	3.650894	1.482418	14	1.85	0.1
	x	4.334005	0.799307			
3	x^2	2.673754	0.971419	14	1.57	0.2
	x	3.026842	0.618331			
4	x^2	3.211919	0.575768	14	1.35	0.3
	x	3.359720	0.427967			
5	x^2	2.424195	0.815767	14	1.38	0.3
	x	2.647765	0.592197			
6	x^2	1.248294	0.588813	14	1.11	0.7
	x	1.307332	0.529775			

restriction of sampling to the proximal zone. The leptokurtosis cannot therefore be due merely to long "tails" to the distribution curve, but expresses itself also in other parts of the curve. Unfortunately no

single day's observations gives a significantly better x than x^2 regression. If, however, the probabilities for each day are multiplied a compound probability of 0.0001 is obtained. Such a compounding is not strictly justifiable unless the observations are independent of one another, which these are not, being based on a constant fly population. The very low compound probability does, however, give some confidence that the leptokurtosis is real.

The regression appears to agree with the fourth moment in showing a steady approach to normality with successive days. That is, the ratio of the remainder sums of squares for the x^2 and x regressions becomes steadily less significant. There is another purely statistical explanation for this. With each succeeding day the slope of the distribution diminishes. The sum of squares which is available for any linear regression will therefore diminish, while the true error sum of squares will presumably remain about constant. The remainder sum of squares will therefore increase relative to the amount taken out by any regression and to the differences between the amounts taken out by any two regressions. Correspondingly it will be increasingly unlikely that any one linear regression should become significantly superior to any other.

In conclusion, the application of the regression method to the *Drosophila* observations, although confined only to the proximal part of the distribution, confirms the leptokurtosis originally found by an unrelated method. It also appears to confirm a day-to-day reduction in kurtosis, though other interpretations of the results from both methods of analysis are possible.

WIND DISPERSAL

Semilogarithmic curves for pollen dispersal (Bateman, 1947) show straight or even slightly concave curves relating $\log y$ to x (therein called $\log P$ and D respectively). They are also fitted well by regression lines using powers of x equal to or less than unity. In view of the general similarity of all wind dispersal it is probable that spores would be distributed in a similarly leptokurtic way.

CROSS-POLLINATION

Cross-pollination is here used in the sense of pollination between two populations or groups of plants which are separated by distance x .

The distribution of cross-pollination in wind-pollinating species will be merely an extension of pollen dispersal and the same leptokurtosis will apply. One limitation which will rarely have an effect is that there is a maximum degree of cross-pollination *viz.* 100 per cent. This would at most cause a slight depression of the peak of the distribution, without affecting the tails.

The situation in insect-pollinated species is more complex. In the first place it is dependent on the flight of the pollinating insects

which, as we have seen, is leptokurtic. But it can be shown (*ibid.*) that as with each visit the insect not only deposits pollen but replaces part of its load by pollen from the flower it is visiting, the leptokurtosis tends to increase still further. Thus if the distribution of the insect were such as to give a straight line relation between $\log y$ and x , the curve for cross-pollination at distance x from the source would be concave. This expectation was confirmed by the data obtained from experiments on cross-pollination.

THE ORIGIN OF LEPTOKURTOSIS

The following types of gene dispersion have thus been shown to be leptokurtic: automatic movement of airborne insects (Dobzhansky and Wright, 1943, 1947; Bateman, 1947 and present paper); water fleas (Brownlee, 1911); passive movement of airborne organisms (Gregory, 1945; Bateman, 1947); dispersal of pollen by insects (Bateman, 1947). There remain a number of important types of gene dispersion for which the necessary information is not yet available: the movement of terrestrial animals and birds; the movement of plankton; and the dispersal of seeds by explosive and other automatic mechanisms. When such a varied collection of known dispersal types has been shown to be consistently leptokurtic there is no reason to doubt that the balance would be similarly non-normal. It is possible, therefore, that gene dispersion is as a general rule, leptokurtic.

There is an apparent contradiction between such a situation and the general theory of normal distributions. Under this theory normality should result whenever a large number of uncorrelated influences are at work. The contradiction forces us to the conclusion that neither the passive nor active movement of organisms is at random. The leptokurtosis observed in the distribution of pollinating insects after single flights suggests a tendency for flights to be either short or long, though the smoothness of the curve shows that there are also all intermediate flight lengths. A possible explanation would be that some flights involved the continuation of an extremely localised forage whereas others were exploratory and ranged over a wider field.

Even if the distribution of *Drosophila* after one flight had been similarly leptokurtic, only a strong correlation between consecutive flights in direction, length, or both would maintain the leptokurtosis after one day's flying. Since the leptokurtosis remained in evidence for several days the correlation between flights must have remained over that period. Unless, however, the correlation were absolute there would be a day-by-day drift to a normal distribution. If we were to classify the individual flies according to their flight lengths into sedentary and wandering classes the classification would tend to become steadily more blurred with extension of time. A closely parallel classification of hive bees, varying with time, has been postulated by Butler (1945). The leptokurtosis of pollen dispersion

by insects is merely an accentuation of that for the insects themselves and requires no separate consideration.

The explanation of the leptokurtosis produced by wind dispersal is to be found in a theoretical discussion of the subject in which the problem is treated as an example in aerodynamics (Gregory, 1945) and is discussed in terms of atmospheric turbulence or eddy diffusion. Gregory's general formula has been expressed in simpler terms (Bateman, 1947) as

$$y = \frac{ge}{x^{(1+\frac{1}{2}m)}} - \left[\frac{cx^{(1-\frac{1}{2}m)}}{1-\frac{1}{2}m} \right]$$

Where y is the number of particles deposited per unit area, x is the distance from the source, g and e are empirical constants and m is the turbulence factor, varying between the limits of 1.24 and 1.76. The power of x to which e is raised, $1-\frac{1}{2}m$, therefore varies between the limits of 0.38 and 0.12 respectively compared with 2 in a normal distribution. The leptokurtosis produced by the lower power of x in the numerator is still further enhanced by the presence of x in the denominator. All gene dispersion depending on wind could therefore be expected to be highly leptokurtic. Such a dispersion would be characteristic of fungal spores, passively borne insects, pollen grains and wind-borne seeds, amounting in all to a large cross-section of all forms of gene dispersal.

It remains to be seen what genetic effects, if any, leptokurtic gene distribution would have. Such effects would probably arise out of the fact that, compared with a normal distribution having the same standard deviation (same overall degree of inbreeding), leptokurtosis produces more breeding between close relatives and simultaneously more breeding between distant relatives.

It is also possible that the "migration" considered by Wright as a category distinct from dispersal within the neighbourhood unit would prove to be adequately expressed as the tails of the leptokurtic distribution and therefore a part of the same gene dispersal system.

SUMMARY

A number of different forms of gene dispersal: the foraging of bees; the flight of *Drosophila* over several days; the dispersal of spores or pollen by wind; cross-pollination by bees, are all shown to give leptokurtic distributions. That is, they differ from normal distributions in having higher proportions of short and long range, as compared with medium dispersal. The implication is that many if not most methods of gene dispersal produce such distributions.

Acknowledgments.—The author is indebted to Mr J. Riley and to Miss S. Mann for assistance with the collection of the data on bee flights; to Professor Sewall Wright for comments on the manuscript and to the Agricultural Research Council for financing the work.

REFERENCES

- BATEMAN, A. J. 1947. Contamination of seed crops. III. Relation with isolation distance. *Heredity*, 1, 303-336.
- BROWNLEE, J. 1911. The mathematical theory of random migration and epidemic distribution. *P.R.S. Edinb.*, 31, 262-289.
- BUTLER, C. G. 1945. The behaviour of bees when foraging. *J. Roy. Soc. Arts*, 93, 501-511.
- DOBZHANSKY, TH., AND WRIGHT, S. 1943. Genetics of natural populations. X. Dispersion rates in *Drosophila pseudoobscura*. *Genetics*, 28, 304-340.
- DOBZHANSKY, TH., AND WRIGHT, S. 1947. Genetics of natural populations. XV. Rate of diffusion of a mutant gene through a population of *D. pseudoobscura*. *Genetics*, 32, 303-324.
- GREGORY, P. H. 1945. The dispersion of airborne spores. *Trans. Brit. mycol. Soc.*, 28, 26-72.
- HALDANE, J. B. S. 1948. The theory of a cline. *J. Genet.*, 48, 277-284.
- MALÉCOT, G. 1948. *Les mathématiques de l'hérédité*. 63 pp. Masson.
- WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics*, 16, 97-159.
- WRIGHT, S. 1943. Isolation by distance. *Genetics*, 28, 114-138.
- WRIGHT, S. 1946. Isolation by distance under diverse systems of mating. *Genetics*, 31, 39-59.