GENE FREQUENCIES IN A CLINE DETERMINED BY SELECTION AND DIFFUSION

R. A. FISHER

University of Cambridge

1. *The ecological problem.*

IN 1937 (1) the author studied the distribution of the gene ratio in the simple case of an advantageous gene advancing along a linear habitat under a constant selective advantage. In Nature situations must be more complex than in this simple model; in particular it must often occur that the selective advantage itself varies with position. The interesting case arises in which a gene enjoys a selective advantage in one part of a species' range, while in the remainder it is at a selective disadvantage. On the boundary between these regions selection is neutral between two allelomorphic genes.

Cases can be observed in practice in which there is a gradient, or cline, in the frequencies of the genotypes determined by a single factor. Generally such cases will be complicated by inequalities of topography, and by consequent irregularities in the population density, and in the gradient of selective advantage. It is to be expected also that the boundary will in general neither be straight (i.e. a great circle of the earth's surface), nor constant in position under the conditions prevailing in different years. Any model worth discussing from a theoretical standpoint will therefore be a drastically simplified one, playing the part of a basis for comparisons by which the real complexities of each situation may be critically demonstrated.

The purely genetical complication of non-recognition of genotype, due to dominance, is also ignored. This is chiefly because I should regard the occurrence of true dominance in such cases as a danger-signal suggesting the very different genetic situation of a balanced polymorphism; partly because if on more careful examination it is found that the heterozygote is recognisable this will greatly increase the value of the observations.

The problem chosen for discussion, as an extension of my work of 1937, thus differs materially from that considered by Haldane in 1948 (2). Haldane discusses the effects of a discontinuous selective intensity acting on a gene ratio obscured by dominance.

The available data, in the common case, will consist of gene frequencies observed at chosen centres of collection. From such data the primary need is to determine the neutral, or 50%, line, and the distances

TABLE I

FUNDAMENTAL TABLE, GIVING THE SOLUTION OF THE DIFFERENTIAL EQUATION (1)

x	q	x	q	x	q
.00	.5000 0000	1.00	.1004 2286	2.00	.0077 7553
.02	.4895 1717	1.02	.0962 0743	2.02	.0073 3125
.04	.4790 4235	1.04	.0921 3390	2.04	.0069 1050
.06	.4685 8348	1.06	.0881 9955	2.06	.0065 1215
.08	.4581 4852	1.08	.0844 0159	2.08	.0061 3513
.10	.4477 4533	1.10	.0807 3716	2.10	.0057 7840
.12	.4373 8169	1.12	.0772 0335	2.12	.0054 4098
.14	.4270 6529	1.14	.0737 9721	2.14	.0051 2192
.16	.4168 0369	1.16	.0705 1573	2.16	.0048 2031
.18	.4066 0430	1.18	.0673 5591	2.18	.0045 3528
.20	.3964 7438	1.20	.0643 1468	2.20	.0042 6600
.22	.3864 2102	1.22	.0613 8899	2.22	.0040 1168
.24	.3764 5110	1.24	.0585 7578	2.24	.0037 7155
.26	.3665 7130	1.26	.0558 7198	2.26	.0035 4489
.28	.3567 8807	1.28	.0532 7452	2.28	.0033 3101
.30	.3471 0763	1.30	.0507 8036	2.30	.0031 2924
.32	.3375 3596	1.32	.0483 8646	2.32	.0029 3895
.34	.3280 7874	1.34	.0460 8981	2.34	.0027 5954
.36	.3187 4142	1.36	.0438 8742	2.36	.0025 9044
.38	.3095 2916	1.38	.0417 7635	2.38	.0024 3109
.40	.3004 4681	1.40	.0397 5367	2.40	.0022 8099
.42	.2914 9895	1.42	.0378 1649	2.42	.0021 3962
.44	.2826 8985	1.44	.0359 6200	2.44	.0020 0652
.46	.2740 2349	1.46	.0341 8738	2.46	.0018 8124
.48	.2655 0351	1.48	.0324 8990	2.48	.0017 6336
.50	.2571 3327	1.50	.0308 6686	2.50	.0016 5246
.52	.2489 1582	1.52	.0293 1561	2.52	.0015 4816
.54	.2408 5390	1.54	.0278 3358	2.54	.0014 5010
.56	.2329 4992	1.56	.0264 1823	2.56	.0013 5792
.58	.2252 0602	1.58	.0250 6707	2.58	.0012 7130
.60	.2176 2403	1.60	.0237 7771	2.60	.0011 8992
.62	.2102 0546	1.62	.0225 4777	2.62	.0011 1348
.64	.2029 5155	1.64	.0213 7496	2.64	.0010 4171
.66	.1958 6327	1.66	.0202 5705	2.66	.0009 7434
.68	.1889 4129	1.68	.0191 9186	2.68	.0009 1111
.70	.1821 8602	1.70	.0181 7727	2.70	.0008 5178
.72	.1755 9759	1.72	.0172 1122	2.72	.0007 9613
.74	.1691 7592	1.74	.0162 9174	2.74	.0007 4395
.76	.1629 2064	1.76	.0154 1687	2.76	.0006 9502
.78	.1568 3117	1.78	.0145 8475	2.78	.0006 4916
.80	.1509 0672	1.80	.0137 9358	2.80	.0006 0619
.82	.1451 4627	1.82	.0130 4158	2.82	.0005 6594
.84	.1395 4859	1.84	.0123 2707	2.84	.0005 2823
.86	.1341 1227	1.86	.0116 4841	2.86	.0004 9293
.88	.1288 3573	1.88	.0110 0401	2.88	.0004 5988
.90	.1237 1721	1.90	.0103 9235	2.90	.0004 2895
.92	.1187 5479	1.92	.0098 1197	2.92	.0004 0001
.94	.1139 4640	1.94	.0092 6143	2.94	.0003 7294
.96	.1092 8985	1.96	.0087 3938	2.96	.0003 4765
.98	.1047 8282	1.98	.0082 4450	2.98	.0003 2396

TABLE I—*Continued*

FUNDAMENTAL TABLE, GIVING THE SOLUTION OF THE DIFFERENTIAL EQUATION (1)

x	q	x	q	x	q
3.00	.0003 0183	4.00	.0000 0670	5.00	.0000 0009
3.02	.0002 8115	4.02	.0000 0618	5.02	.0000 0008
3.04	.0002 6183	4.04	.0000 0570	5.04	.0000 0008
3.06	.0002 4379	4.06	.0000 0525	5.06	.0000 0007
3.08	.0002 2694	4.08	.0000 0484	5.08	.0000 0006
3.10	.0002 1121	4.10	.0000 0446	5.10	.0000 0006
3.12	.0001 9652	4.12	.0000 0410	5.12	.0000 0005
3.14	.0001 8282	4.14	.0000 0378	5.14	.0000 0005
3.16	.0001 7003	4.16	.0000 0348	5.16	.0000 0004
3.18	.0001 5811	4.18	.0000 0320	5.18	.0000 0004
3.20	.0001 4699	4.20	.0000 0295	5.20	.0000 0004
3.22	.0001 3662	4.22	.0000 0271	5.22	.0000 0003
3.24	.0001 2696	4.24	.0000 0250	5.24	.0000 0003
3.26	.0001 1795	4.26	.0000 0230	5.26	.0000 0003
3.28	.0001 0956	4.28	.0000 0211	5.28	.0000 0003
3.30	.0001 0175	4.30	.0000 0194	5.30	.0000 0002
3.32	.0000 9447	4.32	.0000 0178	5.32	.0000 0002
3.34	.0000 8769	4.34	.0000 0164	5.34	.0000 0002
3.36	.0000 8139	4.36	.0000 0151	5.36	.0000 0002
3.38	.0000 7552	4.38	.0000 0138	5.38	.0000 0002
3.40	.0000 7006	4.40	.0000 0127	5.40	.0000 0001
3.42	.0000 6498	4.42	.0000 0117	5.42	.0000 0001
3.44	.0000 6026	4.44	.0000 0107	5.44	.0000 0001
3.46	.0000 5586	4.46	.0000 0098	5.46	.0000 0001
3.48	.0000 5178	4.48	.0000 0090	5.48	.0000 0001
3.50	.0000 4799	4.50	.0000 0083	5.50	.0000 0001
3.52	.0000 4446	4.52	.0000 0076	5.52	.0000 0001
3.54	.0000 4119	4.54	.0000 0070	5.54	.0000 0001
3.56	.0000 3815	4.56	.0000 0064	5.56	.0000 0001
3.58	.0000 3532	4.58	.0000 0059	5.58	.0000 0001
3.60	.0000 3270	4.60	.0000 0054	5.60	.0000 0001
3.62	.0000 3027	4.62	.0000 0049	5.62	.0000 0001
3.64	.0000 2801	4.64	.0000 0045	5.64	.0000 0000
3.66	.0000 2591	4.66	.0000 0041	5.66	.0000 0000
3.68	.0000 2397	4.68	.0000 0038	5.68	.0000 0000
3.70	.0000 2217	4.70	.0000 0035	5.70	.0000 0000
3.72	.0000 2050	4.72	.0000 0032		
3.74	.0000 1895	4.74	.0000 0029		
3.76	.0000 1752	4.76	.0000 0027		
3.78	.0000 1619	4.78	.0000 0024		
3.80	.0000 1495	4.80	.0000 0022		
3.82	.0000 1381	4.82	.0000 0020		
3.84	.0000 1276	4.84	.0000 0019		
3.86	.0000 1178	4.86	.0000 0017		
3.88	.0000 1087	4.88	.0000 0016		
3.90	.0000 1004	4.90	.0000 0014		
3.92	.0000 0926	4.92	.0000 0013		
3.94	.0000 0855	4.94	.0000 0012		
3.96	.0000 0788	4.96	.0000 0011		
3.98	.0000 0727	4.98	.0000 0010		

from this at which other percentages are to be expected. The scale of these distances is an important observational feature, depending on the diffusional mobility of the species, and on the intensity of the selective gradient.

2. *Mathematical formulation.*

The frequency p of a gene depends only on the coordinate x, as in a linear habitat.

In the case in which the gene with frequency p has a selective advantage i (defined as rate of change of logarithmic gene ratio, (3) pp. 70–72)) proportional to x, and x varies without limit in both directions, the rate of increase due to selection is

$$\frac{dp}{dt} = pqi = pq\ g\ x,$$

where g is the gradient of selective advantage, and $q = 1 - p$.

Suppose also a uniform population density with a diffusion constant k such that the rate of increase in gene frequency (p) at any point is

$$k\frac{d^2p}{dx^2} = -k\frac{d^2q}{dx^2}$$

due to diffusion. In 1937 I discussed the limitations and justification of the analogy with physical diffusion.

Then the gene ratio will adjust itself so as to tend to satisfy the equation of equilibrium,

$$k\frac{d^2q}{dx^2} = pq\ g\ x;$$

we may choose the unit of length in which the position x is measured, so that

$$g = 4k,$$

and the relation between q and x is then given by the equation

$$\left.\begin{aligned} &\frac{d^2q}{dx^2} = 4\ x\ pq, \\ &\text{with the boundary conditions} \\ &x = 0 \qquad q = \tfrac{1}{2} \\ &x = \infty \qquad q = 0 \end{aligned}\right\} \qquad (1)$$

TABLE II

STANDARDIZED DEVIATES (LEGITS) FOR GIVEN GENE PERCENTAGES

Gene percentage	Deviate	Gene percentage	Deviate	Gene percentage	Deviate
50	.00000	20.0	.64827	5.0	1.30643
49	.01908	19.5	.66247	4.8	1.32331
48	.03817	19.0	.67691	4.6	1.34080
47	.05729	18.5	.69161	4.4	1.35896
46	.07645	18.0	.70658	4.2	1.37784
45	.09566	17.5	.72184	4.0	1.39752
44	.11493	17.0	.73740	3.8	1.41807
43	.13429	16.5	.75329	3.6	1.43958
42	.15375	16.0	.76952	3.4	1.46216
41	.17332	15.5	.78612	3.2	1.48594
40	.19302	15.0	.80311	3.0	1.51106
39	.21286	14.5	.82052	2.8	1.53771
38	.23286	14.0	.83837	2.6	1.56609
37	.25304	13.5	.85669	2.4	1.59648
36	.27341	13.0	.87553	2.2	1.62922
35	.29400	12.5	.89492	2.0	1.66474
34	.31483	12.0	.91492	1.8	1.70360
33	.33592	11.5	.93556	1.6	1.74656
32	.35729	11.0	.95691	1.4	1.79468
31	.37897	10.5	.97902	1.2	1.84951
30	.40099	10.0	1.00198	1.0	1.91340
29	.42338	9.5	1.02586	0.9	1.94988
28	.44617	9.0	1.05076	0.8	1.99029
27	.46940	8.5	1.07680	0.7	2.03565
26	.49311	8.0	1.10411	0.6	2.08745
25	.51734	7.5	1.13285	0.5	2.14795
24	.54214	7.0	1.16321	0.4	2.22095
23	.56757	6.5	1.19543	0.3	2.31345
22	.59369	6.0	1.22978	0.2	2.44101
21	.62056	5.5	1.26662	0.1	2.65223

Values of q to eight decimal places, from $x = 0$, by intervals of .02, to extinction, are given in Table I. I owe this tabulation to Dr. M. V. Wilkes and Mr. D. J. Wheeler, operating the EDSAC electronic computer. The last decimal place may be in error by 3 or 4 units. From this primary Table I have calculated Table II, giving the deviation x

TABLE III

APPARATUS FOR FINAL FITTING

Provisional Legit	Working Legit		Weighting Coeff.	Provisional Legit	Working Legit		Weighting Coeff.
	Max.	Min.			Max.	Min.	
.00	.9538	−.9538	1.0992	1.50	1.8891	−10.7156	.2104
.10	.9623	−.9636	1.0903	1.60	1.9776	−13.905	.1708
.20	.9857	−.9960	1.0643	1.70	2.0673	−18.135	.1373
.30	1.0211	−1.0564	1.0224	1.80	2.1577	−23.773	.1093
.40	1.0665	−1.1518	.9669	1.90	2.2489	−31.321	.0863
.50	1.1200	−1.2913	.9004	2.00	2.3407	−41.478	.0675
.60	1.1803	−1.4862	.8260	2.10	2.4331	−55.218	.0524
.70	1.2461	−1.7515	.7469	2.20	2.5261	−73.906	.0403
.80	1.3166	−2.1067	.6659	2.30	2.6194	−99.459	.0308
.90	1.3909	−2.5772	.5858	2.40	2.7132	−134.599	.0233
1.00	1.4685	−3.1965	.5087	2.50	2.8073	−183.17	.0175
1.10	1.5487	−4.0090	.4362	2.60	2.9018	−250.70	.0131
1.20	1.6312	−5.0736	.3697	2.70	2.9965	−345.11	.0097
1.30	1.7156	−6.4693	.3097	2.80	3.0916	−477.90	.0071
1.40	1.8017	−8.3021	.2566	2.90	3.1868	−665.39	.0052
1.50	1.8891	−10.7156	.2104	3.00	3.2822	−931.73	.0038

corresponding with 90 chosen probabilities q, and Table III, supplying the computational apparatus analogous to that used in probit analysis of mortality data in toxicology.

To appreciate this analogy q may be regarded as the probability of a variate exceeding x in a certain symmetrical distribution. x is thus a transformation of q which, under the hypothesis, is linear with—and subject to choice of units as explained above is equal to—the distance from the neutral line.

For the normal distribution with unit variance this probability tends to zero in such a way that

$$\log 1/q \div \tfrac{1}{2}x^2 \to 1$$

as x is increased. For the distribution for which

$$2x = \log p - \log q,$$

$$\frac{dp}{dx} = \frac{1}{2}\operatorname{sech}^2 x,$$

we have

$$\log 1/q \div 2x \to 1.$$

The case with which we are concerned falls between these two, for $\log 1/q$ tends at the limit to equality with $\frac{4}{3} x^{3/2}$.

The relation between the three forms may also be shown from their properties at or near the median. Here we may consider

$$-\frac{d^3p}{dx^3} \div \left(\frac{dp}{dx}\right)^3,$$

which is dimensionless, and therefore a pure measure of form. For this ratio we find

		ratio	
Normal transformation	Probit	2π	6.2832
Logarithmic transformation	Logit	2^3	8.0000
Selection-diffusion transformation	Legit	$(1.90764)^3$	6.9421

Since it has been found convenient to distinguish the deviates in the two first cases as Probits and Logits respectively, a similar term such as *Legit* may be found convenient for the standardised deviate of the distribution obtained above from a uniform cline of selection.

The dimensional nature of the quantities k and g involved in the provisional identity

$$g = 4k$$

is determined by

$$k \propto L^2T^{-1}$$

$$g \propto L^{-1}T^{-1},$$

hence

$$\frac{4k}{g} \propto L^3.$$

In setting this quantity equal to unity, therefore, we are merely adopting a unit of length appropriate to our problem, and this unit of length may now be defined as

$$a^3 = \frac{4k}{g}.$$

a will then be the distance separating the line on which q is 10.04 . . .% from the line of 50%, and that in turn from the line having 89.96 . . .% of the chosen gene.

3. *Fitting the data.*

For any observed gene-frequency the corresponding deviate can be read off from Table II. Since the primary data allow of it, five places of decimals have been given. It should, I think, only be used to three places. Gene frequencies (q) greater than 50% have corresponding negative deviates. Each empirical percentage has been based on twice as many genes as organisms have been classified; these double numbers should be multiplied by the weighting factors of Table III corresponding with the deviate obtained. No high precision of weighting is usually called for, since the data will probably contain some observations with 0 or 100%, and these cannot be included at the first stage. In all other cases the observed sample supplies a deviate, an appropriate weight, and two geographical coordinates λ and μ determining its position.

The first step is then to find a weighted regression equation,

$$X = b_0 + b_1\lambda + b_2\mu,$$

in which the coefficients b_0 , b_1 and b_2 have been adjusted so as to minimise the weighted sum of squares

$$S\{w(x - X)^2\}.$$

The (first) estimated position of the neutral line is then given by the equation,

$$b_0 + b_1\lambda + b_2\mu = 0;$$

and its distance from the parallel lines

$$b_0 + b_1\lambda + b_2\mu = \pm 1$$

provides an estimate of the scaling distance a.

No great labour need be expended on this stage of the work, the purpose of which is to provide provisional deviates X, with which to proceed. Usually indeed somewhat simplified values of b_0 , b_1 and b_2 are next substituted for those actually obtained, using values judged to be sufficiently near to those to be finally obtained.

For what will probably be the final fitting these provisional values, X, corresponding to the position of each sample observed, will be used to enter Table III. If 0% have been observed for q the working Legit has its maximum value as tabulated. Linear interpolation in x is generally sufficient. If 100%, the minimum is used, and for intermediate percentages the difference between these is divided proportionally. It is usually convenient at this stage to multiply the minimum by the actual number of genes recorded, the maximum by the number of allelomorphic

genes, and to divide the total by twice the number of organisms observed, before interpolation for the provisional value x required. The weighting coefficients corresponding with the same provisional values, multiplied by twice the observed number of organisms, give the working weights.

We now have working values for deviates, and weights, sufficiently accurate for a final determination of the regression of Legit on geographical position.

The values tabulated in Table III, in accordance with the general application of the Method of Maximum Likelihood are

$$\text{Maximum} \qquad X + Q\frac{\partial X}{\partial P}$$

$$\text{Minimum} \qquad X - P\frac{\partial X}{\partial P}$$

$$\text{Weighting coefficient} \qquad \frac{1}{PQ}\left(\frac{\partial P}{\partial X}\right)^2,$$

where P, Q stand for the probabilities corresponding to the provisional value X.

Using the analysis in this form, if s samples have been observed, with a sufficient number of both allelomorphic genes in each, the minimised value of $S\{w(x - X)^2\}$ may be taken as χ^2 with $(s - 3)$ degrees of freedom to test the goodness of fit of the hypothesis. In all respects the analogy with probit analysis is close.

The determination of the distance a provides the ratio of the diffusion coefficient k to the selective gradient. Since diffusive activity can sometimes be measured independently, the way is opened for the discussion, subject of course to complicating factors, of the selective gradient.

REFERENCES

(1) R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics, vii*, 355–369, 1937.
(2) J. B. S. Haldane. The theory of a cline. *Journal of Genetics, 48*, 277–284, 1948.
(3) R. A. Fisher. *The Genetical Theory of Natural Selection*. 70-72, 1930.