Bryan F.J. Manly
Department of Mathematics and Statistics
University of Otago
P.O. Box 56
Dunedin, New Zealand

ABSTRACT: The effect of stabilizing selection on quantitative variables is to reduce variances whilst leaving means more or less unchanged. In this paper some statistical methods for detecting and measuring stabilizing selection are discussed. A generalization of Levene's test is suggested for seeing whether one sample is significantly more variable than another. A plotting technique is described which shows graphically how selection is related to the extent to which individuals deviate from the average. Models are proposed for relating selection to the amount of deviation. Methods are illustrated on Bumpus' well-known data on moribund sparrows picked up after a severe storm.

* * *

1. INTRODUCTION

In discussions of natural selection a distinction is often made between stabilizing and directional selection. In terms of the distribution of quantitative variables within a population, the effect of stabilizing selection is to reduce variances while leaving means unchanged. This can be contrasted with directional selection, which has the effect of changing means while variances may remain fairly constant.

It was found in the early days of evolutionary studies that selection acting on natural populations within one generation is often stabilizing. For example, Bumpus (1898) compared various measurements on the survivors and non-survivors in a sample of moribund sparrows (*Passer domesticus*) picked up after a harsh storm and reached the conclusion that "the process of selective elimination is most severe with extremely variable individuals, no matter in what direction the variations may occur". Similar conclusions were reached at about the same time by Weldon (1901, 1903), Di Cesnola (1906) and Thomson *et al.* (1911) from studies of other species.

Clearly there is a need for reliable statistical techniques for detecting and measuring stabilizing selection. However, in practice deciding on these techniques is not a simple matter. Directional selection is much easier to handle, particularly when several variables are being considered together.

There are two particular problems. First, tests to compare variances before and after selection are well known to be sensitive to anomalous observations and errors in assumptions. For example, the standard F-test to compare two sample variances depends heavily on the assumption that samples are fromnormal distributions. Second, in the multivariate context the number of parameters needed to describe stabilizing selection may be excessively large. Thus with p variables, $p + 1$ parameters can be used to model directional selection. However, most models for stabilizing selection require $p(p + 3)/2 + 1$ parameters. For example, Lande and Arnold (1983) suggested the use of a quadratic regression model but were not able to fit it to all of Bumpus' (1898) data because it had 45 parameters to be estimated from the survival of only 49 female and 87 male birds. They restricted their attention to only the first two principal components in order to overcome this difficulty.

The purpose of the present paper is to suggest a practical approach for the analysis of data on stabilizing selection that circumvents these difficulties. To be more precise, a three part approach is suggested involving (a) testing to see whether survivors are significantly less variable than non-survivors, (b) plotting the data to show how the probability of survival is related to the deviation of individuals from the population mean, and (c) estimating fitness functions for

* * *

Evolutionary Theory 7: 205-217 (December, 1985)
The editors thank B. Schultz and another referee for help in evaluating this paper.
© 1985, Biology Department, The University of Chicago

which the probability of an individual surviving selection is a function of its distance from the population mean.

2. TYPES OF DATA

There are two quite different types of data that have to be considered. The first will be referred to as survival data. In this case there is an initial group of individuals that are subjected to stress so that some of them die. Variable values are known for the survivors and the non-survivors. The second type of data will be referred to as two sample data. In this case a large population is envisaged with one sample being taken from this before selection and a second sample taken of the survivors after selection.

In practice two sample data is often obtained by taking a single sample from a population and classifying the individuals as juveniles or adults. The juveniles then provide the "sample" before selection and the adults provide the "sample" after selection. This procedure is valid providing that (1) there is no genetic evolution of the characters in the population, (2) the environment does not change in any way that affects individual development of characters, (3) the effects of emigration and immigration are negligible, and (4) there are no ontogenetic changes in the characters under study between the juvenile and adult stage (Lande and Arnold, 1983).

The difference between survival and two sample data is not of great importance when it comes to testing for stabilizing selection or plotting data. The same methods can be used. However different types of fitness function are appropriate for the two types of data.

3. TESTING FOR STABILIZING SELECTION

With survival data there is an initial group of n individuals, of which n_1 die and n_2 survive. The non-survivors can then be thought of as providing sample 1 and the survivors as providing sample 2. The question to be considered in this situation is then whether there is any evidence that sample 1 is more variable than sample 2.

On the other hand, with two sample data there is a sample of size n_1 taken before selection and a sample of n_2 taken after selection. Again the question to be considered is whether sample 1 is more variable than sample 2. Hence the same tests for stabilizing selection can be used on survival and two sample data.

Suppose that p variables $X_1, X_2, \ldots, X_p$ are measured for each individual. An obvious strategy is then to test each variable separately for less variation in sample 2 than in sample 1, and also test all the variables together. For a single variable there is a wide choice of tests (Van Valen, 1978; Conover *et al.*, 1981; Schultz, 1983). Notwithstanding the advice of Van Valen (1978), most people would probably do an F-test to begin with because this test is well known and simple. This is reasonable providing that is is accepted that a significant result may occur because of non-normality of the distribution of a variable rather than because of different variances before and after selection.

Actually, it is better to use a test that is not so sensitive to non-normal data and yet reasonably powerful in detecting real differences in variances. For reasons discussed by Schultz (1983), Levene's (1960) test is a good choice providing that the variation in samples is measured from medians. Thus, let x_{ijk} denote the value of X_j for the ith individual in the sample k, and let M_{jk} be the median for this variable in the sample. Levene's test then involves transforming the data values for variable X_j to

$$y_{ijk} = |x_{ijk} - M_{jk}| \tag{1}$$

and testing to see whether the sample means

$$\bar{y}_{j1} = \sum_{i=1}^{n_1} y_{ij1}/n_1 \quad \text{and} \quad \bar{y}_{j2} = \sum_{i=1}^{n_2} y_{ij2}/n_2 \tag{2}$$

are significantly different using a t-test. The test statistic is

$$t_j = (\bar{y}_{j1} - \bar{y}_{j2})/\{s_j\sqrt{(1/n_1 + 1/n_2)}\} \quad , \tag{3}$$

where s_j^2 is the pooled within sample variance for the transformed data. Because of the nature of stabilizing selection a one sided test is needed to see whether t_j is significantly large.

A potential problem here is that if the variance of X_j is different for samples 1 and 2 then the within sample variances of the transformed data are also liable to differ. One of the assumptions of the ordinary t-test to compare two means is then not valid. However, simulation studies suggest that this is not an important problem in practice (Schultz, 1983).

When all p variables are considered together a test for more variation in sample 1 than in sample 2 is obviously more complicated. A large-sample likelihood ratio test based upon the assumption of multivariate normality is well known (Srivastava and Carter, 1983, p.333). However this test is also well known to be sensitive to the normality assumption. Furthermore, it relies on measuring variation using the determinants of the sample covariance matrices which is not altogether satisfactory (Van Valen, 1978).

Alternative procedures based upon generalizing the principle behind Levene's test are likely to be more reliable. Thus the original data values x_{ijk} can be converted to absolute deviation from sample medians using equation (1). There are then two multivariate samples of y values and stabilizing selection is indicated if the means of the y values are significantly larger in sample 1 than they are in sample 2.

Following Van Valen (1978) one possibility involves taking

$$D_{ik} = \sqrt{\left\{\sum_{j=1}^{p} y_{ijk}^2\right\}} \tag{4}$$

as the distance of individual i in group k from the median centre of that group. To ensure that all p variables contribute equally to this distance, the original X variables should all be standardized to have equal variances before the y values are determined by equation (1). Having calculated D_{ik} values, a t-test can be carried out to see whether the mean for sample 1 ($\bar{D}_1$, say) is significantly larger than the mean for sample 2 ($\bar{D}_2$, say). The test statistic is then

$$t_D = (\bar{D}_1 - \bar{D}_2)/\{s_D\sqrt{(1/n_1 + 1/n_2)}\} \quad , \tag{5}$$

where s_D^2 is the pooled within sample variance of the D values.

Another possibility is to compare the two groups of y values using a T^2 test (Srivastava and Carter, 1983, pp.47 and 53). However some results given below suggest that this will not provide a better test.

In discussing tests of significance it is worth stressing the value of randomization tests. The general principles behind these are reviewed by Edgington (1980). In the present context they involve randomly allocating the n individuals in the data to samples of size n_1 and n_2 and determining the resulting values of test statistics. By doing the randomization a large number of times the randomization distributions of the test statistics can be determined. An observed test statistic is then significant at the $\alpha\%$ level if it exceeds $(100\alpha\%)$ of the values in the corresponding randomized distribution. The advantage of this approach is that no particular distribution assumptions have to be made. It relies entirely on the observed data.

4. EXAMPLES OF TESTS

Two sets of data will be used to illustrate the testing procedures described in the previous section. As an example of survival data, Bumpus' (1898) results for female birds will be used. Here there are 28 non-survivors and 21 survivors. Table 1 shows the means and standard deviations for these two samples for the eight morphological variables measured by Bumpus and also the results of some tests on these variables. None of the mean values differ significantly between survivors and non-survivors. An F-test and Levene's test show non-survivors to be significantly more variable than survivors for the lengths of the humerus and tibio-tarsus. In addition a significant result is found for the F-test on the length of the keel of the sternum.

Table 1

Comparison between 28 non-survivors and 21 survivors for Bumpus' female data.

Meansurement	Non-survivors Mean	Non-survivors Std. dev.	Survivors Mean	Survivors Std. dev.	Test on means*	F-test on variances**	Levene's test on variances***
Total length (mm)	158.42	3.88	157.38	3.32	1.00	1.36	1.20
Alar extent (mm)	241.57	5.70	241.00	4.18	0.39	1.86	1.18
Length of beak & head (mm)	31.48	0.85	31.43	0.73	0.20	1.37	0.81
Length of humerus (inches)	0.726	0.026	0.728	0.016	-0.35	2.50†	1.91†
Length of femur (inches)	0.710	0.028	0.715	0.020	-0.69	1.97	1.49
Length of tibio-tarsus (mm)	1.132	0.048	1.144	0.030	-0.97	2.53†	1.70†
Width of skull (inches)	0.602	0.018	0.600	0.013	0.31	1.93	1.36
Length of keel of sternum (inches)	0.821	0.045	0.819	0.030	0.12	2.30†	1.46

* Test statistic shown is $t = (\bar{x}_1 - \bar{x}_2)/[s\sqrt{\{1/21 + 1/28\}}]$, where $s^2 = (20s_1^2 + 27s_2^2)/47$, the pooled variance, with 47 degrees of freedom.

** Test statistic shown is s_1^2/s_2^2 , with 27 and 20 degrees of freedom.

*** t_j from equation (3) , with significance determined by comparison with the t-distribution with 47 degrees of freedom.

† Significant at the 5% level.

Taking all eight variables together, the likelihood ratio test (assuming multivariate normality) for the equality of the covariance matrices for survivors and non-survivors (Srivastava and Carter, 1983, p.333) provides a test statistic of $X^2 = 69.37$ which has to be compared with the chi-squared distribution with 36 degrees of freedom. This is significant at the 1% level. There is no indication from the test statistic as to the directions of differences between the covariance matrices although it is clear from Table 1 that the non-survivors tend to be more variable. As mentioned before, the problem with this test is that the significant result could be due to non-normal distributions.

If the eight variables are used to calculate the t_D test statistic of equation (5) then the value obtained is 1.97 . Treated as a t statistic with 47 degrees of freedom this is significantly large at the 5% level but not the 1% level. When 1000 t_D values were generated by randomizing the data it was found that 29 of these values were 1.97 or more. Consequently, the observed t_D value is significant at about the 2.9% level on a randomization test.

When a T^2 test (Srivastava and Carter, 1983, p.47) is used to compare the two multivariate samples of y values for survivors and non-survivors a non-significant result is obtained. The reason for this result is that the T^2 test has very low power in the particular situation being considered. This is demonstrated nicely by a randomization experiment that will not be described.

The following procedure was carried out 1000 times:-

(a) A sample of 28 "non-survivors" was chosen at random from Bumpus' 49 female birds and the remaining 21 birds were taken as "survivors".
(b) Let x_{ij1} denote the value of variable X_j for the ith of the "non-survivors". This was modified to $x_{ij1} + e_{ij}$ where e_{ij} was a random variable with mean zero and standard deviation Δs_j , where s_j denotes the standard deviation of X_j for all 49 birds.
(c) The statistic t_D was calculated from equation (5) and T^2 was calculated as described by Srivastava and Carter (1983).

Values used for Δ were 0 , 0.125 , 0.25 , 0.5 and 1.0 .

With $\Delta = 0$ the original data values were unchanged. This corresponds to a situation where the null hypothesis that sample 1 and sample 2 are equally variable must be true. At the other extreme, taking $\Delta = 1$ amounted to doubling the standard deviations of all variables X_1 , X_2 ,..., X_8 in sample 1. The other Δ values provide situations somewhere between these extremes.

It is quite clear that increasing Δ should lead to an increasing proportion of significant values for t_D and T^2 . This is exactly what happens with t_D but it does not happen with T^2 . Table 2 shows the number of significant values for tests at the 5% and 1% levels of significance. The T^2 test gives very poor results. There is no doubt that the t_D test is far better for detecting stabilizing selection at least with this type of data.

As an example of two sample data, Bumpus' male data can be considered with the 28 young males being regarded as a sample before selection and the 59 adult males being regarded as a sample after selection. Table 3 shows the sample means and standard deviations and tests on individual variables. There are no significant differences between the means for young and adult birds. The F test and Levene's test show significantly more variation in the young birds for the alar extent. Levene's test also gives significantly more variation in the young birds for the total length.

The likelihood ratio test for equality of the covariance matrices before and after selection assuming multivariate normal samples (Srivastava and Carter, 1983, p.333) gives a test statistic of $X^2 = 42.91$ for the eight variables taken together. Compared to the chi-squared distribution with 36 degrees of freedom this is not significant at the 5% level. The test statistic of equation (5) is $t_D = 1.65$ which is not quite significantly large in comparison to the t-distribution with 85 degrees of freedom.

It seems that overall there is no clear evidence that the young males were more variable than the adults even though the standard deviations of the individual variables were larger for the young males for seven out of the eight variables.

Table 2

Comparison of the t_D test and the T^2 test for detecting different amounts of stabilizing selection on data generated from Bumpus' female data. The parameter Δ indicates the proportional amount by which the standard deviation is made larger for sample 1 than for sample 2 and ranges from $\Delta = 0$ (equal standard deviations) to $\Delta = 1$ (sample 1 standard deviations are twice those of sample 2). For t_D, critical values for significance have been determined from the t distribution with 47 degrees of freedom. For T^2, the usual critical values based on the F distribution are not satisfactory. For this test the critical values have therefore been chosen so as to give the correct proportions of significant results with $\Delta = 0$.

	Percentage of significant results on the t_D test		Percentage of significant results on the T test	
Δ	5% test	1% test	5% test	1% test
0	4.5	1.1	5.0	1.0
0.125	5.4	1.2	5.7	0.9
0.25	8.6	2.1	6.7	0.7
0.5	28.0	9.3	7.0	1.5
1.0	97.0	84.4	41.8	18.3

* *

A randomization experiment was carried out with the male data using the procedure described above for the female data. Details will not be given since the results obtained were essentially the same for the males as for the females (Table 2).

To sum up the two examples, it appears that stabilizing selection did take place with the storm survival of the females. However there is no real evidence of stabilizing selection on the males between the juvenile and adult ages.

5. FURTHER POWER COMPARISONS OF TESTS

More extensive comparisons of the power of the t_D, T^2 and likelihood ratio tests are provided elsewhere (Manly, 1985) based upon simulating samples from multivariate normal and non-normal distributions. These comparisons show that the T^2 test is generally inferior to the t_D test for detecting stabilizing selection. The t_D test is also superior to the standard likelihood ratio test for detecting this type of effect unless the variables being tested have correlations of the order of $+0.9$.

6. PLOTTING DATA

It is obviously useful to have some way of plotting data to indicate the manner in which selection depends upon the extent to which individual are abnormal. A technique proposed by Copas (1983) is useful in this respect.

As before it will be assumed that there are two samples, of sizes n_1 and n_2, the first of which is expected to be more variable when stabilizing selection takes place. The extent to which an individual is abnormal can conveniently be measured by its Euclidean distance from the mean for all $n = n_1 + n_2$ individuals in terms of variables standardized to have a variance of unity. Standardization is required to ensure that all p variables contribute equally to the distance. The abnormality of individual i in sample k is then

$$d_{ik} = \sqrt{\left\{ \sum_{j=1}^{p} (x_{ijk} - \bar{x}_j)^2 / s_j^2 \right\}} , \qquad (6)$$

where $\bar{x}_j$ is the mean and s_j^2 is the variance of X_j for all n individuals. Other measures of distance could obviously be used. However, the Euclidean distance is relatively simple and it is frequently used in other contexts.

Copas' technique involves plotting the function

Table 3

Comparison between 28 young males and 59 adult males for Bumpus' data.

Measurement	Young Mean	Young Std. dev.	Adults Mean	Adults Std. dev.	Test on means*	F-test on variances**	Levene's test on variances***
Total length (mm)	160.79	3.61	160.25	3.01	0.73	1.44	1.84†
Alar extent (mm)	247.29	5.76	247.56	3.69	-0.26	2.44†	2.45†
Length of beak & head (mm)	31.64	0.71	31.64	0.62	0.00	1.33	0.67
Length of humerus (inches)	0.739	0.026	0.734	0.022	0.93	1.45	0.93
Length of femur (inches)	0.716	0.027	0.712	0.022	0.74	1.50	0.38
Length of tibio-tarsus (mm)	1.138	0.048	1.129	0.037	0.96	1.66	0.96
Width of skull (inches)	0.604	0.017	0.603	0.013	0.30	1.58	1.33
Length of keel of sternum (inches)	0.848	0.035	0.853	0.036	-0.61	0.98	-0.51

* Test statistic is $t = (\bar{x}_1 - \bar{x}_2)/[\, s\sqrt{\{1/28 + 1/59\}}]$, where $s^2 = (27s_1^2 + 58s_2^2)/85$, the pooled variance with 85 degrees of freedom.

** s_1^2/s_2^2 with 27 and 58 degrees of freedom.

*** t_j from equation (3) with significance determined by comparison with the t-distribution with 85 degrees of freedom.

† Significant at the 5% level.

$$\hat{P}(d) = \sum_{i=1}^{n_2} \exp\left\{-\frac{(d - d_{i2})^2}{2h^2}\right\} \Big/ \sum_{k=1}^{2} \sum_{i=1}^{n_k} \exp\left\{-\frac{(d - d_{ik})^2}{2h^2}\right\} \qquad (7)$$

against d, over the range of values of d in the data. This may appear a strange procedure at first but it becomes more understandable when it is noted that the numerator of $\hat{P}(d)$ receives a contribution from each individual in sample 2 while the denominator receives a contribution from the individuals in both samples. Also, the contributions to $\hat{P}(d)$ are greatest from the individuals with distances close to d. Any individual with a distance equal to d will contribute $+1$, this being the maximum possible contribution. Individuals with distances very different to d will make almost no contribution. From the way it is calculated, $\hat{P}(d)$ can be thought of as an estimate of the probability of an individual with a distance d being in sample 2, given that it is in one of the samples.

The parameter h in equation (7) is a smoothing constant that is at choice. If h is small then $\hat{P}(d)$ has important contributions only from individuals with distances close to d. On the other hand, a large value of h allows individuals with distances rather different from d to have an effect. Copas suggests trying a range of values of h, starting with about ten times the average spacing between d values in the data.

See Copas' (1983) paper for a more rigorous justification of his method of plotting data. An approximate equation for the standard error of $\hat{P}(d)$ can be obtained from the same source.

* * * * * * * * * * * * * * * * * * * *

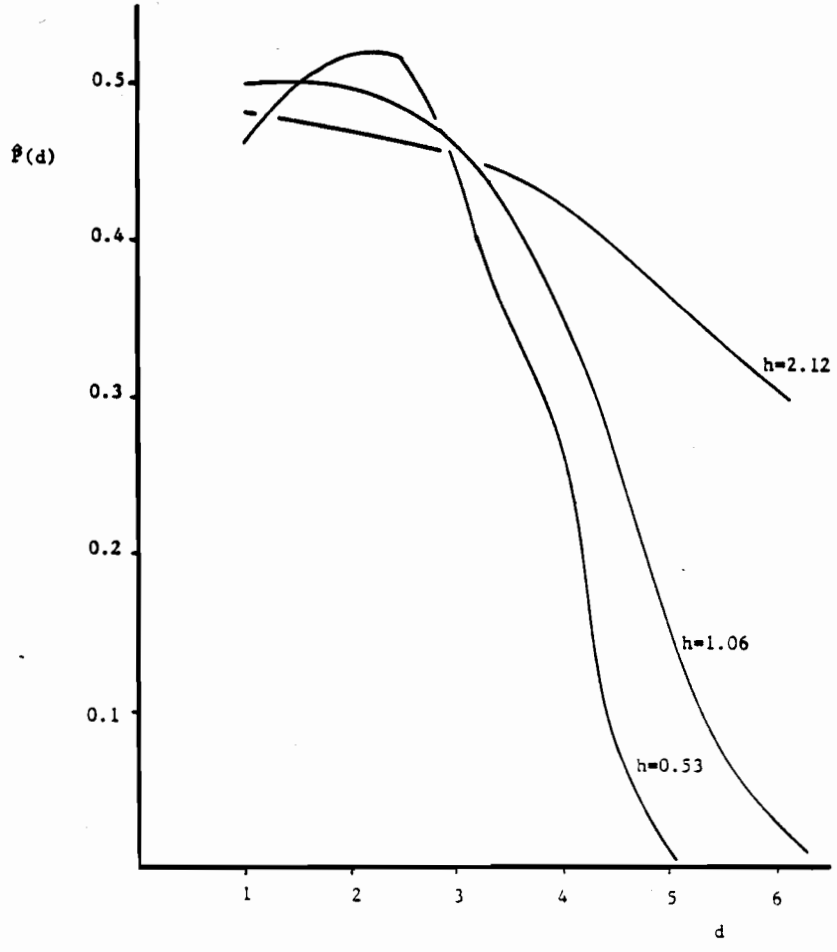

Figure 1. Plot of $\hat{P}(d)$, the estimated probability of survival for a female sparrow distance d from the mean, with three values of the smoothing constant h.

* * * * * * * * * * * * * * * * * * * *

For Bumpus' female data the first sample consists of 28 non-survivors and the second sample consists of 21 survivors. In this case $\hat{P}(d)$ will provide an estimate of the probability of survival for birds with a Euclidean distance of d from the mean bird. The distances in the data range from 1.01 to 6.09 . Ten times the average spacing between these is therefore $10(6.09 - 1.01)/48 = 1.06$. A suitable initial choice for h is therefore 1.06 . Figure 1 shows plots for $h = 0.53$, 1.06 and 2.12 .

The effect of varying the smoothing constant is very pronounced in this example. However it must be remembered that $\hat{P}(d)$ is subject to sampling variation. Even with $h = 2.12$ the standard error of $\hat{P}(d)$ as calculated from the equation given by Copas (1983) is about 0.2 . Under the circumstances it seems best to regard $h = 2.12$ as providing the best plot of the data. It then appears that the survival probability of the birds varied from about 0.5 for "average" birds to 0.3 for the most unusual birds.

For Bumpus' male data, the first sample consists of 28 juveniles and the second sample consists of 59 adults. A plot using equation (7) then indicates how the proportion of adults varies with different values of d . Stabilizing selection is indicated if $\hat{P}(d)$ decreases with increasing d . For this data the minimum value of d is 1.03 and the maximum is 6.10 . Ten times the average spacing between d values is therefore $10(6.10 - 1.03)/86 = 0.59$. Figure 2 shows the plots of $\hat{P}(d)$ for h equal to 0.59 and also half and twice this value.

* * * * * * * * * * * * * * * * * * * *

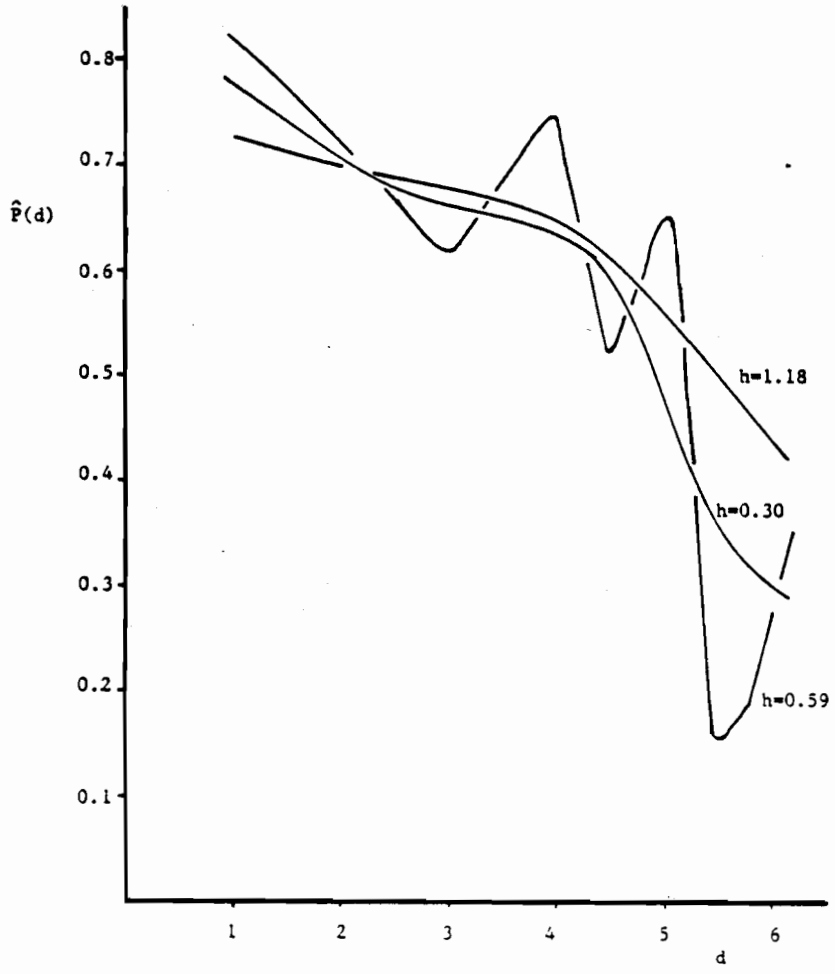

Figure 2. Plot of $\hat{P}(d)$, the estimated proportion of adult birds for males distance d from the mean, with three values for the smoothing constant h .

* * * * * * * * * * * * * * * * * * * *

The plot with $h = 0.30$ is very erratic, indicating a clear lack of sufficient smoothing of the data. Even with $h = 1.18$ the standard error of $P(d)$ values is about 0.2. It therefore seems best to accept this value of h as best. It then appears that the proportion of adults was about 0.7 for "average" birds but declined to about 0.4 for the most unusual birds. This indication of stabilizing selection must be treated with some reservations because of the fact that the tests for stabilizing selection described above give non-significant results with these data.

7. MODELS FOR STABILIZING SELECTION FOR SURVIVAL DATA

For reasons discussed by Manly (1976), a function of the form

$$\phi = \exp\{-\exp(\beta_0 + \beta_1 d)\} \tag{8}$$

in appropriate for representing the relationship between ϕ, the probability of surviving selection, and d, the extent to which an individual deviates from the mean phenotype. For stabilizing selection β_1 must be positive so that large values of d correspond to low values of ϕ.

If $\beta_1 = 0$ then the relationship reduces to

$$\phi = \exp\{-\exp(\beta_0)\} . \tag{9}$$

This model, which will be referred to as *Model 0*, gives the probability of survival to be the same for all individuals. The maximum likelihood estimator of β_0 is

$$\hat{\beta}_0 = \log\{-\log(n_2/n)\} , \tag{10}$$

where n_2/n is the observed proportion dying.

A simple measure for d is the Euclidean distance function (6) that has already been suggested for use with Copas' plotting technique. The model of equation (8) with d determined in this way will be referred to here as *Model 1*. It can be estimated from data by the principle of maximum likelihood using the computer program GLIM (Nelder, 1974; Dobson, 1983, Chapter 8).

A great advantage of *Model 1* over other models for stabilizing selection is that it only needs the single parameter β_1 to describe this selection. However, the price paid for this advantage is a model that may be unrealistic because it treats deviations from means as being equivalent for all variables. Another model which may be more realistic whilst still involving an acceptable number of parameters is

$$\phi = \exp\{-\exp(\beta_0 + \sum_{j=1}^{p} \beta_j d_j)\} , \tag{11}$$

where $d_j = \sqrt{\{(x_j - \bar{x}_j)^2/s_j^2\}}$ is the standardised deviation from the mean for variable X_j for an individual with the value x_j for this variable. This model allows deviations from the mean to have different effects for different variables. It will be referred to as *Model 2*. Like *Model 1*, it can be estimated from data using GLIM.

If GLIM is used to fit *Models 0*, *1* and *2* then the relative goodness of fit of the models can be compared in terms of their deviances. If the deviances are D_0, D_1 and D_2, respectively, then $D_0 - D_1$ can be compared to the chi-squared distribution with one degree of freedom. A significantly large value is evidence of stabilizing selection. Also $D_1 - D_2$ can be compared to the chi-square distribution with $p - 1$ degrees of freedom. A significantly large value is evidence of a different effect for deviations from the mean for different variables.

Another plausible model is *Model 3*, for which

$$\phi = \exp\left[-\exp[\beta_0 + \beta_1 \sqrt{\{\sum_{j=1}^{p} (x_j - \alpha_j)^2/s_j^2\}}]\right] . \tag{12}$$

Here the optimum value of X_j is α_j rather than the overall observed mean, where α_j has to be estimated from the data. Fitting this model may present some problems. It does not fall within the scope of GLIM. However a general maximum likelihood estimation computer program could be used to fit it.

Finally, it is worth noting that the most general model for stabilizing selection that could be entertained would be one something like

$$\phi = \exp\{-\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{i=1}^{p}\sum_{j=1}^{i} \beta_{ij} x_i x_j)\} . \qquad (13)$$

This allows any optimum values for the variables X_1, X_2, ..., X_p and varying effects for deviations from the optimum for each of the variables. Unfortunately, the number of parameters, $1 + p(p + 3)/2$, makes this model impossible to fit unless either p is small or there is a large amount of data.

To illustrate the use of *Models 0*, *1* and *2*, consider yet again Bumpus' female data. For these data the no selection *Model 0* of equation (9) is estimated as

$$\phi = \exp\{-\exp(-0.1657)\} = 0.4286 ,$$

which gives a GLIM deviation of $D_0 = 66.92$. *Model 1* of equation (8) is estimated as

$$\phi = \exp\{-\exp(-1.249 + 0.430d)\} , \qquad (14)$$

with a deviance of $D_1 = 62.00$. The difference in deviance, $D_0 - D_1 = 4.92$, with one degree of freedom, is significantly large at the 5% level when compared to the chi-squared distribution. *Model 1* is therefore a significantly better fit than *Model 0*.

Model 2 was also fitted to the data using GLIM. The deviance was found to be $D_2 = 59.15$. This is hardly any smaller than the deviance of *Model 1*. The difference in deviance, $D_2 - D_1 = 2.85$, with seven degrees of freedom, is not at all significant. It seems therefore that equation (14) best describes the stabilizing selection that seems to have taken place. According to this model, an average female, with $d = 0$, had a probability of about $\phi = 0.75$ of surviving the storm, while the most unusual female, with $d = 6.09$, had a survival probability of only 0.02.

8. MODELS FOR STABILIZING SELECTION FOR TWO SAMPLE DATA.

One approach to modelling two sample data involves assuming that the distribution of the distance of individuals from the mean is $f(d)$ before selection and $\phi(d)f(d)$ after selection. Then $\phi(d)$ is a fitness function which reflects the increase or decrease in the frequency of deviations of d as a result of selection. In that case the probability of an individual with deviation d appearing in the sample before selection, given that it is either in this sample or in the sample after selection, is

$$\begin{aligned} \text{Prob(sample 1}|d) &= f(d)/\{f(d) + \phi(d)f(d)\} \\ &= 1/\{1 + \phi(d)\} , \end{aligned}$$

which does not depend on $f(d)$. A simple assumption is that $\phi(d) = \exp(\beta_0 + \beta_1 d)$, in which case

$$\text{Prob(sample 1}|d) = 1/\{1 + \exp(\beta_0 + \beta_1 d)\} . \qquad (15)$$

The same approach can now be used as suggested in the previous section for survival data. Thus three models in particular can be entertained. *Model 0* is the no selection model for which $\phi(d) = \exp(\beta_0)$ so that

$$\text{Prob(sample 1}|d) = 1/\{1 + \exp(\beta_0)\} . \qquad (16)$$

The maximum likelihood estimator of β_0 can then be shown to be $\hat{\beta}_0 = \log_e(n_2/n_1)$, where n_i is the size of the ith sample. *Model 1* allows β_1 of equation (15) to be non-zero. It should be negative for stabilizing selection.

Finally, for *Model 2*, $\phi(d) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j d_j)$, where $d_j = \sqrt{\{(x_j - \bar{x}_j)^2/s_j^2\}}$, the standardized deviation from the mean for variable X_j. This leads to

$$\text{Prob(sample 1}|d) = 1/\{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j d_j)\} . \qquad (17)$$

All these models can be fitted to data using the logistic fitting capacity of the computer GLIM. If the deviance of *Model i* is D_i then $D_0 - D_1$, with one

degree of freedom, measures the improvement in fit of *Model 1* over *Model 0*. Similarly, $D_1 - D_2$, with $p - 1$ degrees of freedom, measures the improvement of fit of *Model 2* over *Model 1*. Significant improvements can be tested for using the chi-squared distribution.

Another plausible model is *Model 3*, for which

$$\phi = \exp[\beta_0 + \beta_1 \sqrt{\{\sum_{j=1}^{p} (x_j - \alpha_j)^2/s_j^2\}}] \ , \qquad (18)$$

where α_j is the optimum value of the variable X_j , which has to be estimated from the data. Fitting this model is not straightforward. It does not fall within the usual scope of GLIM. However, it can be fitted using a general maximum likelihood estimation computer program.

Finally, note the completely general exponential quadratic model for stabilizing selection, for which

$$\phi = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{i=1}^{p} \sum_{j=1}^{i} \beta_{ij} x_i x_j) \ .$$

Here $p(p + 3)/2$ parameters are required to describe selection.

The example that is being used for two sample data involves the 28 juvenile and 59 adult males in Bumpus' data. Tests have already shown that there is no real evidence of stabilizing selection in this case. Nevertheless, *Models 0 , 1* and *2* have been fitted to the data for illustrative purposes.

The no selection *Model 0* is estimated as

$$\phi(d) = \exp(0.7453) \ ,$$

with a deviance of $D_0 = 109.3$. *Model 1* is estimated as

$$\phi(d) = \exp(1.2258 - 0.0584d) \ ,$$

with a deviance of $D_1 = 105.4$. The difference in deviance is $D_0 - D_1 = 3.9$, with one degree of freedom. As expected, this is not significant.

When *Model 2* is fitted the deviance is $D_2 = 100.9$. The difference $D_1 - D_2 = 4.5$, with seven degrees of freedom, is not significant. There is therefore no evidence of any improvement over *Model 1*.

The model fitting calculations have confirmed the test results. A no selection model describes these data about as well as either of the selection models.

9. DISCUSSION

The approaches to studying stabilizing selection that have been proposed in this paper have a number of limitations. To begin with, the test procedure based upon equations (4) and (5) is somewhat arbitrary. However the test statistic does emphasize the sample differences that are important. The randomization results shown in Table 2 indicate that the test does have reasonable power for the types of data likely to occur in practice and this is confirmed by simulation studies reported elsewhere (Manly, 1985). Also, accurate critical values are available from the t-distribution. Thus the test is simple and reliable.

For plotting data it has been suggested that the deviation of individuals from average be measured by the standardized Euclidean distance function of equation (6). This implies that deviations from the mean are equally important for all variables when they are measured in units of standard deviations. This is unlikely to be exactly true. Nevertheless, it may serve as a useful approximation. There seems no point in using a more complicated distance function that takes into account the correlation between variables such as, for example, the Mahalanobis distance (Srivastava and Carter, 1983, p.232). A selection process may select against extreme individuals but it is difficult to see how it could allow for correlations.

Models for which optimum character values are not equal to means have been suggested (equations (12) and (18)) but fitting procedures have not been discussed. These models have, in fact, been estimated for Bumpus' data and they fit very little better than the other models discussed in Sections 7 and 8. Of course, if optimum values are not equal to means then directional selection will occur as well as stabilizing selection. It may then be appropriate to adopt one of the estimation procedures proposed by Lande and Arnold (1983) and Manly (1976, 1981).

REFERENCES

Bumpus, H.C. (1898). The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. Biological Lectures, Woods Hole Marine Biol. Lab., 11th Lecture, pp.209-226.

Conover, W.J., Johnson, M.E., and Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23: 351-361.

Copas, J.B. (1983). Plotting p against x . *Appl. Statist.* 32: 25-31.

Dobson, A. (1983). Introduction to Statistical Modellings. *Chapman and Hall*, London.

Di Cesnola, A.P. (1906). A first study of natural selection in *Helix arbustorum* (Helicogena). *Biometrika* 5: 387-399.

Edgington, E.S. (1980). *Randomization Tests*. Marcel Dekker, Inc., New York.

Lande, R. and Arnold, S.J. (1983). The measurement of selection on correlated characters. *Evolution* 37: 1210-1226.

Leven, H. (1960). Robust tests for equality of variance. In *Contributions to Probability and Statistics*. (Eds. I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann), pp.278-292. Stanford Univ. Press, Palo Alto, California.

Manly, B.F.J. (1976). Some examples of double exponential fitness functions. *Heredity* 26: 229-234.

Manly, B.F.J. (1981). The estimation of a multivariate fitness function from several samples taken from a population. *Biom. J.* 23: 267-281.

Manly, B.F.J. (1985). Stabilizing selection. In *Proceedings of the Pacific Statistical Congress - 1985*. (Eds. I.S. Francis, F.C. Lam and B.F.J. Manly) North-Holland, in press.

Nelder, J.A. (1974). *Glim Manual*. Numerical Algorithms Group, Banbury Road, Oxford.

Schultz, B. (1983). On Levene's test and other statistics of variation. *Evol. Theory*. 6: 197-203.

Srivastava, M.S. and Carter, E.M. (1983). *An Introduction to Applied Multivariate Statistics*. North-Holland, New York.

Thomson, E.Y., Bell, J., and Pearson, K. (1911). A third cooperative study of *Vespa vulgaris*. Comparison of queens of a single nest with queens of the general autumn population. *Biometrika*. 8: 1-12.

Van Valen, L. (1978). The statistics of variation. *Evol. Theory* 4: 33-43. (Erratum *Evol. Theory* 4: 202).

Weldon, W.F.R. (1901). A first study of natural selection of *Clausilia laminatu* (Montagu). *Biometrika* 1: 109-124.

Weldon, W.F.R. (1903). Note on a race of *Clausilia itala* (Van Marteus). *Biometrika* 3: 299-307.