Theoretical and Applied Genetics 43, 204—212 (1973)

The Population Genetics of Parthenogenetic Strains of *Drosophila mercatorum*

I. One Locus Model and Statistics[1,2]

ALAN R. TEMPLETON and EDWARD D. ROTHMAN

Department of Human Genetics and Department of Statistics, The University of Michigan, Ann Arbor, Michigan (USA)

Summary. A one locus model has been developed to describe parthenogenetic populations restoring diploidy by central fusion, terminal fusion and gamete duplication. It was found that in the absence of selection all populations become homozygous. With selection, however, it is possible to maintain heterozygotes and homozygotes. The conditions required to yield such an equilibrium are a function of (1) the proportions of the various diploid restoring mechanisms (2) linkage to the kinetochore and (3) the intensity of selection. The model was then used to derive one-generation likelihood functions. These likelihoods were used in deriving estimation procedures for the frequency of gamete duplication which is important in forming isogenic lines and for the probability of a heterozygous female giving rise to a heterozygous zygoid. Next, n-generation likelihood functions with and without selection were calculated. These were used to estimate the selection coefficient and to derive two tests of the hypothesis of no selection versus the hypothesis of selection. The first test is a locally best test in the vicinity of no selection, and the second an "odds" for the hypotheses using a prior distribution on the selection coefficient.

Parthenogenesis has often been termed an evolutionary dead end mainly because many authors felt parthenogenetic populations could not maintain genetic variability. Carson (1967a) challenged this idea when he stated that central fusion of the meiotic pronuclei and absolute linkage to the kinetochore would result in a state of permanent heterozygosity. Asher (1970a) studied this problem in more detail and worked out a deterministic one locus model which showed that a population restoring diploidy with both central and terminal fusion could maintain heterozygosity under a wide range of heterotic values. Asher (1970b) also simulated such populations on the computer to take into account the effect of small population size and found once again that heterozygotes could be maintained.

The theoretical predictions are interesting for they suggest that parthenogenetic species may be useful in solving the more general problem of the role of selection in maintaining genetic variation. Many parthenogenetic species offer an excellent strategy for studying this problem since they share with sexual species segregation of alleles, recombination, and the ability to respond to selection but lack the complex population structure which in sexual populations is superimposed upon this basic genetic foundation. The potential therefore exists to make strong inference about the role of selection in this system since many confounding parameters are eliminated. For this potential to be realized it is very important to couple the theory with reality. A model is useful not only in generating concepts and principles, but also in predicting reality and providing a means of testing hypothesis. All too often in population genetics these latter two uses are neglected. The purpose of this paper is therefore a) to describe a one locus model for the type of parthenogenesis characteristic of the fruit fly *Drosophila mercatorum*, b) to investigate estimation procedures for the various parameters of the model and c) to obtain tests of the hypothesis of selection for heterozygotes versus no selection.

One Locus Model

The experiments of Carson *et al.* (1969) on parthenogenetic strains of *Drosophila mercatorum* indicate that meiosis proceeds normally in the eggs of virgin females producing four haploid egg nuclei. Diploidy appears to be restored in more than 90% of the eggs by the post-meiotic doubling of a single haploid egg nucleus. Carson (personal communication) has termed this phenomenon "gamete duplication". In the remaining eggs diploidy is probably restored by central and/or terminal fusion of two of the four meiotic products.

The genetic consequences of the three fusion mechanisms upon a single locus with two alleles (A and a) will now be considered. First, if the adult is homozygous AA or aa it will produce in the absence of

[1] The experimental work was supported by AEC Contract At (11-1)-1552 awarded to Charles F. Sing, Department of Human Genetics, the University of Michigan.

[2] This work was carried out while Templeton was a recipient of an NSF predoctoral fellowship.

mutation only homozygotes no matter which mechanism is used to restore diploidy. However, if the adult is heterozygous Aa the situation is more complex. If gamete duplication occurs, diploidy is restored by the fusion of two genetically identical nuclei so that all loci of the zygoid are homozygous. In the absence of selection gamete duplication in eggs from a heterozygous mother will produce half AA zygoids and half aa zygoids. The consequences of central and terminal fusion have been worked out by Asher (1970a). He has shown that if Y is the probability of recombination between locus A and the kinetochore, then central fusion in eggs from a heterozygous mother produces heterozygotes with frequency $1-Y/2$ and each homozygote type with frequency $Y/4$, while terminal fusion produces heterozygotes with frequency Y and each homozygote type with frequency $(1-Y)/2$.

Letting

E_1 = the proportion of eggs developing by terminal fusion

E_2 = the proportion of eggs developing by central fusion

E_3 = the proportion of eggs developing by gamete duplication

$E_1 + E_2 + E_3 = 1$

a heterozygous female in the absence of selection and mutation will produce Aa zygoids with probability $E_1 Y + E_2 (1 - Y/2) = K$, and AA and aa zygoids each with probability $E_1 (1 - Y)/2 + E_2 Y/4 + E_3/2 = \frac{1}{2}(1 - K)$.

The effect of selection will be determined by assuming the fitness values of AA, Aa and aa are W_{AA}, W_{Aa} and W_{aa}. Letting

P_n = the frequency of AA at generation n
Q_n = the frequency of aa at generation n
R_n = the frequency of Aa at generation n
$R_n + P_n + Q_n = 1$

then

$$P_{n+1} \propto \left(P_n + \frac{1}{2}(1 - K)\ R_n\right) W_{AA}$$

$$Q_{n+1} \propto \left(Q_n + \frac{1}{2}(1 - K)\ R_n\right) W_{aa}$$

$$R_{n+1} \propto K\ R_n\ W_{Aa}\ .$$

At equilibrium the condition $P_{n+1}/P_n = Q_{n+1}/Q_n = R_{n+1}/R_n$ holds. This yields

$$P_{eq} = \frac{R_{eq}\ W_{AA}\ (1 - K)}{2\ (K\ W_{Aa} - W_{AA})}$$

$$Q_{eq} = \frac{R_{eq}\ W_{aa}\ (1 - K)}{2\ (K\ W_{Aa} - W_{aa})}$$

$$R_{eq} = \frac{1}{\frac{(1 - K)\ W_{AA}}{2\ (K\ W_{Aa} - W_{AA})} + \frac{(1 - K)\ W_{aa}}{2\ (K\ W_{Aa} - W_{aa})} + 1}\ .$$

The frequency of heterozygotes at equilibrium is a function of (1) the fitnesses of the genotypes, (2) the proportions of eggs developing by the various fusion mechanisms, and (3) the probability of recombination between the locus and its kinetochore. R_{eq} is greater than zero when $W_{AA} < W_{Aa}\,K$ and $W_{aa} < W_{Aa}\,K$. When the fitness of either or both homozygotes approaches the value of $W_{Aa}\,K$ from zero, R_{eq} approaches zero.

Equilibrium phase diagrams can be used to describe the genetic structure of equilibrium populations for varying values of W_{AA} and W_{aa} with $W_{Aa} = 1$ and for a constant value of K. For a given value of W_{AA} and W_{aa} the diagram indicates whether an equilibrium population will be completely homozygous (A and B Fig. 1) or will sustain heterozygosity (C of Fig. 1). The boundaries separating the three areas are dependent upon the value of K which in turn is a function of E_1, E_2, E_3 and Y.

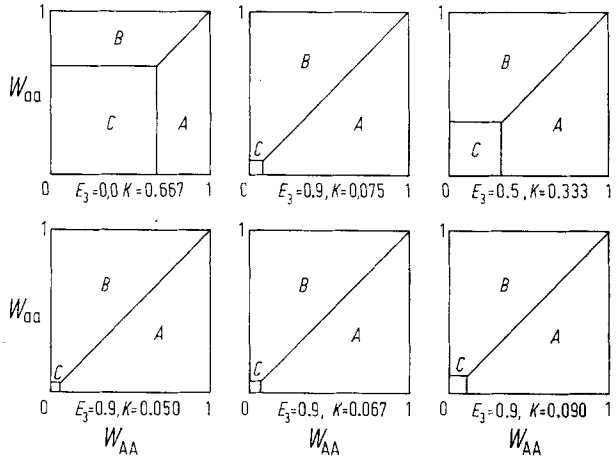

Fig. 1. Equilibrium phase diagrams. In the diagrams in the left-hand column $K = 2/3\,(1 - E_3)$ which corresponds to a locus so loosely linked to the centromere that $Y = 2/3$, or for any locus when $E_1 = 1/3\,(1 - E_3)$. The top diagram of the right-hand column has $E_1 = 0$ and $Y = .50$, the middle diagram $E_1 = 0.1$ and $Y = .50$, and the bottom $E = 0$ and $Y = .20$. Area A indicates populations that are totally homozygous AA; B—homozygous aa; C—with heterozygotes

The boundaries between populations which are completely homozygous (A and B) and those maintaining heterozygotes (C) are given by

$$W_{AA} = K$$
$$W_{aa} = K\ .$$

Within those populations that are completely homozygous the boundary between those all homozygous $AA(A)$ and those all $aa(B)$ is given by

$$W_{AA} = W_{aa} \text{ where } W_{AA} > K \text{ and } W_{aa} > K$$

Since K determines the boundary between equilibrium populations maintaining heterozygotes and those consisting entirely of homozygotes, the evolutionary impact of the fusion mechanisms and linkage to the centromere can be evaluated in terms

of their effect on K. Factors which increase K enhance the maintenance of heterozygotes in the sense that the conditions for $R_{eq} > 0$ are broader for a large K than for a small K. Similarly factors that diminish the value of K hinder maintenance of heterozygotes since in general stronger selection is required to yield $R_{eq} > 0$.

Noting that $K = E_1\ Y + E_2\ (1 - Y/2)$ and $E_1 + E_2 + E_3 = 1$, we see that for fixed Y and E_1, K decreases as E_3 increases. Thus, the presence of gamete duplication puts an added stress on the system in maintaining heterozygotes. When E_3 and Y are fixed and $0 < Y < 2/3$, K increases as the proportion of central fusion increases. Since $0 < Y < 2/3$, central fusion in general enhances maintenance of heterozygotes while terminal fusion hinders it. When $Y = 2/3$, $K = 2\ (1 - E_3)/3$ for all values of E_1 and E_2. Thus for loci that are randomly recombining with their kinetochore ($Y = 2/3$) the proportion of central and terminal fusion has no effect on maintenance of heterozygotes. Similarly, when $E_1 = (1 - E_3)/3$ (random fusion of the pronuclei in those eggs undergoing central of terminal fusion) $K = 2\ (1 - E_3)/3$ so Y has no effect on maintenance of heterozygotes. Finally, for fixed E_3 and E_1 with $E_1 > (1 - E_3)/3$, K increases as Y increases, but for $E_1 < (1 - E_3)/3$ K decreases as Y increases. Thus, when the proportion of terminal to central fusion is greater than 1/2 loci loosely linked to the kinetochore have larger K's than closely linked loci, but when central fusion is more than twice as frequent as terminal fusion the closely linked loci have the larger K's. The extreme case of this is when a locus is absolutely linked to its kinetochore and diploidy is always restored by central fusion. As Carson (1967a) has noted, these conditions can yield a state of permanent heterozygosity; i.e. $K = 1$.

K can also be thought of as an inherent decay rate of heterozygotes in the absence of selection when selective forces are large enough to yield $R_{eq} > 0$, the decay of heterozygotes is gradually dampened until it is zero. It is important to note that even if selection is not strong enough to result in $R_{eq} > 0$ the decay of heterozygotes will be slower than the decay predicted by K alone.

Estimation

To use the model to predict reality and to test hypotheses it is first necessary to estimate the critical parameters of the model. A frequent difficulty in estimation is that there are more parameters than observations. In some cases it is possible to use prior knowledge about the system to eliminate some parameters; in other cases a composite parameter that is estimable directly is all that is actually needed. The following examples will illustrate these points.

One parameter in the model that is of major interest is E_3, the probability of gamete duplication. Since gamete duplication results in total homozygosity, isogenic lines in which all individuals are genetically identical can be formed. For example, the probability that a stock of parthenogenetic *Drosophila mercatorum* formed after n single female generations is not isogenic is $(1 - E_3)^n$. Thus a knowledge of E_3 is useful in forming isogenic stocks.

As demonstrated by Carson *et al.* (1969) insight into the amount of gamete duplication can be gained by observing the proportions of heterozygous and homozygous offspring from virgin heterozygous females. They used bridge stocks (see Carson, 1967b) to obtain parthenogenetic females heterozygous for one or more visible, recessive markers. These females reproduced parthenogenetically and those offspring which were potentially heterozygous at one or more loci (i.e., offspring showing the recessive phenotype for all markers were excluded) were tested. Heterozygotes at one or more of the marker loci would have to come from central or terminal fusions, but homozygotes for all markers could be produced by all three mechanisms. Hence E_3 was estimated by taking the proportion of homozygotes for all markers minus a reasonable guess as to how many the homozygotes were actually produced by central or terminal fusion. Depending upon the number of possible heterozygous markers, 1% to 6.2% of the tested offspring were heterozygous for at least one marker. On this basis they concluded that over 90% of the eggs developed by gamete duplication.

It is possible to use the model presented here and other facts that are known about genetic systems to refine the estimate of E_3 using a design similar to the above.

To estimate E_3 in the parthenogenetic stock *S-l-Im* (see Carson, 1967b), a bridge stock Br_7-*S-v pm vl* was formed that is genetically very similar to *S-l-Im* except for three unlinked autosomal markers (vermillion, plum and veinless) and loci closely linked to them. *S-l-Im* females were mated to bridge males to produce heterozygous females which are then allowed to reproduce parthenogenetically. All these parthenogenetic offspring are then scored for homozygosity at all three loci.

The probability that a fly produced by a heterozygous female will be homozygous for all three unlinked markers is

$$E_3 + E_1\ (1 - Y_1)\ (1 - Y_2)\ (1 - Y_3) + \\ + E_2\ Y_1\ Y_2\ Y_3/8 = G'$$

where Y_1, Y_2 and Y_3 are the probabilities of recombination with the kinetochore for the three loci. Letting N be the total number of offspring and X the observed number of triple homozygotes, then the likelihood of X is

$$f(X\colon Y_1, Y_2, Y_3, E_1, E_3) = \binom{N}{X} (G')^X\ (1 - G')^{N-X}\ .$$

Unfortunately it is impossible to estimate E_3 from this by procedures such as maximum likelihood because

the values of Y_1, Y_2, Y_3 and E_1 are unknown and affect X. However some information about these parameters and their effect on X does exist and this can be used to help get an estimate of E_3.

First, the model developed tells us that for any given value of E_3, the combination of E_1 and E_2 that results in producing the most homozygotes is $E_2 = 0$. In this case a much larger percentage of the observed homozygotes would have to be attributed to terminal fusion and less to gamete duplication than under any other combination of E_1 and E_2. Therefore $E_1 = (1 - E_3)$ with probability one corresponds to a least favorable distribution and amounts to estimating E_3 under the worst possible conditions. Such an estimate would tend to under estimate E_3, but a conservative estimate is more desirable than an over estimate in forming isogenic stocks. Using this prior, one obtains

$$f(X\colon Y_1, Y_2, Y_3, E_3) = \binom{N}{X} (G)^X (1 - G)^{N-X}$$

where $G = E_3 + (1 - Y_1)(1 - Y_2)(1 - Y_3)(1 - E_3)$.

For a prior on the Y's, we first note that since the markers are unlinked the Y's are independent. Y can range from 0 (absolute linkage to the kinetochore) to 2/3 (random recombination with the kinetochore). The markers used are all on major autosomes of *D. mercatorum*. Since it is known in many *Drosophila* species that the chromosome arms of the major autosomes are about 50 to 55 map units long, reasonable priors for the Y's would be independent and identically distributed uniforms on 0 to 2/3.

Finally E_3 itself can vary from 0 to 1. Maximum ignorance of the value of E_3 will be assumed, so the prior on E_3 is a uniform on 0 to 1. An estimate of E_3 can now be obtained by taking a normalized expected value of E_3 given the data; hence

$$E_3(X) = \frac{\int_0^1 \int_0^{2/3} \int_0^{2/3} \int_0^{2/3} E\, G^X (1-G)^{N-X}\, dY_1\, dY_2\, dY_3\, dE_3}{\int_0^1 \int_0^{2/3} \int_0^{2/3} \int_0^{2/3} G^X (1-G)^{N-X}\, dY_1\, dY_2\, dY_3\, dE_3}.$$

The form of the above estimate is actually an extension of a type of estimate known as the Pitman estimate. The Pitman estimate was first derived to estimate location parameters (e.g. the mean of a normal distribution) but can be extended to estimate other types of parameters such as E_3.

The evaluation of this estimator by normal numerical intergration techniques is very costly; therefore an alternative procedure of Monte Carlo integration given in Hammersley and Handscomb (1964) is used. Consider first the one dimensional integral

$$\theta = \int_0^1 f(X)\, d(X)\,.$$

If $\xi_1, \xi_2, \ldots, \xi_n$ are independent random numbers uniformally distributed between zero and one, then

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} f(\xi_i)$$

is an unbiased estimator of θ.

The situation here is four dimensional, but essentially no different. Sets of four independent random numbers are generated — three being uniform 0 to 2/3 and one uniform 0 to 1. The estimator of E_3 becomes

$$E_3(X) = \frac{\sum_{i=1}^{1000} B_{3i}[G(Y_{1i}, Y_{2i}, Y_{3i})]^X\, [1 - G(Y_{1i}, Y_{2i}, Y_{3i}, E_{3i})]^{N-X}}{\sum_{i=1}^{1000} [G(Y_{1i}, Y_{2i}, Y_{3i}, B_{3i})]^X\, [1 - G(Y_{1i}, Y_{2i}, Y_{3i}, E_{3i})]^{N-X}}$$

where $(Y_{1i}, Y_{2i}, Y_{3i}, E_{3i})$ is a set of random numbers. The estimators of the numerator and denominator are unbiased, but the ratio of these estimators is a biased estimator of the ratio. This bias will also underestimate E_3.

In an actual experiment heterozygous F_1 females from the cross *Br*$_7$-*S-v pm vl* males *X S-l-Im* females produced at total of 110 offspring, 107 of which were homozygous at all three marker loci. The estimate of E_3 is

$$E_3(X) = .94\,.$$

This estimate is in agreement with the conclusion given in Carson *et al.* (1969) that over 90% of the eggs develop by gamete duplication.

The fundamental parameters of the model are E_1, E_2, E_3 and Y, but if we are interested in the rate of decay of heterozygosity it is not necessary to estimate all of these but only

$$K = E_1 (1 - Y)/2 + E_2\, Y/4\,.$$

It is very important to measure K accurately if one desires to detect selection since, as will be shown in the next section, it is necessary to treat K as a constant in testing for selection.

Since K is the decay rate of heterozygosity in the absence of selection, it is best measured under conditions that minimize selection such as raising the flies under low densities in a uniform environment, sampling as early as possible and performing a one generation experiment. A one generation experiment is desirable because small selective differences are magnified over several generations. Also, in *D. mercatorum* it is possible to obtain isogenic parthenogenetic and bridge stocks so that the stocks used in an experiment can be very nearly genetically identical. Thus there are no hereditary differences in fecundity in a one generation experiment and the only place where selection can enter is the viability of the zygotes to the time of sampling. This would also make replicates in different bottles true replicates since there would be no genetic background effects. Therefore an optimal design would be to create a F_1 heterozygous female stock by crossing isogenic bridge males to isogenic impaternate females, raise the F_2 impaternate offspring under optimal conditions and sample as early as possible. Under these conditions, the model would predict that if N flies were sampled from

each bottle the number of heterozygotes in the i^{th} bottle, Y_i, would be

$$Y_i = K\,N + \xi_i \qquad i = 1, 2, \ldots, r$$

where ξ_i is a random error term and r the number of bottles.

The distribution of ξ_i will now be determined. Since the initial population is entirely heterozygous it will produce zygotes in the proportion of K heterozygotes to $1 - K$ homozygotes. From this zygoid pool in each bottle N flies are drawn at random and hence the distribution of Y_i is binomial with mean $K\,N$ and variance $N\,K\,(1 - K)$. Now $\xi_i = Y_i - K\,N$ and by the central limit theorem for large N

$$\frac{\xi_i}{\sqrt{N\,K\,(1-K)}} \sim \eta(0, 1)\,.$$

Therefore

$$\xi_i \sim \eta(0, \sigma^2)$$

where

$$\sigma^2 = N\,K\,(1 - K)\,.$$

The Gauss Markoff estimator of K is

$$\hat{K} = \frac{\sum_{i=1}^{r} Y_i}{r\,N}$$

and

$$Var\,\hat{K} = \sigma_k^2 = \frac{1}{r^2 N^2} \sum_{i=1}^{r} Var\,Y_i$$

$$\sigma_k^2 = \frac{K\,(1 - K)}{r\,N}\,.$$

Since $K < 1$, $K\,(1 - K) < \frac{1}{4}$ so

$$\sigma_k < \frac{1}{2\sqrt{r\,N}}\,.$$

From the normality a 95% confidence interval would be

$$K \pm 2\,\sigma_k$$

so with at least probability .95, the true K will be in the interval $(K - 1/\sqrt{r\,N},\ K + 1/\sqrt{r\,N})$. Since K will be treated as a constant in the next section, it is important to get an accurate estimate. If it is desired to get $K \pm x$, the total number of flies needed to be sampled is at the maximum $1/x^2$. For example, if $x = .02$, $r\,N = 2500$.

However there is strong reason to believe $E_3 = .90$ or larger. Since $K < (1 - E_3)$, one can be fairly confident that $K < .10$ and therefore $K\,(1 - K) < < .09$. Under these conditions $\sigma_k < .3/\sqrt{r\,N}$ and it is only necessary to sample 900 flies to obtain $2\,\sigma_k < .02$ and a sample of 2500 flies would yield $2\,\sigma_k < .012$.

The knowledge about the value of E_3 can also be used to modify the estimate of K itself. K can theoretically take on any value from 0 to 1, but the experimental evidence indicates that it will be between 0 and .1. We can therefore put a prior distribution on K that gives most of its mass to the interval (0, .1). A prior that accomplishes this is the beta distribution $\Pi(K) = \beta(3/320, 57/320)\ (K)$ which has a mean of .05 and a variance of .04. Using this prior $\Pi(K)$ one can obtain the Bayes estimate of K with squared error loss as

$$\hat{K} = \frac{\sum_{i=1}^{r} Y_i + 3/320}{r\,N + 60/320}\,.$$

It is obvious that for large $r\,N$ the two estimates will give nearly identical results. Furthermore, a 95% confidence interval $(\alpha < K < \beta)$ is given by

$$\int_{\alpha}^{\beta} T(K|Y_1, \ldots, Y_r)\,dK = .95 \qquad (1)$$

where

$$T(K|Y_1, \ldots, Y_r) = \frac{f(Y_1, \ldots, Y_r|K)\,\Pi(K)}{\int_0^1 f(Y_1, \ldots, Y_r|K)\Pi(K)\,dK}$$

$$= \frac{K^{\Sigma Y_i - 317/320}(1 - K)^{rN - \Sigma Y_i - 260/320}}{\beta\left(\sum_{i=1}^{r} Y_i + \frac{3}{320},\ r\,N + \frac{57}{320} - \sum_{i=1}^{n} Y_i\right)}\,.$$

This is a beta distribution on K and tables exist for it. The values of α and β are chosen to make (1) true. It is convenient to choose them in a symmetrical fashion so $\alpha = \hat{K} - \mu$ and $\beta = \hat{K} + \mu$.

Comparing Hypotheses

One of the major problems in population genetics is the role of selection in the maintenance of genetic variability. One reason there is so much controversy in this area is that it is often difficult to sort out the effects of inbreeding, effective population size, migration, non-random mating and other parameters from selection. As already stated, parthenogenetic populations of *Drosophila mercatorum* could prove useful in studying this problem. Consequently it is important to develop statistical procedures to selection in parthenogenetic strains.

To test for the presence of heterosis consider the model in which a parthenogenetic population of *D. mercatorum* of size N with discrete generations is followed at a marker locus for several generations with the following fitnesses

Genotype	AA	Aa	aa
Fitness	1	$1 + s$	1

$$X_{eq} = \frac{1}{\frac{(1 - K)}{(1 + s)\,K - 1} + 1}\,.$$

When $(1 + s)\,K > 1$, $X_{eq} > 0$, but when $(1 + s)\,K < 1$, there will be no heterozygotes at equilibrium. However, selection will retard the decay of heterozygosity if $s > 0$.

If such a population is followed for n gnerations a random sequence is generated $X = X_1, X_2, \ldots, X_n\}$ where X_i is the number of heterozygotes at generation i. The first step in deriving tests to detect selection

is to obtain the joint likelihood of such a sequence when $s = 0$ and when $s \neq 0$. Let the initial population be completely heterozygous so $X_0 = N$. Then in the absence of selection the next generation is chosen from the zygoid pool with probability K of being heterozygous; therefore

$$f(X_1) = \binom{N}{X_1} K^{X_1} (1 - K)^{N - X_1}$$

and in general

$$f(X_i | X_{i-1}, \ldots, X_1) = f(X_i | X_{i-1}) = \binom{N}{X_i} \left(\frac{K X_{i-1}}{N}\right)^{X_i} \left(1 - \frac{K X_{i-1}}{N}\right)^{N - X_i}.$$

Using

$$f(X_i, X_{i-1} / X_{i-2}, \ldots, X_1) = f(X_i | X_{i-1}, \ldots, X_1)\, f(X_{i-1} | X_{i-2}, \ldots, X_1)$$

the likelihood function when $s = 0$ is obtained as

$$L_0(X_1, \ldots, X_n) = \prod_{i=1}^{n} \binom{N}{X_i} \left(\frac{K X_{i-1}}{N}\right)^{X_i} \left(1 - \frac{K X_{i-1}}{N}\right)^{N - X_i}.$$

Similarly when $s \neq 0$

$$L_1(X_1, \ldots, X_n | s) = \prod_{i=1}^{n} \binom{N}{X_i} \left(\frac{K X_{i-1}(1 + s)}{N + s K X_{i-1}}\right)^{X_i} \times \left(1 - \frac{K X_{i-1}(1 + s)}{N + s K X_{i-1}}\right)^{N - X_i} = \prod_{i-1} \frac{(1 + s)^{X_i}}{(N + s K X_{i-1})} \times N \binom{N}{X_i} (K X_{i-1})^{X_i} (N - K X_{i-1})^{N - X_i}.$$

Notice that in both likelihood functions K is treated as a known constant. This is necessary because K and s are not estimable from the same data set since they can confound each other. For example, a population with a small K and $s > 0$ could behave similarly to a population with a larger K and no selection. Because K is treated as a constant it is very important to obtain an accurate, independent estimate of K using the procedures of the preceding section.

It is easy to obtain a maximum likelihood estimate of s from the likelihood function by differentiating it with respect to s and setting this equal to zero. One obtains

$$f(s) = \sum_{i=1}^{n} X_i - (1 + \hat{s}) K \sum_{i=1}^{n} \frac{X_{i-1}}{1 + \frac{s K X_{i-1}}{N}} = 0.$$

An initial approximation when s is small and hence $1 + \frac{s K X_{i-1}}{N} = 1$ is

$$\hat{s}_0 = \frac{\sum_{i-1}^{n} X_i - K \sum_{i-1}^{n} X_{i-1}}{K \sum_{i=1}^{n} X_{i-1}}.$$

A more exact solution of $\hat{s}$ can be obtained by the Newton-Raphson method using the same initial approximation and the iteration formula

$$\hat{s}_{r+1} = \hat{s}_r \frac{f(\hat{s}_r)}{f'(\hat{s}_r)}$$

$$f'(\hat{s}_r) = \frac{(1 + \hat{s}_r) K^2}{N} \sum_{i=1}^{n} \frac{X_{i-1}^2}{\left(1 + \frac{\hat{s}_r K X_{i-1}}{N}\right)^2} - K \sum_{i=1}^{n} \frac{X_{i-1}}{1 + \frac{\hat{s}_r K X_{i-1}}{N}}.$$

This iteration is continued until the s_r's converge to the desired level of accuracy.

Despite the desirability of obtaining an estimate of s, it is more important to decide whether or not selection is operating on the system; i.e. one must test the hypothesis of selection versus the hypothesis of no selection. A common method of deciding which hypothesis is correct is to form an index function of the data and calculate its probability distribution under the null hypothesis $s = 0$. One would reject the null hypothesis when the probability of the observed value or larger values of the index function is sufficiently small; say reject H_0 when $I(X) > i$ where $P(I(X) > i) = \alpha$. A serious limitation of this approach is that it ignores the likelihood of $I(X)$ under the alternative hypothesis. This can sometimes lead to paradoxical conclusions. For example, consider the hypothetical situation given in Fig. 2 where L_0 is the likelihood of the index function I under a null hypothesis and L_1 the likelihood under an alternative hypothesis.

One rejects H_0 when $I < i$ where $P(I < i) = \alpha$. If an experiment was performed and the observed value of I was I_1, the test would lead us to accept H_0 even though it is much more likely to obtain I_1 under H_1. On the other hand, if the observed value was I_2 the test would lead us to reject H_0 even though it is more likely to get I_2 under H_0 than H_1. Unfortunately situations such as this could easily arise when testing for the presence of selection since many experiments have shown that values of s tend to cluster around 0 and selection often reduces genetic drift yielding a likelihood function with most of its mass concentrated near $s = 0$.

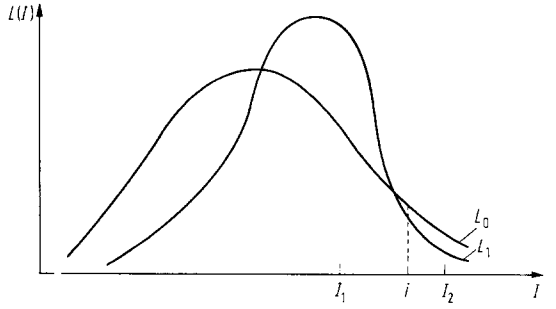

Fig. 2. Hypothetical likelihoods of an Index function $I(X)$ under two different hypotheses; H_0 and H_1. Explanation in text

In order to avoid such difficulties the likelihood ratio or odds for the two hypotheses can be calculated. Such an approach would give which hypothesis is more likely and by how much and hence gives the experimenter more information on which to base a decision. However, in our case we are comparing a simple hypothesis $s = 0$ against the complex hypothesis $s \neq 0$ which has an infinite number of possible values for s and hence an infinite number of likelihood functions. Therefore, to compare the likelihoods it is first necessary to eliminate somehow the dependence upon s of the likelihood function when $s \neq 0$. One method of accomplishing this is to find a sufficient statistic for s since the likelihood given the sufficient statistic would be independent of s. Unfortunately a sufficient statistic for s has not been found. Another way to eliminate s is to investigate the likelihood ratio in the vicinity of $s = 0$ and ignore the ratio under other values of s. The rationale for this is that most confusion will arise when s is close to zero and it is therefore desirable to device a test which has optimal properties in that vicinity. A third approach is to use prior knowledge about the distribution of s and weight the likelihoods with this distribution. Both of these latter two approaches are used.

The first test to be derived is the locally best test in the vicinity of $s = 0$. The basic principle of a locally best test is to find a test with maximum power at some point in the parameter space out of all tests of the same size at that point. Consider testing the hypothesis $H_0: s < 0$ against $H_1: s > 0$. If a test is written as

$$\varphi(X) = \begin{cases} 1 \text{ if the test function is in the critical region} \\ 0 \text{ otherwise} \end{cases}$$

meaning reject the null hypothesis with probability one if the test function is in the critical region and accept it with probability one if it is not, then the power function can be written as

$$\beta_\varphi(s) = E_s\, \varphi(X) = \int \varphi(X)\, f(X: s)\, dX$$

where $f(X:s)$ is the probability density function of the data X assuming s is the true value of the parameter. We also assume the distribution of s is such that one continuous derivative may be passed beneath the integral sign

$$\beta_\varphi(s) = \int \varphi(X)\, \frac{\partial}{\partial s} f(X: s)\, dX\,.$$

A test φ_0 for $H_0: s < 0$ against $H_1: s > 0$ is said to be the locally best test if for any other test φ for which $\beta_\varphi(0) = \beta_{\varphi_0}(0)$ we have $\beta'_\varphi(0) < \beta'_{\varphi_0}(0)$; i.e. the locally best test has maximum slope of the power function at $s = 0$ out of all tests of the same size at $s = 0$.

It has peen shown (see Ferguson, 1967) that the form of the locally best test is

$$\varphi_0(X) = \begin{cases} 1 \text{ if } \dfrac{\frac{\partial}{\partial s} L(X:s)}{L(X:s=0)}\Big|_{i=1} > C(\alpha, n) \\ 0 \text{ otherwise} \end{cases}$$

where C is a function of n and α, the size of the sample and the size of the test respectively. Now

$$\frac{L_1(X|s)}{L_0(X)} = \prod_{i=1}^{n} \frac{(1+s)^{X_i}}{\left(1 + \frac{s\,K\,X_{i-1}}{N}\right)^N} = H(s)$$

$$H(s) = (1+s)^{\sum_{i=1}^{n} X_i}\, e^{-N \sum_{i=1}^{n} \ln\left(1 + \frac{sKX_{i-1}}{N}\right)}.$$

So

$$H'(s)|_{s=0} = \sum_{i=1}^{n} X_i - K \sum_{i=1}^{n} X_{i-1}\,.$$

Hence the locally best test is

$$\varphi_0(X) = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} X_i - K \sum_{i=1}^{n} X_{i-1} > C(\alpha, n) \\ 0 \text{ otherwise} \end{cases}$$

In order to evaluate C for a given n and α, it is necessary to find the probability distribution of $\sum_{i=1}^{n} (X_i - K\,X_{i-1})$ or some function of it. As is shown in the appendix

$$\frac{\sum_{i=1}^{n} (X_i - K\,X_{i-1})}{N \sum_{i=1}^{n} K^i(1 - K^i)} = U$$

is approximately distributed as a normal with mean zero and variance one. Therefore the locally best test is to reject H_0 when

$$U > \Phi(\alpha)$$

where $\Phi(\alpha)$ is a function of α only and is obtained from a standard normal distribution table.

Another way of distinguishing between $s = 0$ and $s \neq 0$ is to calculate "odds" of a given data set and use a prior distribution on s to weight the likelihoods. Before deriving the odds some similarities between this approach and the more usual likelihood ratio test must be pointed out. For simplicity assume for the present that we are interested in distinguishing between two simple alternatives, H_0 and H_1. The usual likelihood ratio test is of the form

$$\text{reject } H_0 \text{ when } \frac{L_1(X)}{L_0(X)} > k$$

where $L_0(X)$ is the likelihood under H_0 and $L_1(X)$ the likelihood under H_1. Now let p be the prior probability that H_0 is true and $1 - p$ the prior probability that H_1 is true. Then the odds using these priors are

$$\frac{(1-p)\, L_1(X)}{p\, L_0(X)}$$

One would tend to accept that hypothesis which is more likely, i.e. reject H_0 when $(1-p)\, L_1(X)/p\, L_0(X) > 1$. Notice that both "tests" are identical when $p = k/(1+k)$. This illustrates an extremely important point — that even the usual likelihood ratio test has an implicit prior distribution on the hypotheses. Thus the fact that a prior distribution

has to be chosen to calculate the odds is not a weakness of this approach, but rather a strength since outside information can be used to chose a prior that is more realistic than the arbitrary prior implicit in the likelihood ratio test.

In the case of distinguishing between $s = 0$ and $s \neq 0$ it is first noted that s can vary from -1 (heterozygote lethal) to ∞ (homozygote lethal). Many studies have been done on measuring selection coefficients (Dobzhansky and Spassky, 1954; James, 1959; Jain and Allard, 1960; Muller, 1950; and Timofeeff-Ressovsky, 1940), and the outcome of these studies suggests that most selection coefficients for a given locus are close to zero while lethals and semi-lethals are relatively rare. Therefore a realistic prior on s would be one which gives most of its mass in the vicinity of $s = 0$ and gives much less mass to the semi-lethal and lethal regions. A prior distribution that accomplishes this is a gamma distribution on $1 + s$. Therefore the prior on $1 + s = t$ is taken to be

$$\Pi(t) = \frac{t^{r-1} e^{-t}}{\Gamma(r)} .$$

The mode is at $s = 0$ when $r = 2$.

In general the odds have the form

$$0(X) = \frac{L_0(X)}{\int_{-1}^{\infty} L_1(X|S)\, \Pi(S)\, dS} = \frac{1}{\int_{-1}^{\infty} H(S)\, \Pi(S)\, dS} .$$

If N is large

$$0(X) \doteq \frac{1}{\int_{-1}^{\infty} (1+s) \sum_{i=1}^{n} X_i\, e^{-sK \sum_{i=1}^{n} (X_{i-1}) \Pi(s) ds}} .$$

Using the chosen prior with $r = 2$

$$0(X) = \frac{e^{-\beta}\, (\beta+1)^{\alpha+2}\, \Gamma(2)}{\Gamma(\alpha+2)} = \frac{e^{-\beta}\, (\beta+1)^{\alpha+2}}{(\alpha+1)!}$$

where $\alpha = \sum_{i=1}^{n} X_i$ and $\beta = K \sum_{i=1}^{n} X_{i-1}$. When H_0 is true $\alpha \doteq \beta$. In this case the odds are

$$0(X) \doteq \frac{e^{-\beta}\, (\beta+1)^{\beta+2}}{(\beta+1)!} .$$

If we further assume that the experiment has used a population size and continued long enough to insure that β is large, Stirling's approximation can be used:

$$(\beta+1)! = \sqrt{2\Pi}\, (\beta+1)^{\beta+3/2}\, e^{-\beta-1}$$

$$0(X) \doteq \frac{e}{\sqrt{2\Pi}} (\beta+1)^{1/2} .$$

Since it is assumed β is large and $e/\sqrt{2\Pi} > 1$, the odds will be very much greater than one. Thus when $\alpha \doteq \beta$ the odds are greatly in favor of H_0 as expected. Notice that the odds, the locally best test and the estimator of s all depend on the data in a similar fashion since the locally best test is to reject H_0 when $\frac{\alpha - \beta}{\sqrt{N \sum K^i (1 - K^i)}} > \Phi(\alpha)$ and $\hat{s}_0 = \frac{\alpha - \beta}{\beta}$. Basically both tests measure the "goodness of fit" of the data generation by generation with its predicted value under the null hypothesis.

These tests and the estimation procedures of the previous section provide a statistical framework to deal with one locus data and suggest optimal experimental designs for detecting selection in parthenogenetic populations. The appropriate stocks of *Drosophila mercatorum* are presently being formed to carry out such experiments. However, one limitation of the work presented here is that it deals with an individual locus as the unit of selection, but this may be an erroneous concept (see Franklin and Lewontin, 1970). It is therefore desirable to extend this analysis to two loci. Asher (personal communication) has recently developed a two locus deterministic model for parthenogenetic populations. The statistical implications of this model with respect to estimation, detection of heterotic selection, and measurement of correlations between non-alleles are currently being investigated. The development of these will yield a very thorough statistical framework which could deal with many questions and further emphasizes the usefullness of parthenogenetic populations in approaching basic evolutionary problems.

Acknowledgements

The authors wish to thank Dr. Charles F. Sing and Dr. James H. Asher, Jr. for their critical evaluations and guidance during the preparation of this manuscript, and to Dr. Hampton L. Carson for providing the stocks of *Drosophila mercatorum* used in these investigations.

Appendix; Derivation of the Distribution of U

The joint density of $X_1, X_2, \ldots, X_n$ under H_0 is given by

$$f(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \binom{N}{X_i} (K X_{i-1}/N)^{X_i} (1 - K X_{i-1}/N)^{N - X_i} .$$

For large N the binomials can be approximated by normals

$$f(X_1, \ldots, X_n) = \left(2\Pi K X_{i-1} (1 - K X_{i-1}/N)\right)^{-1/2} \times e^{-\frac{(X_i - K X_{i-1})^2}{2K X_{i-1}(1 - K X_{i-1}/N)}}$$

Let

$$Y_i = \frac{X_i - K X_{i-1}}{\sqrt{N}} .$$

This is a one to one transformation, and the inverse of the transformation is

$$X_i = \sqrt{N}\, Y_i + K \sqrt{N}\, Y_{i-1} + \cdots + K^{i-1} \sqrt{N}\, Y_1 + + K^i N .$$

The Jacobian of the transformation is N^n. Thus

$$f(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} \left[2\pi \left(\frac{K\, Y_{i-1}}{\sqrt{N}} + \frac{K^2\, Y_{i-2}}{\sqrt{N}} + \cdots + K^i \right) \left(1 - \frac{K\, Y_{i-1}}{\sqrt{N}} - \cdots - K^i \right) \right]^{-1/2} \times \exp \left[\frac{-Y_i^2}{2\left(\frac{K\, Y_{i-1}}{\sqrt{N}} + \cdots + K^i \right)} (1 - \ldots - K^i) \right]$$

and as N gets large

$$f(Y_1, \ldots, Y_n) \prod_{i=1}^{n} [2\pi K^i (1 - K^i)]^{-1/2} \times \exp \left[\frac{-Y_i^2}{2 K^i (1 - K^i)} \right].$$

From the central limit theorem

$$f(Y_i) \simeq \left(2\pi K^i (1 - K^i)\right)^{-1/2} \exp\left(-Y_i^2 / 2 K^i (1 - K^i)\right).$$

Therefore

$$f(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f(Y_i).$$

Since the joint density is asymptotically the product of the marginals, the Y's are asymptotically independent $\eta\left(0, K^i (1 - K^i)\right)$.

Let $T = \sum_{i=1}^{n} Y_i$. We can use the characteristic functions of the Y's and their asymptotic independence to calculate the density of T.

$$\psi_T(m) = \prod_{i=1}^{n} \psi_{Y_i}(m)$$

where

$$\psi_{Y_i}(m) = e^{-1/2 K^i (1 - K^i) m^2}$$

so

$$\psi_T(m) = e^{-1/2 m^2 \sum_{i=1}^{n} K^i (1 - K^i)}$$

which implies that

$$f(T) = \eta\left(0, \sum_{i=1}^{n} K^i (1 - K^i)\right)$$

Hence, if $U = T \Big/ \sqrt{\sum_{i=1}^{n} K^i (1 - K^i)}$

$$f(U) = \eta(0, 1).$$

Literature

1. Asher, J. H., Jr.: Parthenogenesis and genetic variability. II. One-locus models for various diploid populations. Genetics **66**, 369–391 (1970a). — 2. Asher, J. H., Jr.: Parthenogenesis and genetic variability. Doctoral thesis, University of Michigan, Ann Arbor, Michigan (1970b). — 3. Carson, H. L.: Permanent heterozygosity. Evol. Biol. **1**, 143–168 (1967a). — 4. Carson, H. L.: Selection for parthenogenesis in *Drosophila mercatorum*. Genetics **55**, 157–171 (1967b). — 5. Carson, H. L., Wei, I. Y., Niederkorn, J. A., Jr.: Isogenicity in parthenogenetic strains of *Drosophila mercatorum*. Genetics **63**, 619–628 (1969). — 6. Dobzhansky, Th., Spassky, B.: Genetics of natural populations. XXII. A comparison of the concealed variability in *Drosophila prosaltans* with that in other species. Genetics **39**, 472–487 (1954). — 7. Ferguson, T. S.: Mathematical Statistics: a Decision Theoretic Approach. New York: Academic Press 1967. — 8. Franklin, I., Lewontin, R. C.: Is the gene the unit of selection? Genetics **65**, 707–734 (1970). — 9. Hammersley, J. M., Handscomb, D. C.: Monte Carlo Methods. London: Methuen and Co. 1964. — 10. Jain, S. K., Allard, R. W.: Population studies in predominately self-pollinated species. I. Evidence for heterozygote advantage in a closed population of barley. PNAS **46**, 1371–1377 (1960). — 11. James, A. P.: The spectrum of severity of mutant effects. I. Haploid effects in yeast. Genetics **44**, 1309–1324 (1959). — 12. Muller, H. J.: Radiation damage to the genetic material. I. Effects manifested mainly in descendants. Am. Scientist **38**, 33–59 (1950). — 13. Timofeeff-Ressovsky, N. W.: Mutations and geographical variation, pp. 73–136. In: The New Systematics, edited by J. Huxley. Oxford: Clarendon Press 1940.

Received April 3, 1972

Communicated by R. W. Allard

Alan R. Templeton
Department of Human Genetics
The University of Michigan
Ann Arbor, Michigan 48104 (USA)

Dr. Edward D. Rothman
Department of Statistics
The University of Michigan
Ann Arbor, Michigan 48104 (USA)