MATHEMATICAL ANALYSIS OF GENETIC PROBLEMS IN PLANT AND ANIMAL BREEDING

by

R. E. COMSTOCK

Institute of Statistics
Mimeograph Series No.52

MATHEMATICAL ANALYSIS OF GENETIC PROBLEMS IN PLANT AND ANIMAL BREEDING

CHAPTER I

Perspective

The basic objective of the breeder is the production of genetic populations that in specific ways are superior to ones already in being. The materials at his command for creation of such populations are the genes that can be introduced into one inter-breeding group of organisms via cross-fertilization. Within the limitation of this material the task is conceptually, though not actually, simple; it is to produce individuals whose genotypes represent optimum combinations of the available genes and to devise breeding systems with which these superior genotypes can be "re-produced" at will.

Breeding procedures can be divided into two categories: (1) those concerned with inter-population transfer of genes, and (2) those concerned with the synthesis of optimum genotypes from the genes co-existent in a particular inter-breeding population.

Because genes must first be present in the same genetic population before they can be incorporated into a single genotype, barriers to inter-population gene transfer are limiting factors to the breeders accomplishments. Gene transfer depends on the production of fertile progeny from inter-population crosses. The nature of barriers to cross-fertility varies greatly, (Dobzhansky, 1951; Stebbins, 1950). Hybrids may for a variety of reasons be impossible to obtain, while in other instances they can be produced but are infertile. One of the best understood causes of hybrid infertility is meiotic failure due to lack of chromosome homology. Some of these barriers, when understood, can be overcome by appropriate devices or procedures.

Work in this area by cyto-genetecists and students of reproductive physiology has contributed greatly in plant breeding by making it possible to aggregate genes from more diverse sources than was at one time possible.

The process of synthesizing useful genotypes is hampered by the fact that the genes, which are the building units to be combining in constructing the edifice that is the genotype, can neither be examined nor manipulated individually. The properties of genes that are of concern to the breeder are their effects on the phenotype of the organism, but the normal organism carries an entire gene complement and no aspect of its phenotype can be described as the property of a single gene. Rather each aspect of the phenotype represents the net effect of the entire gene complement.* So far as manipulation is concerned the genetic mechanism is such that the operational unit must be a set of genes (ordinarily those contained in a gamete, but in some special situations those in a chromosome or piece of chromosome). Moreover the only control that can be brought to bear on the gene composition of these sets is by the indirect route of selection, predicated on phenotype, of the individuals from which the sets are to be derived. As a consequence, some deleterious genes invariably move with the useful genes and the "optimum genotype" can be no more than a goal to be struggled toward.

Whatever the effectiveness of selection may be, it is the only directional force that the breeder has for guiding genotype composition. It is the force he must rely on; there is no other. In only three respects are alternatives

---

*Evidence continues to accumulate to the effect that even character which appears to vary in response to segregation at a single locus are in reality conditioned not alone by the alleles of that locus by those genes acting against the background provided by the entire genotype.

open to him. Within limits there is usually a choice of foundation stocks, which is to say that judgement may be exercised in establishing a basic gene pool. The second choice to be made is the nature of units among which to select. The unit may be the individual or one of many types of families (homozygous lines, F_1's of homozygous lines, half-sib or full-sib families, vegetative clones, etc.). If it is to be a family, then its size is (again within limits) at the breeders discretion. It should be noted that choice of the unit of selection amounts to choice of a mating system since production of a specific type of individual or family implies a specific system of mating. The third choice is of characters to be emphasized in selection and the relative emphasis to be placed on each. In this he will be guided by the relative economic values that will accrue from improvement in various characters and the relative ease with which he anticipates they can be altered.

This book will be devoted to the problems that the breeder faces in making these decisions.

The Analytical Approach

What we shall call the analytical approach is primarily a deductive process. It starts with a model and supplies the expectations associated with a specified set of operations. A simple illustration will serve for clarification. Assume a mono-factorial character in a diploid organism that exhibits regular meiosis, and further that the character can be expressed in but two alternative ways of which one exhibits dominance. This is the model; it consists, as it always does, of a set of assumptions about the genetics of the organism and trait. Now suppose that the following operations

are to be performed: (1) two individuals, one homozygous for the dominant and the other for the recessive expression of the trait, are to be mated, (2) the progeny are to be mated at random among themselves, and (3) the progeny of these matings that possess the character in its dominant form are to be mated at random among themselves. The situation is such an elementary one to students of genetics that the expectations are perceived with almost no mental effort. They are, of course, as follows:

1. Offspring from the first mating that are all heterozygous.
2. In the next generation, 3:1 segregation between the dominant and recessive expressions of the trait.
3. In the final generation, an 8:1 ratio between numbers of offspring possessing the trait in its dominant and recessive forms. Fifty percent heterozgates among individuals exhibiting the dominant form.

Knowledge of such expectations serves both the breeder and the genetecist. The latter finds in them the basis for tests of the parent model, i.e. of genetic hypotheses. The expected 3:1 ratio in the second generation of the above illustration is, for example, a time honored test of the monofactorial, complete dominance, hypothesis. Of course, it is not a completely critical test. There are other models that yield the same expectations in the case of the operations leading here to the 3:1 ratio. This is typical. It is the rule rather than the exception that a series of tests is required before all alternatives to a particular hypothesis or model can be excluded as possible explanations of the phenomenan being studied.

The breeders most immediate interest is in the expected consequences of systems of breeding. Typically these are obtained as functions of one

or more parameters (values that are constant for a given population, but may vary from one population to another) that characterize genetic variables which bear on the success to be anticipated by the system in question. A companion objective must then be to discover sets of operations with expectations such that data obtained can be used for estimation of these genetic parameters.

Unfortunately, genetic models that are general enough to be completely realistic usually pose very difficult mathematical problems. Consequently, most results dealing with all but the simplest problems, that have been produced, are based on models that have been simplified in one or more important ways. This is sometimes pointed to as the fatal weakness of the mathematical approach in genetics, the implication being that since, for the most part, only non-realistic problems can be solved, the approach may as well be abandoned. However, those who take this point of view dodge the question their argument raises concerning a more useful alternative approach. The answer appears to be that there is no alternative so far as the basic genetic questions are concerned. A problem for which theoretical analysis in symbolic language is impossible can never be solved experimentally. As Benjamin (1949) has so clearly outlined, the critical function of experimentation in an "explanatory science" lies in verification of hypotheses by checking whether the observed results of operations agree with expectations obtained deductively. Obviously if the expectations cannot be deduced, this verification is impossible.

There is an alternative so far as breeding problems are concerned. Breeding systems can be compared empirically and, if this is done in a thorough fashion, with completely satisfactory results so far as the practical aspect is concerned. In fact, progress in breeding methods in the past

has stemmed in part from the empirical approach. There are, however, two counts against turning completely to empiricism. First, its cost is, if not prohibitive, at least so great that information would accumulate very slowly. Consider, for example, the time and expense that would be involved in learning whether the hybrid system as used with corn can be effective with beef cattle. Then add the cost of evaluating the other systems that would suggest themselves as possible superior alternatives. In terms of money and facilities that have been available in the past, the total cost involved would be staggering and results of any great consequence would hardly be obtained during the lifetime of the present cattleman. Second, a large portion of the value of empirical studies in the past has stemmed from the light thrown on basic genetic problems. But this has been so because theory has made it possible to relate empirical observations to underlying causative mechanisms. Abandonment of the analytical approach would mean that results of much empirical work could no longer contribute to our basic understanding of genetics.

Admitting that a complete theory may forever be beyond our grasp, there is nevertheless a brighter side. Expectations premised on simplified models constitute approximations that can be used to advantage if their deficiencies are understood. For example, expectations associated with some operations are not seriously disturbed by model simplification and where this is true, data obtained using these operations can be very useful. In addition there is more than one way in which models can be simplified enough to work with, and by putting together results based on different models it is sometimes possible to synthesize an adequate basis for constructive action. Actually,

it appears certain, rather than otherwise, that continued efforts to expand mathematical theory in genetics will be fruitful and will contribute to the advance of the entire science.

Quantitative Characters

Quantitative characters will be defined to include (1) all those exhibiting continous variation, and (2) those for which variation is discontinuous but partly non-genetic in origin. A few examples will serve to clarify the distinction intended between quantitative and qualitative characters.

Coat color in Aberdeen Angus cattle varies discontinuously; it is either black or red. In addition, non-genetic variation is not exhibited. It happens that the trait is controlled by a single pair of alleles, one of which is recessive to the other. So far as is known, homozygous recessives are always red and other individuals are always black. This character is clearly of the qualitative category. Many more such examples could be cited. Eye, flower, and coat color contrasts in many species; the horned-polled contrast in cattle; the starchy-sugary contrast in maize, and the blood groups of man and cattle are of this sort.

At the other extreme are such traits as the growth rates and mature weight of animals and the grain yield of various crop plants. These exhibit continuous variation and are universally referred to as quantitative characters.

Another sort of character that differs superficially from those just discussed is exemplified by litter size in swine (or any of the multiparous animals). In single expressions it is discontinuous in its variation, being limited to whole numbers that only infrequently are as large as

fifteen. However, phenotypic variation is not entirely under genetic control, but results more from environmental than from genetic effects. This becomes obvious when it is observed that numbers vary among litters produced by the same female. In fact, the averages of litter size by individuals approach a continuous distribution as the number of litters per individual increases. This is in contrast to the usual behaviour of flower color where all flowers produced by a particular plant are similar in color. Examples of other characters in the litter size category are digit number of guinea pigs (Wright, 1934), kernel row number in maize (Emerson and East, 1913), and locule number in tomatoes or cotton. These characters differ from the obviously quantitative ones only because phenotypic expression is limited by its nature to specific points of its range. It seems entirely appropriate to include these with the other quantitative characters.

There has been a tendency in the genetic literature to suggest that quantitative characters as a class are distinguished by being conditioned by numerous genes which individually have small effects. In this connection the term, multi-factorial character, is rather generally accepted as synonymous with quantitative character. More recently Mather (1941) has coined the term, polygenic character, as a substitute for multi-factorial character. While there can be little doubt that most quantitative characters are actually affected by numerous genes, it is worth noting that this is not a necessary attribute of such traits. Any mono- or di-factorial character sufficiently affected by ordinary variation in environment to cause genetic differences to be obscured in the phenotype is perfore a quantitative character. For this reason if for no other the term quantitative character is to be preferred to multi-factorial or polygenic character as a name for that class

of characters for which the classical procedures of factorial analysis are ineffective.

The essential point of difference between quantitative and qualitative characters turns out to be the magnitude of single gene effects relative to environmental effects. If the former are large enough so that they are but rarely obscured by the latter, the character involved is qualitative, otherwise it is quantitative. In this connection it is worth noting that when genes with unusually large effects are segregating, characters usually considered quantitative become qualitative in so far as the variation produced by those genes is concerned. This is true of plant height in corn when one of more of the known dwarfing genes are segregating. It is true also of viability in many plants and animals as affected by many lethal genes. The corollary is that for many characters in which major variation is qualitative, there is intra-class variation of the sort identified with quantitative characters. For example, among Angus cattle that are black there is undoubtedly variation in color intensity that is partly genetic in nature, but phenotypically continuous.

Quantitative Characters and Applied Genetics

The breeders concern is for the most part with the great variety of traits that condition the economic value of the organism. Examples are growth rate in meat animals; milk production in dairy cattle; quality and yield of wool in sheep; reproductive rate; resistance to various pathogenes and parasites in both plants and animals; climatic adaptation; and acre yield of cereals, cotton lint, sugar, or coffee. These are all quantitative characters, and many more of that same class could be added to the list. Qualitative characters having economic significance can also be

listed. Examples are coat color in farm animals (of considerable importance in some situations), presence or absence of horns in cattle, sugary and waxy endosperm in corn, and presence or absence of awns in wheat and barley. It is obvious, however, that most breeders must deal continuously with quantitative characters. Moreover, the latter are the more difficult to manipulate. This is due to greater complexity in mode of inheritance and the masking of genetic by environmental effects.

It is worth noting in passing that quantitative characters are also of prime significance in other fields of genetic application. The course of organic evolution is determined by the relative "survival values" of species and the individuals of which species are comprised. Survival value in turn represents a complex of traits of which many of the more important ones are quantitative. And turning to pathology we find that such traits as predisposition to cancer and tuberculosis, and longevity are also quantitative.

Scope of the Book

This is intended as an elementary treatment of the subject. The objectives are:

1. To demonstrate the analytical process at a mathematical level that will allow the ordinary student of genetics and breeding to understand each step involved.
2. To present, with derivation, an organized body of theory relating to breeding problems, and to discuss it with reference to breeding procedure and the investigation of quantitative inheritance.

Most of the material to be presented is available, though scattered, in the literature in scientific journal articles. In addition much of the

theory itself is summarized in texts by Lerner (1950), Mather (1949), and Lush (1943). From my contacts with students I have gained the impression that many have read the theory or had portions of it presented to them but that they fail to fully comprehend its implications because they do not understand its bases and derivation. My own efforts to obtain a grasp of the subject from the literature have often been hampered by the variation in symbolism, definition, and mathematical devices employed by different authors, and by the ommission of intermediate steps in derivation so often necessitated in journal articles. For example, some of the most important early papers involving statistical ideas employ statistics and terminology that have been superceded in most modern courses in statistics. Thus, the emphasis to be placed on derivation is deliberate in the belief that it will facilitate the students grasp of what seems to me a vital aspect of a satisfactory training in genetics.

Attention will be devoted primarily to problems encountered with quantitative characters. Reasons are (a) in general, the breeder encounters his most important, and at the same time, his most perplexing problems in his work with quantitative characters, (b) texts devoted to mathematical analysis of problems with qualitative characters are already available (Hogben, 1946; Li, 1948)

Literature cited:

1. Benjamin, A. Cornelius (1949), On Defining "Science". The Scientific Monthly, LXVIII:192-198.

2. Dobzhansky, Theodosius (1951), Genetics and the Origin of Species, 364 pp. 3rd ed. Columbia University Press, New York.

3. Emerson, R. A. and E. M. East (1913), The Inheritance of Quantitative Characters in Maize. Nebr. Agric. Exp. Sta. Res. Bull. 2.

4. Hogben, Lancelot (1946), An Introduction to Mathematical Genetics, 260 pp. W. W. Norton and Co. Inc., New York.

5. Lerner, I. Michael (1950), Population Genetics and Animal Improvement. 342 pp. Cambridge University Press. Cambridge, England.

6. Li, Ching Chun (1948), An Introduction to Population Genetics, 321 pp. National Peking University Press, Peiping, China.

7. Lush, Jay (1943), Animal Breeding Plans, 2nd ed. 437 pp. Iowa State College Press, Ames, Iowa.

8. Mather, K. (1941), Variation and Selection of Polygenic Characters. J. Genet. 41:159-193.

9. Mather, K. (1949), Biometrical Genetics. 158 pp. Dover Publications, Inc., New York.

10. Stebbins, G. Ledyard, Jr. (1950), Variation and Evolution in Plants, 643 pp. Columbia University Press, New York.

11. Wright, Sewall (1934), The Results of Crosses Between Inbred Strains of Guinea Pigs, Differing in Number of Digits. Genetics 19:537-551.

CHAPTER III

The General Form of the Breeding Problem

The Nature of Selection in Breeding Programs

Breeding programs vary greatly in detail but, with few exceptions, the essential ingredients of all are

(1) development of genetically variable material, and

(2) selection among the units of that material.

The operational unit in selection will be called the selection unit; it may be the individual organism or any of various kinds of families, e.g. full-sib families, clones, pure lines, etc. The variable material within which selection is practiced will be referred to as the genetic sample and the units actually selected as the selections.

When the selections are interbred to obtain a new genetic sample from which to re-select, the breeding program will be called cyclic. All breeding programs are potentially cyclic but, in some, single cycles engage the breeder for much longer periods of time than in others. For example, in animal breeding by the common outcrossing system each generation represents a cycle and in swine and poultry a new generation can be produced each year. In sharp contrast is the hybrid system in the form so widely used with corn. Here a cycle includes both the development of a series of inbred lines and the laborious task of identifying the best of the hybrids that can be made from them. This requires no less than eight or ten years and may engage the breeder much longer depending on how thoroughly he feels he should exploit his original foundation materials.

The selection practiced in a single cycle of a breeding program may or may not be all in one stage. For example, in corn breeding, potential inbred lines are discarded in each successive generation of self-fertilization. Some are lost because they are barren and many others are culled because they exhibit such undesirable agronomic characters as lodging, undue susceptibility to specific diseases or insects, or poor pollen production. As homozygosity is approached it is common practice to eliminate a portion of the remaining lines on the basis of general combining ability (Jenkins and Brunson, 1932). Next, because the number of double-crosses possible using a small number of lines is very great,[1] only a fraction of

[1] Hayes and Immer (1942) point out that with n inbred lines $\frac{n(n-1)(n-2)(n-3)}{8}$ different double crosses are possible.

the double-crosses, that could be produced with the lines remaining, are actually made up for field testing. Those to be tested are frequently chosen on the basis of yield prediction as outlined by Jenkins (1934). Finally, a limited number of double-crosses are made up and compared in field trials.

The same step-wise process which we will call multi-stage selection is used in work with other organisms. For example, in swine breeding it is common to make a preliminary culling of males at weaning time (when the less promising ones are castrated) and final selection at from five to eight months of age.

The criterion for selection may be a single character or may involve several characters as suggested above in connection with the elimination of partially inbred lines of corn. In strawberries, for instance, the breeder must pay attention to color, shape, size, and firmness of berry; flavor; disease resistance; etc. in addition to yield. The breeder is always hoping to find individuals or families that excel in all of the traits considered important, but this hope is rarely realized. Instead, he is continually confronted with having to weigh weakness in one or more traits against strength in others. In practice, this is usually done subjectively. The approximation of optimum weights to be given such a series of characters will be considered in a later chapter.

The measurements that provide the basis for selection may be made on the selection units under consideration or on genetically related materials. In field comparison of double-cross corn hybrids data is taken on the hybrids themselves whereas data on related single-crosses provides the basis for selecting double-crosses to be tested in the field. In beef cattle, sires are almost always selected for their own attributes, while the production of dam, sisters, and daughters is called on in evaluating dairy sires.

The Measure of Genotypes

Accomplishment in breeding programs consists of creating and identifying better genotypes than existed before. This implies that there is a "value" of genotypes that can be measured on a quantitative scale. Before going on to discussion of the accomplishment or progress that may be anticipated in breeding programs, we must obviously have a clear notion of how the merit of genotypes is properly measured.

Dobzhansky (1951) defines the genotype as "the genetic constitution, the sum total of the hereditory factors received by the organism from its ancestors". In brief a genotype is no more than a particular complement of genes, and has no intrinsic value. It is rather the phenotype which develops under stimuli from the genotype that reflects "value" back on the latter. We learn in elementary genetics that the phenotype is, in the words of Sinnott, Dunn, and Dobzhansky (1950), "always a result of the interaction of a genotype with an environment", that the same genotype may, in different environments, yield different phenotypes. While qualitative characters are relatively stable under environmental variation, quantitative characters are responsive to numerous aspects of environments that are beyond control by the agronomist or animal husbandman. For this reason we cannot resort to definition of a "standard environment" to be used as a background in measurement of the "values" of genotypes. As an alternative we must be satisfied with defining the "value" of a genotype with reference to not one, but a population of environments.

At this point it is well to consider what is meant by an "environment". The environment of a plant can only be defined as encompassing all factors, other than the genotype of the plant, that affects its development. This includes all the physical and chemical attributes of the soil in which the plant grows, all aspects of the atmosphere which surrounds the plant, the amount and distribution of solar radiation which strikes the plants, and the number and kind of biological organisms (facteria, fungi, insects, etc.) to which the plant is exposed. This list is imposing as it stands but the nature of environment becomes still more complex when we consider, as we must, the infinite variety of patterns in time that each of these factors can present through the life of the plant. Animals may be less affected by soil variables than are plants but this is more than compensated by the fact that animals possess highly organized nervous systems with the result that events and the action behaviors of other organisms become a part of their environments. Sinnott, Dunn, and Dobzhansky (1950) point out that with but 100 loci segregating for only two alleles per locus the potential number of different genotypes is 3^{100} and that this number is greater than that of all the atoms of the earth. Considering the great variety of ways in which environment can vary it is obvious that the potential number of different environments possible for, say, corn plants in a single county of Iowa must be some such awe-inspiring figure as

that which Sinnott and his colleagues have used to impress us with the unlimited scope of genetic variation. The probability of finding two individuals in an interbreeding population that are identical in genotype, infinitessimal though it be, is probably no less than that of finding two for which environment has been the same.

Definition of the value of a genotype in terms of its average phenotypic expression in the environments of a specified population establishes "value" as a parameter since a particular genotype can never be observed in more than a sample of those environments.

Awareness that genotypes have no absolute value but are good or poor with respect to the milieu in which the individuals possessing them are placed is inherent in various practices and guiding principles of the breeder. A variety is said to be "adapted" in one region but not adapted in another, which is a succinct way of saying that the average of "values" of the genotypes represented in the variety is high with reference to the population of environments encountered in the first region but low with respect to the population of environments of the second region. Recognition that this sort of thing may be so prevents the breeder from recommending a strain, variety, or breed for use in a region or circumstances for which it has not been tested. And before a variety is considered to be adequately tested its performance must be observed under several different circumstances (as in different years and locations), a tacit acknowledgement that it is not one but a population of environments that must always be dealt with.

The rigor of definition necessary to mathematical analysis of the problem of estimating the "values" of genotypes can be achieved as follows. Let the individual environments in a population of environments be numbered 1, 2, 3, ... m, and the individual genotypes of a population of genotypes be numbered 1, 2, 3, ... n. There is no sacrifice of generality if it is assumed that any aspect of phenotype can be quantitatively described on some numerical scale of measurement, since even the expressions of qualitative characters can be represented numerically and at the same time meaningfully if the scale is judiciously chosen, Let P_{ij} symbolize the number (whole or

fractional) that describes the phenotype obtained with the i-th genotype in the j-th environment. (For example, if the organism in question were swine and the aspect of phenotype concerned were daily growth rate measured in pounds, the major portion of the P_{ij}'s would be numbers in the range 1.1 to 1.9.) Now visualizing that an individual of each genotype is grown in each environment, the symbolic numerical descriptions of the nm genotypes can be organized in a two-way table as shown below.

	Environment					
Genotype	1	2	3	j	m	Mean
1	P_{11}	P_{12}	P_{13}	P_{1j}	P_{1m}	Y_1
2	P_{21}	P_{22}	P_{23}	P_{2j}	P_{2m}	Y_2
3	P_{31}	P_{32}	P_{33}	P_{3j}	P_{3m}	Y_3
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
i	P_{i1}	P_{i2}	P_{i3}	P_{ij}	P_{im}	Y_i
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
n	P_{n1}	P_{n2}	P_{n3}	P_{nj}	P_{nm}	Y_n
Mean	E_1	E_2	E_3	E_j	E_m	μ_P

Note that μ_P is not only the mean of all values in the body of the table (the P_{ij}'s) but also of the row means (the Y_i's) and of the column means (the E_j's). Y_1, the phenotypic expression for genotype 1 averaged over all environments, is, in accord with notions elaborated above, the <u>value</u> of genotype 1. In like manner Y_2 is the value of genotype 2, and in general Y_i is the value of the i-th genotype. The value of a genotype as a deviation from the population mean will be referred to as the <u>effect</u> of the genotype and symbolized by the small case letter <u>y</u>. Thus,

$$y_i = Y_i - \mu_P = \text{the } \underline{\text{effect}} \text{ of the i-th genotype.}$$

The <u>value</u> and <u>effect</u> of environments will be defined in analogous fashion: E_j is the <u>value</u> of the j-th environment, and $e_j = E_j - \mu_P$ is the <u>effect</u> of the j-th environment.

A logical symbolic statement about an individual genotype can be written as follows:

$$P_{ij} = \mu_P + (Y_i - \mu_P) + (E_j - \mu_P) + (P_{ij} - Y_i - E_j + \mu_P) = \mu_P + y_i + e_j + (ye)_{ij}$$

$$\text{or } P_{ij} - \mu_P = p_{ij} = y_i + e_j + (ye)_{ij} \qquad (3.1)$$

in which $(ye)_{ij}$ symbolizes the effect of interaction between the i-th genotype and the j-th environment. As is obvious from the above equations, (ye) the effect of interaction, is the amount by which the deviation of the phenotype from the population mean differs from the sum of the effects of the genotype and environment that produced the phenotype in question.

It should be noted for future reference that since $\sum p_{ij} = 0$, $\sum y_i = 0$, and $\sum e_j = 0$ (because the sum of the deviations of things from their mean is always zero), $\sum (ye)_{ij}$ must also be equal to zero.

Progress in Breeding Programs

The practical measure of the success of a breeding program is the amount by which the derived product excels the one or more varieties, breeds, or hybrids that it is intended to replace in commercial use and which will be referred to collectively as the control.

Breeding programs can be classified in two categories depending on whether the selection unit is a type of family that in commercial production may be used as a "variety". Instances in which this is so well be referred to as Type A breeding programs. Examples are,

1. Corn breeding by the hybrid system. The hybrid is the selection unit and also, for commercial purposes, the "variety".
2. Breeding of asexually propagated crops (potatoes, strawberries, sugar cane, brambles, apples, etc.). The clone is at once the selection unit and, for agricultural purposes, the "variety".
3. Breeding of the naturally self-pollinated crops (oats, barley, tobacco, beans, tomatoes, etc.). The selection unit may be a pure line or a family of such lines either of which may be used as a variety.

The opposing class (Type B) is represented by programs in which the new or improved variety or breed is obtained by interbreeding the selections and is continued or expanded by intra-population crosses or matings. The selections in the breeding

program are not potential "varieties"; instead the breed or variety is a segregating population established by interbreeding the selections. Examples of this class are

1. Out-crossing programs for the improvement of strains or breeds of livestock and poultry. These include programs in which inbreeding is used to differentiate families of which selected ones are later to be recombined in a single interbreeding population as suggested by Wright (1939).
2. Mass selection and ear-row selection in corn. The selection unit is the individual or a half-sib family while the product of the breeding program is a segregating population initiated by interbreeding the selections.
3. Synthetic variety development in corn or alfalfa. In corn the selection unit is an inbred line (sometimes not completely homozygous); in alfalfa it may be either a partially homozygous inbred line or a clone. In either event the selections are interbred to establish the synthetic variety which is then continued by more or less random cross-pollination within the variety.

The significance of the above distinction is that in the Type A programs the merit of the derived product is represented by the average value of the genotypes represented in the selections of the final cycle of the program. In Type B programs, on the other hand, the derived product is a segregating population obtained by interbreeding the selections of the final cycle, and its merit is represented by the average value of the genotypes in that population. This will in general be different from the average value of the genotypes of the selections that served as the parent stock.

Resorting to symbolism, a more concise statement about progress, and one that will expedite later discussion of the prediction of progress, can be provided. Let,

C be the merit of the control

$\bar{Y}_{sn}$ be the average value of genotypes represented in selections of the final cycle of the program, and

$\bar{\bar{Y}}_{n+1}$ be the average value of genotypes in the segregating population established by interbreeding the selections of the final cycle of the program.

Then for our first class of programs total progress is

$$D = \bar{Y}_{sn} - C$$

and for the second class of programs it is

$$D = \bar{\bar{Y}}_{n+1} - C$$

The idea of total progress being the sum of that in successive cycles is made explicit as follows. Let,

$\bar{\bar{Y}}_k$ symbolize the average value of the genetic sample of the k-th cycle, i.e. the average value of genotypes represented in that genetic sample, and

$\bar{Y}_{sk}$ symbolize the average value of the selections of the k-th cycle.

Then, assuming $\underline{n}$ cycles, we can write progress in Type A programs as,

$$D = \bar{\bar{Y}}_1 + (\bar{\bar{Y}}_2 - \bar{\bar{Y}}_1) + (\bar{\bar{Y}}_3 - \bar{\bar{Y}}_2) + \ldots\ldots (\bar{\bar{Y}}_n - \bar{\bar{Y}}_{n-1}) + (\bar{Y}_{sn} - \bar{\bar{Y}}_n) - C$$

$$= \bar{\bar{Y}}_1 + \sum_{k=1}^{n-1} (\bar{\bar{Y}}_{k+1} - \bar{\bar{Y}}_k) + (\bar{Y}_{sn} - \bar{\bar{Y}}_n) - C \qquad (3.2a)$$

Now note that the difference in average value of the genetic samples of successive generation can be partitioned as follows.

$$\bar{\bar{Y}}_{k+1} - \bar{\bar{Y}}_k = (\bar{Y}_{sk} - \bar{\bar{Y}}_k) + (\bar{\bar{Y}}_{k+1} - \bar{Y}_{sk})$$

The first of these two parts is the difference in average value between the selections and the genetic sample of the k-th cycle; the second is the difference in average value between the selections of the k-th cycle and the genetic sample of the (k+1)-th cycle, which is made up of the genotypes produced by interbreeding the selections of the k-th cycle. D can now be written differently as,

$$D = \bar{\bar{Y}}_1 + \sum_{k=1}^{n} (\bar{Y}_{sk} - \bar{\bar{Y}}_k) + \sum_{k=1}^{n-1} (\bar{\bar{Y}}_{k+1} - \bar{Y}_{sk}) - C \qquad (3.2b)$$

The significance of this expression becomes most evident when it is written

$$D = (\bar{\bar{Y}}_1 - C) + \sum_{k=1}^{n} (\bar{Y}_{sk} - \bar{\bar{Y}}_k) + \sum_{k=1}^{n+1} (\bar{\bar{Y}}_{k+1} - \bar{Y}_{sk}) \qquad (3.2c)$$

Parallel expressions for progress in Type B programs are as follows.

$$D = \bar{\bar{Y}}_1 + \sum_{k=1}^{n} (\bar{\bar{Y}}_{k+1} - \bar{\bar{Y}}_k) - C \qquad (3.3a)$$

$$D = \bar{\bar{Y}}_1 + \sum_{k=1}^{n} (\bar{Y}_{sk} - \bar{\bar{Y}}_k) + \sum_{k=1}^{n} (\bar{\bar{Y}}_{k+1} - \bar{Y}_{sk}) - C \qquad (3.3b)$$

$$D = (\bar{\bar{Y}}_1 - C) + \sum_{k=1}^{n} (\bar{Y}_{sk} - \bar{\bar{Y}}_k) + \sum_{k=1}^{n} (\bar{\bar{Y}}_{k+1} - \bar{Y}_{sk}) \qquad (3.3c)$$

A word of explanation is required concerning the intended meaning of $\bar{\bar{Y}}_{k+1}$ in these expressions, in the special case when $k = n$. By our definitions it can mean the average value of either the genetic sample of the (n+1)-th cycle (which can hardly be intended since only n cycles were assumed) or the average value of the product of the breeding program, i.e. the population obtained by interbreeding the selections of the n-th or final cycle. It does not really matter, however, because the same array of genotypes is obtained regardless of whether they are to be used as the genetic sample for an additional cycle in the breeding program or as the first generation in being of an improved breed or variety for commercial use.

At this point it will be useful and instructive to examine the above expressions for progress in breeding programs with reference to the practical questions that arise in connection with the design of such programs. These expressions are complete statements about progress in the sense that the effects of any variation that can be made in breeding procedure will be reflected in one or more of their terms.

The difference $(\bar{\bar{Y}}_1 - C)$ in merit between the Control (the varieties or breeds in commercial use at the time the breeding program is initiated) and the first genetic sample in the program is always an important consideration. If it is large and in favor of the Control, then if there is to be some positive accomplishment the rate of improvement must be great or the program must be a long one. This matter frequently presents itself with special force when very wide crosses are being considered as the basis for improving a specific character of a variety that is superior in most other respects. An example is provided by the so-called triple-hybrid approach to improvement of fibre strength in cotton discussed by T. R. Richmond at the 1951 meeting of the American Society of Agronomy. Upland cotton, G. hirsutum, is an allotetraploid with 26 pairs of chromosomes. There are several wild American species of Gossypium that are potential sources of fibre strength

genes but these are all diploids with 13 chromosome pairs which in direct crosses with G. Hirsutum yield only sterile hybrids. However, by crossing two wild species and then doubling the chromosome complement of the hybrid, a synthetic allotetraploid can be produced that will yield fertile hybrids when crossed with Hirsutum (the amount of meiotic irregularity and hence the level of fertility in such hybrids varies with the wild species used). A bridge is thus provided by which fibre strength genes can be transferred from the wild species to Hirsutum. However, the triple-hybrids and their F_2 or backcross progeny while very superior in fibre strength (at least in the case of some wild species that have been used) are quite inferior in yield and certain other traits such as size and structure of bolls, i.e. with reference to overall merit $\bar{\bar{Y}}_1 - C$ is negative and large. The result is that while considerable effort has gone into this approach and progress has been recorded, it appears that it may still be a long time before strains can be derived that, all thing considered, are superior to present commercial types. There are, in fact, cotton breeders who believe the effort would be more wisely expended on genetic materials drawn from within the species Hirsutum with which the time rate of progress would admittedly be less but in which the initial level of performance, $\bar{\bar{Y}}_1$, might be very close to the Control as represented by present commercial varieties. The same problem is often faced in connection with the use of crosses with wild species as a source of disease resistance in various kinds of plants. In animal breeding a familiar example is the use of Brahma-English breed crosses as the foundation material for the development of superior strains of beef cattle. Brahman cattle bring a great deal of vigor to the cross but the early generation hybrids are typically inferior in conformation and often undesirable in temperament. In other words, with respect to conformation and temperament, $(\bar{\bar{Y}}_1 - C)$ is negative, which is a disadvantage in the use of this cross as the basis for the development of improved beef cattle.

On the other hand, $(\bar{\bar{Y}}_1 - C)$ is not necessarily negative. When two breeds of livestock have been developed for essentially the same commercial use the population obtained from inter se matings of the cross between them is frequently about as good as either of the original breeds and may be somewhat better. A number of useful new breeds have been established from such a genetic foundation. Among them are the Minnesota No. 1 breed of swine (Winters et al, 1943), the Columbia and Corriedale breeds of sheep, and the Australorp variety of chickens.

As still another example it is anticipated on theoretical grounds (and confirmed by experience; Shull, 1910) that the hybrids of inbred lines extracted from the same open-pollinated variety of corn will have average performance at least as great as that of the parent variety. And when lines extracted from different varieties are crossed the average merit of hybrids is expected to be greater than that of the parent varieties.

The second term in D (3.2c or 3.3c) is the summation for all cycles of the average amount by which selections are superior in genotype to the entire genetic sample from which they are taken. It is well known by geneticists and breeders that this difference in any single cycle of a breeding program is a function of (1) the genotypic variation among members of the genetic sample (2) the extent to which phenotypic variation stems from environmental variation, and (3) the intensity of selection. What is not so well known is the exact form of the function. This has specific significance, however, with respect to two of the important aspects of the design of breeding programs. These are (1) the selection of foundation material, i.e. the initial genetic sample, and (2) the balance to be struck between number of selection units to be observed and attention to reduction of the environmental contribution to the phenotypic variation of the selection units.

The change to be anticipated in $(\bar{Y}_{sk} - \bar{\bar{Y}}_{k})$ in successive cycles is obviously decisive with respect to the value of cyclic programs. It is generally held that except in special circumstances $(\bar{Y}_{sk} - \bar{\bar{Y}}_{k})$ should be expected to decrease in successive cycles. However, there is considerable uncertainty and probably a degree of misconception concerning the rate of decrease. In general, the rate of decrease is probably over-estimated. Lush (1943) points out that when the number of segregating genes is moderately high and environmental variation is considerable, expected values of $(\bar{Y}_{sk} - \bar{\bar{Y}}_{k})$ cannot decrease rapidly.

The final term in D is a summation over cycles of the difference in average value between the genotypes of the genetic sample and those of the selections which were interbred as the first step in creation of the genetic sample. This difference will ordinarily be negative and will vary in magnitude with the degree of non-additivity in gene action. It will tend towards zero as gene action approaches the state of being completely additive and will become large as the non-additive effects of genes become important in the variation of the genetic sample (additive and non-additive effects of genes will be defined in a very specific fashion in a

later chapter). The crux of the matter is that when non-additive gene effects are important, i.e. when the effects of individual genes vary greatly depending on what other genes are contained in the genotype, a fraction of the superiority of the genotypes of selections is the result of particularly fortunate combinations of genes. When the selections are interbred these unique combinations are broken up by segregation and the separate genes recombined in new ways. While some fraction of these recombinations may be as good or better than the origianl ones, they will on the average, be inferior. A striking example is provided when commercial varieties of potatoes are crossed to provide segregating material in which to search for genotypes superior to those of the parent varieties. While a limited number of such superior genotypes may be found, the average merit of the progeny is invariably much less than that of the parents.

The expressions for D can usefully be written in one more form. For Class A breeding systems we may write,

$$D = (\bar{\bar{Y}}_1 - C) + \sum_{k=1}^{n-1} \left[(\bar{Y}_{sk} - \bar{\bar{Y}}_k) - (\bar{Y}_{sk} - \bar{\bar{Y}}_{k+1}) \right] + (\bar{Y}_{sn} - \bar{\bar{Y}}_n) \qquad (3.2d)$$

and for Class B sytems,

$$D = (\bar{\bar{Y}}_1 - C) + \sum_{k=1}^{n} \left[(\bar{Y}_{sk} - \bar{\bar{Y}}_k) - (\bar{Y}_{sk} - \bar{\bar{Y}}_{k+1}) \right] \qquad (3.3d)$$

Three type quantities are involved. If they can be evaluated by observation or prediction, the net progress to be attained by a breeding system can be predicted. The first of these quantities, $\bar{\bar{Y}}_1 - C$, the average difference in genotypic merit between the Control and the genetic sample of the first selection cycle, is a characteristic of the specific genetic stocks involved and will usually be evaluated by observation. The other two sorts of quantities depend on the details of the breeding program and it is to the advantage of the breeder to be able to predict them as functions of the specific procedures to be followed. The first of these, $(\bar{Y}_{sk} - \bar{\bar{Y}}_k) - (\bar{Y}_{sk} - \bar{\bar{Y}}_{k+1})$ represents the average change in merit in one complete cycle or generation of the breeding program. The second, $\bar{Y}_{sk} - \bar{\bar{Y}}_k$, represented by the final term in equation (3.2d), is the average genotypic superiority of selections over the genetic sample in any specified cycle of the program. The following chapters will be devoted directly or indirectly to the problem of predicting quantities of these last two types. The procedure will be to derive

prediction formulas appropriate to different breeding plans, note the parameters involved, and develop bases for estimation of the parameters.

As noted earlier, the merit of economic plants and animals is invariably a function of sets of characters rather than of single characters. At the same time basic principles are most easily developed in terms of progress to be anticipated in a single character when only that character is considered in selection. For that reason, attention will be limited at the outset to one character situations.

Literature Cited

Dobzhansky, Theodosius (1951) Genetics and the Origin of Species, Columbia University Press, New York.

Hayes, H. K. and F. R. Immer (1942) Methods of Plant Breeding, McGraw-Hill Book Company, New York and London.

Jenkins, M. T. (1934) Methods of Estimating the Performance of Double Crosses in Corn. Jour. Am. Soc. Agron. 26:199-204.

Jenkins, M. T. and A. M. Brunson (1932) Methods of Testing Inbred Lines of Maize in Crossbred Combinations. Jour. Am. Soc. Agron. 24:523-530.

Lush, Jay L. (1943) Animal Breeding Plans. 2nd Ed. The Iowa State College Press. Ames, Iowa.

Shull, G. H. (1910) Hybridization Methods in Corn Breeding. Am. Breeders Magazine 1:98-107.

Sinnott, E. W., L. C. Dunn, and Th. Dobzhansky (1950) Principles of Genetics. 4th Ed. McGraw-Hill Book Co. New York, Toronto, and London.

Winters, L. M., R. E. Comstock, and D. F. Dailey (1943) The Development of an Inbred Line of Swine (Minn. No. 1) From a Crossbred Foundation. Jour. An. Sci. 2:129-137.

Wright, Sewall (1939) Genetic Principles Governing the Rate of Progress of Livestock Breeding. Proc. Am. Soc. Anim. Prod. 32nd. pp. 18-26.

Chapter V

The Genotypic Superiority of Selections

5.1 Introduction

The genotypic merit of selections relative to average merit of the entire genetic sample from which the selections were chosen was discussed in the preceeding chapter as a component of total progress in cyclic breeding programs. In addition it has particular importance whenever the nature of the material is such that a selection in any cycle is a potential "variety" in which the entire merit of the selection can be retained. This point is apparent from comparison of equations ( ) and ( ).

The discussion of Chapter IV was framed in terms of genotypic improvement relative to a single character, though it was pointed out that the breeder must usuall be concerned with two or more characters of the organism he is attempting to improve. We shall now generalize our notation by letting Z symbolize the index of genotypic merit, whatever it is to be, in terms of which improvement is to be measured. Z may be simply genotypic value in a single specified character or any function of genotypic values in two or more characters. In the first case, $Z = Y$. In the second, Z will be some function of two or more Y's such as $a_1Y_1 + a_2Y_2$, $Y_1^2Y_2$, or $a_1Y_1 + a_2Y_2Y$ in which the a's are constants and subscripts specify characters.

In the general discussion of this chapter it will not be necessary to specify particular cycles of the breeding program. Hence the subscript used for that purpose in the last chapter will not be employed. The mean genotypic merit of all selection units in the genetic sample will be designated simply as $\bar{Z}$ and the average genotypic merit of selections as $\bar{Z}_s$. Then $\bar{Z}_s - \bar{Z}$ will represent average difference in genotypic merit between selections and the entire genetic sample. This difference will frequent be referred to as the genotypic superiority of selections.

Natural selection is an automatic consequence of the differential reproductive rates of the individuals of a population. Hence, in nature, selection operates in terms of characters having a bearing on reproduction. The same is true in man-controlled populations but here the connection between character and reproduction will frequently be an artificial one. For example, the weight of fleece produced by sheep may have little bearing on capacity for reproduction. However, if the breeder is selecting for wool production he arbitrarily selects as parents for the next generation those individuals that have produced the heaviest fleeces and imposes zero reproduction on individuals that produced light fleeces. Fleece weight then has a tremendous bearing on reproductive rates (as allowed or imposed by the breeder) but the connection is one which stems from the objective of the breeder rather than from physiological relations. To distinguish it from natural selection, man-directed selection will hereafter be referred to as artificial selection.

The essential feature of selection in breeding programs (artificial selection) is that measured phenotypic value of the selection units is used as the basis for estimating genotypic value which itself is not directly observable. The nature of the phenotypic value employed varies with the nature of the selection unit. If the unit is an individual plant or animal, phenotype may be represented by one measurement (for example, the grain yield of a corn plant or the weight gained by a pig in the interval from 8 to 24 weeks of age) or by the average of measurements taken in successive time intervals (for example, average milk production per lactation by dairy cows or average annual yield in the case of coffee trees). If the unit is some sort of family, phenotypic value is represented by the average of measurements made on sepaeate members or subsets of members of the family (for example, the aveage egg production of pullets having the same sire, or the average forage yield of several plots of the same polycross family in alfalfa).

Phenotypic value as defined for purposes of selection in a specific breeding program, i.e. the selection criterion may also vary character-wise. As noted earlier the breeder may choose to select on the basis of any single character or he may adopt a criterion based on any two or more characters.

The magnitude of $\bar{Z}_s - \bar{\bar{Z}}$ depends on (a) the amount of variation in genotypic value and (b) the success of the breeder in identifying the best units of the genetic sample. The variation in genotypic value among units in the genetic sample depends in the long run on that in the population from which the genetic sample is drawn. Thus the population variance of genotypic values of selection units, which will be symbolized as σ_z^2, sets the upper limit of what can be accomplished by selection from the genetic population in question. For example, assuming normal distribution, 2.23% of the units in the population would excel the mean by $2\sigma_z$ or more, 0.13% would excel the mean by $3\sigma_z$ or more, and less than .01% would excel the mean for the population by as much as $4\sigma_z$. Since the limit of what can be accomplished by selection is set by the genotypic value of the best unit in the genetic sample, it rarely would be in excess of $3\sigma_z$ and would probably fall considerably short of that in many samples of moderate size.

It is often convenient to measure potential superiority of selections in percent of the mean. Assuming for practical purposes that the best selection unit contained in a genetic sample will rarely excel the population mean by more than $3\sigma_z$, the limit to accomplishment, approximated in percent of the mean, becomes $100\ (3\sigma_z/\mu)$. With this in mind Burton (1951) defined $100\ (\sigma_z/\mu_z)$ as the genetic coefficient of variation and proposed its use as a criterion of the limit of improvement through selection in a non-cyclic program. Burton defined this coefficient with reference to variation in a single character, but the idea is equally applicable to any index of genotypic merit.

While the breeder is interested in knowing the maximum that he may possibly accomplish, it is as much or more to the point to inquire how close he may expect on the average to approach that maximum. This will depend partially on the size of the genetic sample (the larger the sample the greater the chance that it will contain extreme deviates from the population mean) and partly on success in picking out the best selection units in the genetic sample. If more phenotypic information is obtained on each selection unit a better job of picking out the best units can be done. Increased information is obtained in the case of families by increasing the number of individuals per family on which measurements are made, by increasing the number of observations made on each individual if the trait is one that is expressed more than once by a single individual, by decreasing the error of measurement, or through use of a more effective selection criterion. Only the last two of these can be accomplished without reduction in the total number of selection units that can be observed without increase in work, facilities, or both. Since number of units observed is also a factor in what can be accomplished, it is clearly important to establish a basis for judging the optimum balance to be stuck between size of genetic sample and amount of phenotypic information to be obtained on each selection unit. Specifically, we need a quantitative expression for prediction, i.e. estimation, of the genotypic superiority of selections in terms of the factors which affect its magnitude.

5.2. The model.

The first step toward derivation of a mathematical relationship is specification of premises to be adopted, i.e. the model (see section 1.2). This may be done in terms of words only or in terms of words and symbols.

It is necessary to note that in practice selection is usually for genotypic value with respect to some relatively large population of environments (often only

losely defined) whereas the observations on which selection is based are made by necessity in what by comparison is a restricted set of environments. For example, the breeder of purebred swine selects on the basis of performance in the environments provided on his farm under his own management system but aspires to work genetic improvement with respect to performance in the greater variety of environments to be encountered on farms where descendants of his animals will be grown. For our immediate purposes, the aspect of this to be noted is that total environment is divisable into a portion that is constant for all selection units and a portion that is variable from one unit to another. These portions will be designated F and V, respectively. In the case of the swine breeder, F is composed of elements of environment that are constant for his farm and management system but variable among farms and management systems. All other elements are encompassed by V. Note that the environments in which this hypothetical breeder raises his swine belong to a particular (not a random) subset of the entire population of environments with respect to which genotypic value is defined. For ease of reference the latter population will be symbolized as A and any subset having common elements in addition to those of the entire population w ill be designated as A'. As an example more specific than that cited above, consider the value of corn hybrids for such a geographic area as Southern Minnesota. Here A would represent the population of environments possible in that area. Now suppose that a series of hybrids are compared, as the basis for selection, in a specific year and location within the area. Then A' is the sub-set of environments possible in that year and location. The particular environments in which the hybrids are a ctually grown represent a sample from the sub-set A' and F is composed of the elements of environment that are constant throughout all environments of the sub-set. Next, suppose the series of hybrids are compared at three locations (within the area) in each of two years. Then six subsets

of A are sampled. They can be designated as A'_{11}, A'_{21}, A'_{31}, A'_{12}, A'_{22} and A'_{32} where the first subscript refers to location, the second to year. In this case each hybrid is grown in a number of environments (oridinarily the number would be the same for each hybrid), i.e. a sample from A_1 and F would be constituted of the portion of environmental elements contributing to the sample that is constant for all hybrids.

As pointed out earlier the selection criterion may be phenotype with respect to any single character or any function of phenotype for two or more characters. Let the criterion be symbolized as X, its mean value for the i-th selection unit as $\bar{X}_i$ and its mean for the genetic sample as $\bar{\bar{X}}$.

Whatever the index, Z, of genotypic merit, its mean for the i-th selection unit will be represented as $\bar{Z}_i$ and its mean for the entire genetic sample as $\bar{\bar{Z}}$.

The assumptions of the model on which following derivations will be based are listed below.

(1) The genetic sample is randomly drawn from its parent population. The implication is that results to be obtained will be strictly applicable for only the first stage of multi-stage selection.

(2) The composition of phenotypic value with respect to any single character is

$$\bar{P}_i = \mu + f + \bar{y}_i + (f\bar{y})_i + \bar{e}_i \qquad (5.1)$$

where

$\bar{P}_i$ is the observed phenotypic value of the i-th selection unit with respect to the designated character,

μ is the mean genotypic value (with respect to A) of the population from which the genetic sample is drawn,

f is the effect of F, the constant portion of the environments in which the selection units are observed,

$\bar{y}_i$ is the average genotypic effect (with respect to A) of the member of the i-th selection unit,

$(f\bar{y})_i$ is the average of effects of interation of genotypes of members of the i-th selection unit with F,

$\bar{e}_i$ is the sum of all other effects that contribute to $\bar{P}_i$. It includes effects of random elements of environment, effects of interaction of genotype with random elements of environment, and errors of measurement.

(3) The $\bar{e}_i$'s are random members from a common population. The vital implications of this are (a) that the $\bar{e}_i$'s vary independly with respect to the $\bar{y}_i$'s and the $(fy)_i$'s, and (b) that an equal amount of phenotypic information is to be obtained on all selection units so that all $\bar{P}_i$'s can be considered equally good as estimates of genotypic values of the various selection units.

(4) The regression of $\bar{Z}_i$ on $\bar{X}_i$ is linear.

5.3. The prediction formula

Under our assumption of linear regression

$$\beta_{\bar{Z}\bar{X}} (\bar{X}_i - \bar{\bar{X}})$$

(where $\beta_{\bar{Z}\bar{X}}$ is the coefficient of regression of $\bar{Z}_i$ on $\bar{X}_i$) is an unbiased estimate of $(\bar{Z}_i - \bar{\bar{Z}})$. Averaging this quantity for all selections provides an estimate of $\bar{Z}_s - \bar{\bar{Z}}$ which may be written

$$E(\bar{Z}_s - \bar{\bar{Z}}) = (\bar{X}_s - \bar{\bar{X}})\, \beta_{\bar{Z}\bar{X}} \qquad (5.2)$$

For covenience in writing, S will from here on be used as the equivalent of $E(\bar{Z}_s - \bar{\bar{Z}})$. The quantity, $\bar{X}_s - \bar{\bar{X}}$, is the average difference in phenotypic value (as reflected by the selection criterion) between the selections and the entire genetic sample; it will be referred to as the selection differential. This differs slightly from the definition given by Lush (1943) in that his implies that by selection differential

he means the average difference between selections and the genetic population. However, while the mean difference between selections and population would rarely be exactly equal, in individual cases, to that between selections and sample, they approach equality on the average.

Now note:

(1) that $\beta_{\bar{Z}\,\bar{X}} = \sigma_{\bar{z}\bar{x}}/\sigma^2_{\bar{x}}$, and

(2) that if the selection differential measured in terms of σ_x is symbolized by k, i.e. $k = \frac{\bar{X}_s - \bar{\bar{X}}}{\sigma_{\bar{x}}}$, then $\bar{X}_s - \bar{\bar{X}} = k\sigma_{\bar{x}}$.

It is apparent that (5.2) can be written in the following equivalent forms.

$$S = E(\bar{Z}_s - \bar{\bar{Z}}) = (\bar{X}_s - \bar{\bar{X}})\,\beta_{\bar{Z}\,\bar{X}} = (\bar{X}_s - \bar{\bar{X}})\,\frac{\sigma_{\bar{z}\bar{x}}}{\sigma^2_{\bar{x}}}$$

$$= k\sigma_{\bar{x}}\,\frac{\sigma_{\bar{z}\bar{x}}}{\sigma^2_{\bar{x}}} = \frac{k\sigma_{\bar{z}\bar{x}}}{\sigma_{\bar{x}}} \qquad (5.3)$$

This expression has great generality and is very compact, only three quantities (k, $\sigma_{\bar{z}\bar{x}}$ and $\sigma^2_{\bar{x}}$) being involved. Variations in the size of the genetic sample and the nature of information on which selection is based affect S through their effects on k, $\sigma_{\bar{z}\bar{x}}$ and $\sigma^2_{\bar{x}}$. Thus if the expression is to be given its fullest significance the sources of variation in these later three quantities must be understood.

5.4. The composition of $\sigma_{\bar{z}\bar{x}}$ and $\sigma^2_{\bar{x}}$.

An initial grasp of the composition of these quantities will be gained most easily by many persons in terms of specific examples [and the exact implications of (5.3) for any specific situation can hardly be perceived without examining how $\sigma_{\bar{z}\bar{x}}$ and $\sigma^2_{\bar{x}}$ are constituted in that situation]. At the same time there is some value (doubtless varying with the reader) in more abstract considerations. A

partially abstract illustration of the make-up of $\sigma_{\overline{zx}}$ and $\sigma^2_{\overline{x}}$ will be offered here and this will be supplemented in the following chapter by illustration in terms of rather specific examples.

Consider a situation where the selection criterion is a linear function of phenotype as expressed in two characters. It can then be represented as

$$\overline{X} = c_1 P_1 + c_2 P_2$$

so that

$$\overline{X}_i = c_1 \overline{P}_{j1} + c_2 \overline{P}_{j2} .$$

The numerical subscript identifies the characters. Suppose further that genotypic improvement is to be measured in terms of some linear function in genotypic value with respect to three characters of which one is not considered in selection. The index of genotypic merit can then be written as

$$Z = a_1 Y_1 + a_2 Y_2 + a_3 Y_3$$

so that

$$\overline{Z}_i = a_1 \overline{Y}_{i1} + a_2 \overline{Y}_{i2} + a_3 \overline{Y}_{i3}$$

Finally assume that the selection units are observed in a random sample from a sub-set of environments for which the portion F of total environment is constant.

$\overline{X}$ and $\overline{Z}$ being expressed as linear functions, $\sigma_{\overline{zx}}$ and $\sigma^2_{\overline{x}}$ are easily written in terms of variances and covariances of the variables in those functions (see Chapter II).

$$\sigma_{\overline{zx}} = a_1 c_1 \sigma_{\overline{y}_1 \overline{p}_1} + a_1 c_2 \sigma_{\overline{y}_1 \overline{p}_2} + a_2 c_1 \sigma_{\overline{y}_2 \overline{p}_1} + a_2 c_2 \sigma_{\overline{y}_2 \overline{p}_2} + a_3 c_1 \sigma_{\overline{y}_3 \overline{p}_1} + a_3 c_2 \sigma_{\overline{y}_3 \overline{p}_2}$$

$$\sigma^2_{\overline{x}} = c_1^2 \sigma^2_{\overline{p}_1} + 2 c_1 c_2 \sigma_{\overline{p}_1 \overline{p}_2} + c_2^2 \sigma^2_{\overline{p}_2}$$

The various quantities involved in $\sigma_{\overline{zx}}$ and $\sigma^2_{\overline{x}}$ are typified by $\sigma_{\overline{y}_1 \overline{p}_1}$, $\sigma_{\overline{y}_1 \overline{p}_2}$, $\sigma^2_{\overline{p}_1}$, and

$\sigma_{\bar{P}_1\bar{P}_2}$. These can be further expanded if we note from equation (5.1) that $\bar{P}_{i1}$ and $\bar{P}_{i2}$ can be written

$$\bar{P}_{i1} = \mu_1 + f_1 + \bar{y}_{i1} + (f\bar{y})_{i1} + \bar{e}_{i1}$$

$$\bar{P}_{i2} = \mu_2 + f_2 + \bar{y}_{i2} + (f\bar{y})_{i2} + \bar{e}_{i2}$$

It follows that

$$\sigma_{\bar{y}_1\bar{P}_1} = \sigma^2_{\bar{y}_1} + \sigma_{\bar{y}_1(f\bar{y})_1}$$

$$\sigma_{\bar{y}_1\bar{P}_2} = \sigma_{\bar{y}_1\bar{y}_2} + \sigma_{\bar{y}_1(f\bar{y})_2}$$

$$\sigma^2_{\bar{P}_1} = \sigma^2_{\bar{y}_1} + \sigma^2_{f\bar{y}_1} + \sigma^2_{\bar{e}_1} + 2\sigma_{\bar{y}_1(f\bar{y})_1}$$

$$\sigma_{\bar{P}_1\bar{P}_2} = \sigma_{\bar{y}_1\bar{y}_2} + \sigma_{\bar{y}_1(f\bar{y})_2} + \sigma_{\bar{y}_2(f\bar{y})_1} + \sigma_{(f\bar{y})_1(f\bar{y})_2} + \sigma_{\bar{e}_1\bar{e}_2}$$

Note that the assumption (of our model) that $\bar{e}_i$ varies randomly with respect to $\bar{y}_i$ implies that $\bar{e}_i$ varies randomly with respect to genotype and hence that all covariances of $\bar{e}$'s with $\bar{y}$'s or $(f\bar{y})$'s will be zero. It is for this reason that all such covariances are neglected in the above expression. By contrast covariance between $\bar{e}_1$ and $\bar{e}_2$ cannot be assumed to be zero. The reason is that while $\bar{e}_1$ and $\bar{e}_2$ represent effects on different characters, these effects occur in a common series of individuals (the members of the i-th selection unit). Since correlation between characters in the response of individuals to their environments may exist (usually does) it would be entirely unrealistic to assume $\sigma_{\bar{e}_1\bar{e}_2} = 0$.

It will be observed that covariances of the types, $\sigma_{\bar{y}_1(f\bar{y})_1}$, $\sigma_{\bar{y}_1(f\bar{y})_2}$ are assumed to exist (are not neglected as equal to zero). In Chapter III it was specifically noted that the sum of products of y_i, value of a specific genotype,

and $(ye)_{ij}$, the effect of interaction between a specific genotype and a specific environment, when taken over the entire imaginary population of phenotypes that would be obtained of each of a population of genotypes were grown if each environment of the population of environments, would equal zero. Symbolically

$$\sum_{j}\sum_{i} y_i (ye)_{ij} = 0 .$$

It was noted on the contrary that when the product of y_i and $(ye)_{ij}$ is summed over genotypes for a single environment the sum need not be zero.

$$\sum_{i} y_i \, (ye)_{ij} \neq 0$$

It follows that if $\sum_{i} y_i (ye)_{ij}$ is summed over some subset of environments the sum again need not be zero.

$$\sum_{j=1}^{n} \sum_{i} y_i (ye)_{ij} \neq 0$$

where $\underline{n}$ is a number smaller than the total number of environments in the population.

Since covariance is the mean product between two variables, covariance cannot be zero if the sum of products is not zero.

The student will note that the above applies to single values of y and (ye), whereas the variables being dealt with in this chapter, $\bar{y}$ and $(f\bar{y})$, are means of single values. Likewise the (ye) of Chapter III is the effect of interaction between a specific genotype and the entirety of a specific environment whereas $(f\bar{y})$ represents the mean over a <u>selection unit</u> of the effects of interaction of genotypes with a portion of total environment that is constant in a subset of environments otherwise variable. Nevertheless, the issue with respect to covariances among such variables as $\bar{y}_{i1}$, $\bar{y}_{i2}$, $(f\bar{y})_{i1}$ and $(f\bar{y})_{i2}$ will not be examined further at this point. Instead we shall content ourselves here with the suggestion that the statement that these covariances need not be zero be tentatively accepted on the basis of analogy with

what has been shown concerning y and (ye).

The foregoing was presented to illustrate derivation of the composition of $\sigma_{\overline{zx}}$ and $\sigma^2_{\bar{x}}$. With this purpose in mind an example was chosen which at the same time would draw attention to all the types of quantities that can contribute to $\sigma_{\overline{zx}}$ and $\sigma^2_{\bar{x}}$ when only cases where Z and X are linear functions are considered. While it is sufficiently general to identify all the sorts of factors that bear on S (except selection intensity, the effect of which is reflected in k), it is not a particularly favorable example for use in preliminary discussion of the relative importance of those factors or as a background for relating symbolic expressions that have been developed with concepts most frequently encountered in the literature. A variety of specific issues will be considered, in terms of less general examples, in Chapter VI.

5.5. Evaluation of k

Selection as discussed in this book will always have reference to choice of the phenotypically most extreme members of a genetic sample. Usually these will be the best, but occasionally the poorest, members. The number of standard deviations by which the phenotypic mean for the selections should be expected to exceed that for the sample depends on

(a) the fraction of the sample that is selected,

(b) the nature of the distribution of phenotypic values in the parent population, and

(c) the size of the sample in which selection is practiced (this will be brought out below)

It is quite obvious that greater phenotypic superiority of selections will result if a smaller fraction of all units are selected. (For example, the best two in a sample will always have a higher mean than the best three unless the second and third are

units are identical in phenotype.) Regarding (b) consider the populations characterized by the two curves below

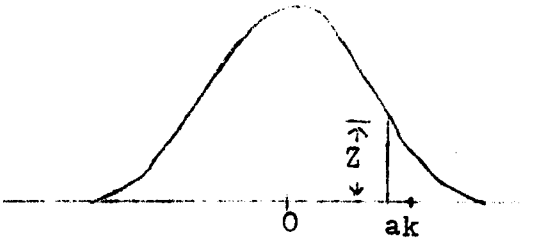
(1) (2)

Intuition tells us immediately that if small fractions of samples are being selected, k (the phenotypic superiority of selections measured in standard deviations) will on the average be at least moderately larger in the case of the population represented by the second curve.

When considering selection that has been done, the selection differential, $k\sigma_{\bar{p}}$, can be computed directly whereever complete data on the genetic sample has been recorded. On the other hand there are times when estimates of $\sigma_{\overline{zx}}$ and $\sigma^2_{\bar{x}}$ are available and it is desired to use equation (5.3) to predict the effect of selection to be practiced in the future. In such instances one can resort to use of the expected value of $\underline{k}$ for the proposed selection intensity under assumption of a specified distribution in the genetic population. (Cochran, 1951; Dickerson and Hazel, 1944; Smith, 1936). The common practice is to evaluate $\underline{k}$ assuming "normal" distribution in the parent population. This basis will usually be quite adequate for the many characters that approach "normality" in distribution, though a question can be raised when the fraction selected is very small (the reason is that there is little definite evidence that the many biological measurements that are approximately symmetrical in distribution, and are ordinarily said to be approximately "normal" in distribution actually are very close to "normality" in the tails of their

distributions). It is known of course that there are characters for which the distribution may be badly skewed uhless a scale of measurement specially devised to avoid skewness is chosen. However, evaluation of k will be considered only under the assumption of approximate "normality".

Consider first the purely hypothetical case of selection of a fraction(w) from the population itself. This can be depicted graphically by erecting an ordinate at a point a such that the area under the "normal" curve is divided into fractions w and (1-w), see below.



Let the fraction of the total area to the right of the ordinate be w, and distances on the abscissa be measured in standard deviations from the mean of the distribution. Then the mean for all individuals having values greater than a, the selected individuals, will be k and k will of course be greater than a. If we are dealing with the normal distribution, it has been shown (Pearson, 1931) that the numerical value of k will be h/w where h is the height of the ordinate dividing the two areas. Fortunately, associated values of Z and w are extensively tabled. A convenient source of such values in the Handbook for Chemistry and Physics published by the Chemical Rubber Publishing Company. The table in question is titled "areas, ordinates and derivatives of the normal curve of error". The two columns of interest are headed "area" and "ordinate". The area tabulated is (.5-w) for $w < .5$ or (w -.5) for $w > .5$. The ordinate tabulated is the one specified above. Consider selection of .05 of the population. There $w = .05$ and $.5 - w = .45$. Going to the table we find that .45 does not appear in the area column. The closest values listed are .4495 and .4505.

The corresponding ordinates are .1040 and .1023. Dividing these by $\pm$ we obtain 2.08 and 2.046. For practical purposes either is a sufficiently close approximation to the value sought but if, for any reason, more precision is desired we can use linear interpolation as follows:

$$2.08 + \left[\frac{.45 - .4495}{.4505 - .4495} \right] (2.046 - 2.08) = 2.063$$

to obtain a slightly more exact value. We find then that if the best 5 percent is selected from a "normally" distributed population the selections will excel the entire population by an average of 2.063 standard deviations.

Proceeding now to the realistic case of selection from a sample of the population, intuition tells us that the distribution of "large" samples will approach that of the population and hence that it will suffice to assume k equal to what it would be for selection from the population. There remains a question of how large the sample must be before this procedure is satisfactory. Before discussing this further we will consider "small" samples.

In their Table XX Fisher and Yates (1938) present the average values (deviations from the sample mean in standard deviations) of the ranked members of samples of sizes from 2 to 50. The table assumes samples from a normally distributed population. For samples of size 10, the value for the best member is 1.54 and for the second best is 1.00. Thus for selection of the best 2 from a sample of 10, the proper value to assign k is $(1.54 + 1.00)/2 = 1.27$. To be perfectly clear about the meaning of this value it will be stated as follows:

(a) As the number of samples from a normally distributed population is increased, the average deviation of the largest member of a sample of 10 from the mean for the sample will approach 1.54 where the unit of measure is the standard deviation of the population.

(b) The same deviation for the second largest member of such samples will approach 1.00.

(c) Since $\underline{k}$ has been defined as the expectation (value to be expected on the average) of the mean deviation (measured in terms of the population standard deviation) of selections from the sample mean, its proper value for selection of the best 2 from samples size 10 is 1.27, the mean of 1.54 and 1.00.

As further examples consider selection (a) of the one best member of samples of size 50, and (b) the five best members from samples of size 50. From the table in Fisher and Yates we find the expected devaition of the best member to be 2.25 standard deviations. The mean of the expected deviations for the best five members is (2.25 + 1.85 + 1.63 + 1.46 + 1.33)/5 = 1.704. Thus 2.25 and 1.704 are the proper values of $\underline{k}$ for the two cases. Now let us see how they compare with values that would be adopted assuming selection of 1/50 = .02 and 5/50 = .10 from the populations itself or from samples large enough to assume that $\underline{k}$ would be virtually the same as for selection from the population. For selection of .02 of the population h would be .0484 making $\underline{k}$ = .0484/.02 = 2.42. For selection of .10 of the population h is .1755 and $\underline{k}$ is .1755/.1 = 1.755. Thus for selection of 1 from samples of 50 we would overestimate $\underline{k}$ by 7.5 percent if we assumed 50 to be "large" enough to justify using the value that would apply for selection from the population. For selection of 5 from 50 the comparable overestimation would be 3 percent. In neither case is the over-estimation great enough to be critical for most applications. Thus in practice it will suffice if $\underline{k}$ is determined from the table in Fisher and Yates when selection in samples of 50 or less is being considered and under the assumption of selection from the population when samples larger than 50 are to be used. The error involved in the latter case will be samll compared to that which might arise due to "non-normality" of the parent population.

Literature Cited

Burton, Glenn W. Quantitative Inheritance in Pearl Millet (Pennisetum Glaucum). Agron. Jour. 43:409-417, 1951.

Dickerson, G. E. and L. N. Hazel. Effectiveness of Selection on Progeny Performance as a Supplement to Earlier Culling in Livestock. Jour. Agric. Res. 69: 459-476, 1944.

Smith, H. Fairfield. A Discriminant Function for Plant Selection. Ann. Eugenics 7:240-250, 1936.

Cochran, W. G. Improvement by Means of Selection. Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability. pp. 449-470. Univ. Calif. Press. 1951.

Handbook for Chemistry and Physics. Chemical Rubber Publishing Company, Cleveland.

Pearson, K. Tables for Statisticians and Biometricians, Part 2. Cambridge Univ. Press. Cambridge 1931.

Fisher, R. A. and F. Yates. Statistical Tables for Biological, Agricultural, and Medical Research. Oliver and Boyd. London and Edinburgh, 1938.