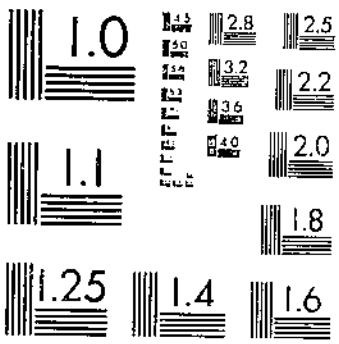
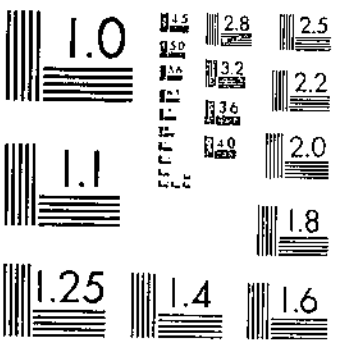
START



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

INTRODUCTION TO QUANTITATIVE GENETICS IN FORESTRY

by
Gene Namkoong, Principal Plant Geneticist
Southeastern Forest Experiment Station

Technical Bulletin No. 1588

Forest Service
United States Department of Agriculture

Washington, D. C. February 1979

Namkoong, Gene.
1979. Introduction to quantitative genetics in forestry. U. S. Dep. Agric., Tech. Bull. No. 1588, 342p.

Presents information forest geneticists need about applied mathematics, statistics, population biology, and genetics in order to design breeding programs. College-level training in mathematics, statistics, and genetics is required to understand most of the book.
KEY WORDS: Quantitative genetics, tree breeding, genetic selection, population genetics.
OXFORD NO. 165.3—015.5

ACKNOWLEDGMENTS AND DEDICATION

A diversity of people and skills is required to publish a book like this, but the distribution of credit is highly restricted. While quite inadequate for a just distribution, recognition must be given to the scientific critiques given by P. C. Burrows, W. J. Libby, J. H. Roberds, E. B. Snyder, J. P. Van Buijtenen, J. W. Wright, and the students in forest genetics at North Carolina State University. Their efforts have enhanced the quality of presentation and have reduced the multitude of errors which I had perpetrated to their present level.

In addition, this work rests on the foundation built by foresters and geneticists who pioneered in developing forest genetics, often in isolated pursuit and with meager recognition. In particular, the joyous dedication to forest genetics and to scientific truth and the gentle ethics of Philip C. Wakeley have been both directly useful and a continuing source of inspiration for more of us than he can know. It is to his great humanity that I am personally indebted, and to him that I dedicate this book.

CONTENTS

*Graduate-level statistical training required for thorough understanding.

INTRODUCTION

Man's explosive intrusion into forest ecosystems has not only affected the present character of our forests but in a more profoundly disturbing way has also affected the evolution of our future forests. Not only are the trees growing today different from those of past decades, but we have often lost the resilient capacity of this renewable resource to respond to the changing demands of nature and man. When whole forests are lost, the genes are lost, and replanting the land cannot recover the potential of any extinct genes. Even during breeding, the genetic resource may often be so reduced that future evolution is halted. Today, forests are being more intensively exploited and the forester has an obligation to safeguard the future of his resource. He can constructively direct the evolution of forests toward increased productivity within a genetic system that is capable of cumulative improvement and of meeting the varying and uncertain demands of the future.

If the genetic resource is to be effectively used and forest composition extensively controlled, we must look for ways to optimally control the evolutionary system of the whole species, and not just the transient status of any one generation. The forest scientist is thus obliged to understand the forces which have controlled or can control the evolving forest and to predict the consequences of directed or accidental changes in both the genetic and ecological systems. The potential benefits of tree breeding are widely recognized, and forest tree breeders will undoubtedly have at least partial control of the genetic basis of future forests. The forest geneticist must therefore understand the genetic materials and the manipulative techniques available. Quantitative genetics can help him to rationalize his tactics and strategies. It provides a means to construct unifying and explicit theoretical structures and testable hypotheses of alternate theories and practices.

During the past two decades, forest geneticists have devoted most attention to observing inheritance patterns, correlations among traits, and developmental relations among traits and between juvenile and mature tree performances. Much work has also been devoted to estimating the apportionment of genetic differences between and within seed sources, the utility of hybrids, and the economic and biological constraints of forest trees which affect breeding operations. Thus, the forest geneticist has begun to develop a greater understanding of the organisms handled and

a pool of materials for starting a process of controlled evolution. However, the past two decades have also produced major developments in the science of genetics and the theoretical foundations of evolutionary and breeding theory. Thus, the tree breeder often finds that his initial efforts have provided him with a good basis for directing the evolution of future forests but that there now exists a vast array of new selection theories.

This book is a guide for forest geneticists to the more useful techniques and theories of that collection of applied mathematics, statistics, population biology, and genetics which is collectively called quantitative genetics. Those parts of the theoretical and analytical techniques which can be useful in forestry are reviewed. However, there is no general review of the forest genetics literature. No detailed instructions on breeding mechanics or seed production are given, nor are many specific population or provenance studies reviewed except to illustrate how the basic principles and theories are applied. Within most chapters, a skeletal guide to the necessary concepts is given with applications to forestry. Often, topics which are not immediately applicable to forestry are discussed because of their potential future importance as our scientific knowledge increases. For the most part, the book requires undergraduate college-level mathematics, statistics, and genetics. Several special topics require a background in graduate-level statistics, but these are not essential to the continuity of subjects. Such topics are labeled with an asterisk.

The chapters are grouped into two sections. The first section is devoted to the breeding and population genetic theories applicable to forest tree breeding. The first chapter is devoted to the basic models of gene effects and genetic variances which form the basis for selection and breeding theories. Chapter 2 is devoted to the application of those statistical and population genetic concepts to the study of selection effects and how selection can be made effective in tree breeding. Then, chapter 3 considers selection theory as applied to plant breeding and tree breeding in particular. The strategy of breeding is discussed with respect to the objectives and the tactics available in chapter 4. Some special problems in developing an optimal breeding program in forestry are discussed in chapter 5.

The second section is devoted to a deeper examination of the population ecology models on which the genetic models are built and the statistical models and methods used. These areas of currently expanding research can clearly affect the breeding operations of foresters in the near future. Chapter 6 is devoted to the population ecology related to forest trees. Chapters 7 and 8 are devoted to the statistical developments which can directly affect tree breeding. Chapter 9 is a more detailed examination of population genetic theories related to forest trees. Finally, chapter 10 considers research in forest genetics needed to fulfill the forester's obligation to create an optimal evolutionary system for future forests.

CHAPTER 1
MODELS OF GENE ACTION

To the unpracticed eye, forests may at first seem to be monolithic, immutable masses with uniform shape and behavior. However, a closer look readily reveals tremendous variations in age, size, and species of trees—even a single stand of trees cannot be completely characterized by any single concept or measure. Variations exist around some average form or behavior, and an acute observer may discern a pattern in the individual-tree deviations from the norm. Scientists are interested in determining causes for some of those deviations, and they have found that clusters of performance types exist. Thus, our knowledge of the nature of forests has advanced from a perception of uniformity to a concept of an average with variation, and to an analysis of the sources of variation. In this scientific search for causes of variation, models are formulated and tested against reality, and better models formulated. In forest genetics, we have generally passed the stage of estimating means and are now estimating variations and evaluating the relative importance of different sources of variable behavior.

In this chapter, simplistic concepts of tree populations are described, along with effects of genetic differences on these populations. The concepts of mean and variance are used in population models to ascribe variation to environmental and genetic causes. The essentials of population genetics are then introduced as a basis for the subsequent chapters on selection and breeding. These statistical and population concepts are explored in greater depth in chapters 6 and 7.

DESCRIPTIVE STATISTICS

Almost any collection of trees varies considerably in a multitude of traits. The responses of trees to even the same sequence of environmental conditions usually differ sufficiently to produce recognizable variations in height, weight, color, odor, or other measurable traits including the physiological response system itself. Since trees also grow under different environmental sequences, even in managed plantations, large variations in individual tree behavior commonly exist. Silviculturists traditionally

have recognized and used some of the causes of these differences. The recent history of silvics is largely devoted to effects of such factors as age, spacing, and soil type on tree behavior. Even accounting for these major sources of variation, considerable variations still remain unexplained and can often mask even major site effects.

If general groups of behavioral types can be recognized, then it is useful to know the average performance for each group as well as how tightly clustered the groups are. Traditional and useful descriptors are the mean (μ) and variance (σ^2), which are defined as:

$$\text{Mean} = \Sigma f_i x_i = \mu$$

$$\text{Variance} = \Sigma f_i x_i^2 - \mu^2 = \sigma^2,$$

where f_i is the frequency of the i^{th} type, and x_i is the value of that type.

If the measurements are made on a continuous scale, the definitions become:

$$\text{Mean} = \int x f(x)\ dx = \mu$$

$$\text{Variance} = \int x^2 f(x)\ dx - \mu^2,$$

where $f(x)$ is the probability density function of x, and x is the value over the whole range.

Once we recognize that the population is not a single, uniform entity which can be described by a single statement, the above two descriptors often suffice for a statement of central location and degree of dispersion. However, once causes of that variation are considered to exist and hypothesized in a conceptual model to affect the trait, then the mean depends on the level of the causal mechanism and the variance depends on whether we consider the dispersion of the whole population or only that around the mean at one level of the causal factor. If soil fertility affects diameter of a tree at a given age, then we might conceive of a consistent increment in size for unit increments in fertility. The population mean and total variance measured in ignorance of soil fertility remain as they were, but the informed forester would be interested in descriptions of the mean for each fertility level and the variation around those means. He would also be interested in describing the relationship between the fertility levels and those means. The regression is a useful way to describe these relationships according to the conceived model of cause and effect, and it is a useful third measure for describing the true state of the world.

Foresters have traditionally been interested in environmental or silvicultural control of tree behavior and have frequently used regression first to describe effects of environmental factors and then to modify the forest environment for improved performance. Thus, if potassium levels, for example, affect tree size within a plantation, and if soil samples can be taken, the potential to

improve growth may exist. While other factors may continue to cause variations in size even among trees at the same fertility levels, the total variance can be partitioned into a part due to those unexplained other sources of variance, and a part due to variations in potassium. If the unexplained causes of variation are unrelated to potassium effects and if they occur independently of potassium levels, the total variance would simply be the sum of the two variances. The forester would presumably conclude that increased yields would follow from increased potassium applications, and he might even be able to eliminate that as a source of variance and have a more uniform stand. If he were a scientist, he would check his deductions against results and would likely find his initially conceived models inadequate. He might then propose better models of growth and fertilizer response and develop this branch of science.

In genetics, a similar sequence of development is involved and can be described by similar kinds of parameters. It is clear that genes do affect growth behavior, and for some populations of trees part of the variation in size is due to differences in genes possessed by individual trees. There may well exist considerable variations in behavior, even for the same genetic state, but the total variance would still be partitionable into a part due to genetic causes and a part due to other effects, such as fertility, and other unexplained differences. The forester would then also be justified in concluding that fixing the correct genes could give him behavioral improvements.

However, two major differences exist between genetic and environmental sources of variation. First, the genetic sources of variation are often caused by so many genes that, through proper breeding, they constitute a renewable resource which can continue to yield cumulative improvements. Unlike fertilizer treatments, the objective of gene management often is not just to fix the best available genotype but to use genetic recombination to generate more useful variations. The second major difference is our inability to directly observe and control most genes and hence our inability to directly create an ideal genotype, even if one could be defined. Therefore, to thoroughly understand and use the genetic resources of tree populations, we require more sophisticated concepts of breeding than simply picking and fixing the best. We must understand how genes act; we must formulate explicit models before we can establish anything near ideal breeding procedures.

GENETIC SOURCES OF VARIATION

The description of gene actions which shall be used here is based on the simple Mendelian model of two alleles at a genetically active locus and the genotypes which may thus exist. Considering genotypes to be fixed at any one time and describing the variation caused by the effect that such genotypic differences would have on average performance differences is similar to describing the variation due to any other source, such as soil fertility.

If stem-volume growth averages 100 units per tree and no recognizable fertility differences exist in the population, the trees may still vary in performance due to unknown causes. A sample of these trees might measure 101, 104, 93, 102, 97, 99, 103, etc., and carry an average near 100, a range of 11, and a variance due to unidentified causes of around 10. If the population contained variations in genetic composition such that some of the trees had a growth average of 95 units, then a sample of trees with that genotype might be 96, 98, 99, 94, 88, 92, 97, etc., carrying an average near 95 and a variance due to those same unidentified causes (residual variance) of around 10. If another genotypic variant existed at random in the same population and had an average growth of 100, a sample of its trees might be 93, 96, 99, 102, 102, 104, 101, etc., with residual variance around 10. If a third genotype existed and its trees measured 106, 109, 107, 108, 98, 102, 104, etc., averaging 105 with the unexplained residual variance of around 10, it can be observed that the total variance has increased if all genotypes are included in the same population sample. Whereas the range of variation in the initial population was from 93 to 104, the range now is 88 to 109. The actual variance in this more variable population would then depend on the relative frequencies of the genotypes. If almost all were of any one type, the variance might not be much different than originally, but if almost all were equally split between the extreme types, then the variance would be considerably larger.

Consider that the three types described above may be the three genotypic variants generated from two alleles, A and A', namely, $A'A'$, $A'A$, AA. If they were equally frequent in the population, then the trees from all types would be roughly equally sampled and the variance due to genotypic differences would be 16–2/3. The total variance for a sample, including the residual variance, would be the sum of the genetic and residual variances, 26–2/3, if genotypes were randomly located with respect to those unidentified sources of variance. More typically, however, the relative frequency of the genotypes is not equal but is dependent on other factors such as the relative frequency of the alternate alleles and mating patterns. If the alleles were equally frequent (0.5 each) and mating was random, the relative genotypic frequencies would be expected to be 0.25, 0.5, and 0.25, respectively, for $A'A'$, $A'A$, and AA. The variance due to genetic differences would be then 12.5. However, if matings were arranged such that only 0.5 $A'A'$ and 0.5 AA existed, then the variance due to there being just the two extreme types would be 25.

Gene frequency can affect the variance due to genotypic differences even with the same model of gene effects. If one allele, say A, were at very high frequency in the population, and if mating were random, then almost all trees would be of genotype AA, and few of the $A'A'$ or $A'A$ would exist or be sampled. Then, the population mean would be close to 105 and the variance not much more than the residual level of 10. The same, of course, holds true

if A were at low frequency, though then the mean would be closer to 95.

Another factor that can affect genetically caused variance is the gene-action model itself. Clearly, if the mean differences were 80, 100, and 120, the variance would be much greater than if they were 99, 100, and 101. Also, if genes acted such that the heterozygote, $A'A$, did not yield an intermediate value between the homozygotes, then the total variance would change. For example, if dominance existed, the genotypes $A'A'$, $A'A$, and AA might have values like 95, 105, and 105, and the variance can be larger than for 95, 100, and 105. If the values were 93.5, 101.5, and 103.5 and frequencies were 0.25, 0.5, and 0.25, the total genotypic variance would be 14.5 instead of 12.5 as above, even though the mean stayed at 100 and the difference between extremes remained at 10.

To describe these effects in simple models, it is useful to partition the genetic sources of variation into parts ascribable to classical types of additive and dominance types of gene action. This can be done in several ways, as detailed in chapter 7, and one particularly convenient method uses the following definitions and assumptions:

Genotypes are $A'A' : A'A : AA$

Let q=frequency of one of the alleles, say q_A.

Assume random mating, which then implies genotypic frequencies $(1-q)^2 : 2q(1-q) : q^2$.

The measured difference between $A'A'$ and AA is u so that if the variable being measured is 93.5 for $A'A'$ and 103.5 for AA, $u=5$.

The value of the heterozygote $A'A$ is $a \cdot u$, a multiple of u and a factor "a" which determines how much greater or less the heterozygote is than an intermediate, or no dominance position. If complete, classical dominance exists, $A'A$ and AA are identical, then $a=1$ and $a \cdot u=u$. If no dominance exists, $A'A$ is intermediate between $-u$, and $+u$, and $a=0$. If overdominance exists, then $A'A$ is larger than AA and its measure, $a \cdot u$, has a value larger than u, and hence $a>1$. If A'A is 101.5 exhibiting only partial dominance, as in the above example, where $u=5$, then $a=0.6$, lying between 0 and 1.

Under these conditions, the portion of the genotypically caused variance called the additive genetic variance is $\sigma_A^2=2q(1-q)u^2[1+(1-2q)a]^2$. This is the part of the total genetic variance which can be described as having been caused by the average effect of substituting one allele, say A, for the other. Hence, it is a measure of how an allelic substitution in a tree would cause variations in a tree's performance, and is similar to the variation in performance caused by a unit change in fertilizer application. In the

above example, $\sigma_A^2=12.5$. The complementary portion of the genetic variation is the dominance genetic variance:

$$\sigma_D^2=4q^2\ (1-q)^2\ a^2\ u^2.$$

This is the part of the total variation due to the heterozygotes' failure to behave in a simple intermediate manner. If variations in performance due to genotypic differences at this locus are not describable or completely accounted for by a simple model which adds a unit in yield for an allelic substitution, then dominance exists, and its effect on genotypic variance is σ_D^2. In the above example, $\sigma_D^2=2.25$.

The above two partitions of the total variation are analogous to a linear and quadratic partitioning of the variance due to any ordinary type of causal or regression variable. In a soil fertilizer experiment, it is common to use a few levels of application of a particular nutrient, say potassium, and to describe its effectiveness in terms of sums of squares or variance due to the nutrient and to the linear and quadratic portions of that variance. In such experiments, it is also common to use other nutrients such as nitrogen to study their effect on trees and to similarly describe their total effect in terms of variances accounted for or caused by those variations. It is often valuable to know the interactions among nutrient effects as well as the linear and quadratic effects of nitrogen. The form of the effect of potassium may change with nitrogen level. In a similar way, the combined effects of two genetic loci can be described even if they are not as easy to control or change as soil fertility is.

MULTIPLE-GENE LOCUS MODELS

Consider two loci with roughly the same kinds of gene action as described above. Each has some average homozygote and heterozygote yields, and hence some average effect of alleles which is measured over all variations in external environments and over all variations in genetic differences at other loci. With this model, it may be more difficult to perceive average genotypic differences at any one locus, because the background variations are larger due to genetic variations of other loci in addition to the otherwise unidentified variations. Similar gene actions would cause similar variations, but the genetic variations would include an σ_A^2 and σ_D^2 at each locus. In addition, if interactions between loci occur as between potassium and nitrogen, then additional effects and their description in terms of variances must be defined. These genetic interactions are collectively known as epistasis, and they can be statistically described as:

additive-by-additive epistasis (linear-by-linear interaction),
additive-by-dominance epistasis (linear-by-quadratic interaction), and
dominance-by-dominance epistasis (quadratic-by-quadratic interaction).

The classical genetic concepts of epistatic interaction, such as complementary or multiplicative gene action, would be reflected in the existence of variations in performance above those expected on the basis of models assuming independent gene actions.

Greater complications are introduced into the model if three loci are considered, since not only are more two-way interactions generated but triple interactions of various sorts may also exist. Such extensive models would indeed be complicated, and if we wished to analyze the detailed interactions, our problems would increase dramatically with each new locus added. Experimental models on silvicultural treatments with three kinds of variables are usually as much as can be handled, and certainly four or more variables soon become impossible to interpret. Yet in genetical situations, we often deal with effects which cannot be easily handled physically, are often masked by unidentified variations, and involve the actions of many genes. In such situations, if single-gene effects are important, the geneticist will try to isolate those effects by fixing all other sources of variations including genetic and environmental sources. More commonly, however, the single-gene effects are not easily studied and greater concern is centered on the total cumulative effect of all genes which influence a trait. Thus, if 20 loci affect growth, the statistic of main interest is the sum of variances due to the additive gene actions at all loci, or the sum of variances due to the dominance actions at all loci. The interactions may also be of interest and use, but again, with many loci, the sum of all two-way interactions, as for example all the additive-by-additive epistasis, is of greater interest than the form of the effects at any one pair.

This approach makes the description and analysis of gene effects much easier, but also submerges many substantive questions about the actual interaction of genes. Within a small range of gene actions and small changes in the frequency of the genes in any population, the consolidated statistics may accurately and consistently describe gene actions. For many breeding systems, gene frequency at each locus changes slowly, though the total impact of all loci on the phenotype may be large. Through mating and recombination, the genetic variance at each locus may change, but total variance may remain fairly constant. For any real population, however, very complicated interactions are likely to occur among loci and are likely to change whenever any one locus changes much in genotypic composition. Since foresters commonly deal with traits with fairly complicated morphogenesis, not only may many genes affect a single-behavior mode, but many physiological systems may be interacting to produce the composite trait of growth, resistance, etc. Thus, while genetic variance statistics are highly useful in condensing meaningful data and modeling population behavior, a complete knowledge of genic systems requires far deeper and more extensive research. It will eventually be necessary to recognize and study the genic interactions of traits

and how genes and physiological systems interact in composite traits.

For many practical purposes, the gross statistics of the collective genetic variances and environmental variances are useful descriptors of factors affecting forest tree behavior, and many studies have indicated that substantial amounts of genetically related variation exist. Thus, it is reasonable to consider the variation of tree behavior in most forests to be due to many effects, which are simultaneously varying. The genetic sources of this variance, which can be used and still provide more variation for cumulative gains, are the focus of interest in this book.

Unfortunately, genetic sources are complex and difficult to measure and use. Other sources of variation, such as soil fertility, can be examined by chemical and structural analysis of the soil, and the relation between those variables and growth determined by experimental control and test. Genes cannot often be measured and are generally known only by their action on the trait being measured. Therefore, instead of directly manipulating genes, other relationships have to be used to infer something about their effects.

One kind of relationship useful in analyzing the strength of genetic variation is the tendency of close relatives to be genetically more similar than distant relatives or unrelated trees. If genes have any effect on the trait being studied, then the trait should show a lower degree of variation within close family groups than between unrelated trees. Trees with the same ancestors will share more common genes and hence will behave more similarly to each other than trees with dissimilar ancestries. However, if the genes do not affect the trait being studied, then values for the trait will not be clustered within families. The geneticist, therefore, has an instrument by which he can measure the importance of genetic sources of variation. By comparing the degrees of variation between related and unrelated trees, the genetic differences can be seen to have a strong effect if the data cluster in family groups, or a weak effect if family clusters are diffuse. If the geneticist can control the degree of relatedness, an exact relationship between genetic variance and family differences can be obtained. The closer the family relatedness and the higher the genetic variance, the higher the variance between the families. If either relatedness or genetic variance is weak, the variance between families relative to that within families is small. The relationship is a multiplicative one:

$$\sigma_f^2 = r\sigma_g^2,$$

where σ_f^2 is variance among families, r is coefficient of relationship, and σ_g^2 is genetic variance.

This form of relationship is important not only for analyzing the relative strength of genetic sources of variance, but for also selecting and breeding. Therefore, before discussing the selection

process and the relation of gain from selection to gene action and to genetic variance, let us consider the experimental design and analysis possible with gene effects.

ESTIMATING GENETIC SOURCES OF VARIATION

In most studies of response to some manipulated variable, the variation caused by or attributable to the variable is separated from the variation caused by other effects or other unidentified sources. For genetics experiments in which the only controllable factor is the degree of relatedness within families, the family groups are the experimental sources of variance which can be controlled and analyzed. The degree of relationship and the strength of the genetic effects determine the physical distinctiveness of the family groupings. By knowing or controlling the degree of relationship, we can study gene effects measuring the similarity of family members. If variation is largely the result of gene effects, then close relatives like parent-offspring or sib-sib will be very similar, as compared with unrelated pairs. The variation in an offspring population will be correlated with parental-behavior variations. This condition can be expressed in terms of statistical regression as a high covariance of relatives. In such cases, the behavior of trees is predictable from the behavior of their siblings. If both parents are common between sibs (full-sibs), the covariance is higher than if only one parent (half-sibs) is the same. In turn, the half-sib covariance is higher than for more distantly related pairs.

If variation among trees is largely the result of nongenetic factors and is nearly random with respect to ancestral relationships, then the degrees of relationship can change as above, but the behavioral correlations would be lower. More variations due to nongenetic effects would reduce the measured covariance of those relatives. These relationships are derived more extensively in chapter 7 but can be summarized as follows:

$$\text{Cov (parent-offspring)} = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_{AA} + \ldots$$

$$\text{Cov (full-sibs)} = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \tfrac{1}{4}\sigma_{AA}^2 + \tfrac{1}{8}\sigma_{AD}^2 + \ldots$$

$$\text{Cov (half-sibs)} = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \ldots$$

From this point of view, the covariance of relatives reflects the relative similarity of family associations and hence increases as family groups become more distinctive due to close relationship, high genetic variation, or both. The genetic variation is reflected in the variation between family groups, which increases as the covariance of relatives increases within groups. Then, by constructing family groups, the variation between them is a measure of the covariance within families. Since the covariances are known functions of the genetic variances as given above, the genetic variances can then be estimated.

COEFFICIENTS OF RELATIONSHIP

Genetic variances can be estimated from measures of common ancestry. Measures of relationship can also indicate the degree of inbreeding from matings of relatives. Common ancestries are expressed in terms of probabilities that the trees involved have alleles derived from common ancestors. Consider that for any two individuals, a covariance would exist and can be written in terms of genetic effects if there is some probability that identical genetic effects occur other than solely by chance in random mating. If pairs of individuals are randomly chosen from a large population, then their alleles are expected to occur in the frequencies expected of the general population. If the pairs have closer relationship, then the degree of nonrandomness can be measured by the probability that the alleles in the two individuals are identically derived and exactly alike. Thus, for a linear model of average and dominance effects, as we have previously defined, we can derive the covariance between two individuals, X and Y, according to the probabilities that their alleles are the same:

$$X=\mu+\alpha_{x_\delta}+\alpha_{x_\varphi}+\delta_{x_\delta x_\varphi}$$

and $$Y=\mu+\alpha_{y_\delta}+\alpha_{y_\varphi}+\delta_{y_\delta y_\varphi}$$

where α_{x_δ} is average effect of male parent gene contributed to x, α_{x_φ} is average effect of female parent gene contributed to x, α_{y_δ} is average effect of male parent gene contributed to y, α_{y_φ} is average effect of female parent gene contributed to y, $\delta_{x_\delta x_\varphi}$ is dominance deviation of parental genes contributed to x, and $\delta_{y_\delta y_\varphi}$ is dominance deviation of parental genes contributed to y.

As developed in chapter 7, $\sigma_\alpha^2=\frac{1}{2}\sigma_A^2$, and $\sigma_\delta^2=\sigma_D^2$. If the male parentage of X and Y is identical, nonrandom, or related in some way, then a certain probability exists that $\alpha_{x_\delta}=\alpha_{y_\delta}$, and the covariance of X includes $Pr(X_\delta=Y_\delta)\ (\frac{1}{2})\ \sigma_A^2$. If the female parentage was somehow nonrandom or related, the $Pr(X_\varphi=Y_\varphi)\neq 0$ and the variance contains $\Sigma_{ij}Pr\ (X_i=Y_j)\ (\frac{1}{2})\sigma_A^2$. Note that if we took the probability of a random allele from x and random allele from y being identical by descent, this probability is $\Sigma_{ij}\ \frac{1}{4}\ Pr(X_i=Y_j)$, which is Malécot's (1969) coefficient of coancestry f_{xy}. Therefore, $2f_{xy}=\frac{1}{2}\ \Sigma_{ij}Pr(X_i=Y_j)$ which can be used as the coefficient for the σ_A^2 contribution to the covariance of relatives. If both male and female parentage of X and Y are related, then $E(\delta_{x_\delta x_\varphi}\delta_{y_\delta y_\varphi})=Pr(x_\delta=y_\delta$ and $x_\varphi=y_\varphi)\ \sigma_\delta^2+Pr(x_\delta=y_\varphi$ and $x_\varphi=y_\delta)\ \sigma_\delta^2$. Then for any kinds of relationship, we can trace the various probabilities and determine the contributions of these genetic variances to the covariance of relatives. For

example, if the female parent of X and Y were the same, then the only nonzero probability would be $Pr(X_{♀}=Y_{♀})$, and it would depend on how the choice of gametes is made in the production of eggs of the common mother. If the choice is random, then the probabilty is ½ that the same allele (either one) is chosen and the contributions of the genetic variance to the covariance of these half-sibs are ¼ σ_A^2. If both the male and female parents of X and Y were common, then $Pr(X_{♀}=Y_{♀})=Pr(X_{♂}=Y_{♂})=½$ and the probability that both are identical is $(½)\cdot(½)=¼$, and the other probabilities are zero. Therefore, the genetic variance contributions to the covariance of full-sibs is ½ $\sigma_A^2+¼$ σ_D^2. For the case of parent-offspring covariances, if we take the parent as X and the offspring as Y, the $Pr(X_{♂}=Y_{♀})=Pr(X_{♀}=Y_{♀})=½$ and all other probabilities are zero. Then the covariance of parent and offspring is $(½)$ σ_A^2.

If additional genetic loci affect the genetic variances and covariances among relatives and if they are independent loci, then the probabilities of identity by descent for multiple-locus effects can be added over the genetic variances at each locus. For multiple-locus epistatic effects, the probabilities of joint identities by descent are products of the independent probabilities. In such cases, for any kinds of relatives which have the additive genetic variance coefficient of a and a coefficient for σ_D^2 of d, the general covariance due to all types of genetic variance can be written as:

$$\text{Cov}=a\sigma_A^2+d\sigma_D^2+ad\sigma_{AD}^2+a^2\sigma_{AA}^2+d^2\sigma_{DD}^2+a^2d\sigma_{AAD}^2+\ldots$$

or in general $\text{Cov}=\Sigma_{i,j}\, a^i d^j\, \sigma^2_{A^iD^j}$.

Inbreeding nullifies the independence assumptions and the probabilities of drawing identical alleles. It is clear, for example, that if F is defined as the probability that the two alleles at a locus are identical by descent, the probability that two randomly drawn alleles are identical is ½ $(1+F)$ instead of $(½)$. With a parental inbreeding coefficient of F, even with random choice of parents and hence no inbreeding of the offspring, the a and d coefficients used to compute the covariance of relatives are increased by factors of $(1+F)$ and $(1+F)^2$, respectively. The problem remains, however, that the σ_A^2 and σ_D^2 themselves require specification with respect to the inbreeding generation they refer to.

Linkage can also affect the probabilities of some gametic combinations, the contributions of the epistatic gene effects, as well as how the additive variances are summed over loci. The manner in which they affect the covariance of relatives is not an easily derivable relationship (Cockerham 1956). Nevertheless, if we wish to exactly define and estimate meaningful parameters, the broad effects of such factors as linkage and inbreeding must be considered.

It is also clear that hybrid populations will engender genetic variances and covariances among relatives with quite unique

effects and probabilities of drawing gametic contributions. The effects of dominance types of intralocus gene actions are unique, and all types of interlocus epistatic interactions are unique since the entire genome is a hybrid combination. In addition, gametic frequencies depend on the differences in gene frequency between the populations and on the linkage disequilibrium so induced (Stuber and Cockerham 1966). In our brief review, all of these effects will be neglected and we shall assume large random-mating populations with independent loci.

DESIGNS FOR ESTIMATION

In the kinds of designs useful with forest trees, it is often possible to derive estimates of variances due to family differences where families are structured into half- or full-sib groups. For example, if female parents are chosen and a different set of male parents is chosen for each female, then the variation among offspring in different female parent groups is the same as the variation among half-sibs. Similarly, the variations among the families of different males within the same female family group is the variance among full-sibs within half-sib groups. It is thus the variance among full-sibs, less the variance among half-sibs. If both estimators are available, then we can estimate as follows:

$$\text{Variance (female half-sibs)} = (\tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \ldots)$$

$$\text{Variance (male full-sibs within female half-sibs)} = (\tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \tfrac{1}{4}\sigma_{AA}^2 + \ldots - \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \tfrac{3}{16}\sigma_{AA}^2 + \ldots)$$

Thus, the female family variance contains only ¼ of the additive genetic variance and a small fraction of additive types of epistasis, and the male family variance contains that much plus ¼ of the dominance variance. The difference between them therefore contains ¼ of the dominance genetic variance and small fractions of the epistatic variance.

Since many experimental mating designs can be constructed to provide similar estimates, populations can be examined for their genetic sources of variation. Not only are analysis of variance estimators available, but regressions of offsprings, clones, etc., on parental performances also allow one to estimate the variances. Since precise estimates require large experiments, efficient experimental design is highly desirable. For purposes of this chapter, recognition of the existence, descriptive forms, and estimability of genetic variance parameters are sufficient.

Using various experimental procedures, large estimates of genetic sources of variation in forest trees have often been derived. How have forests evolved such a system? It behooves us to consider the mechanisms by which variations are generated and maintained. An understanding of the dynamics of forest systems is desirable for its own sake as well as to help us design more

efficient manipulative mechanisms to serve the long-term interests of forests and man.

POPULATION GENETIC BASIS

The basic forces which have molded the system of genetic variability have been mutation, migration, selection, and random events. Mutation has rarely been successfully used in breeding programs and though it is the basic originating mechanism for new alleles and is occasionally useful, it will be ignored in this chapter.

Migration, or its lack, and consequent subdivision of the population into intraspecies subgroups, has also been an important factor in evolution but is not a significant manipulative factor except for constructing or crossing among subpopulational groupings. The lack of complete migration of genotypes through a species leads to separate evolutionary paths being taken by subpopulations as they respond to selection differences or chance sampling events. The use of variations among these subdivisions directly as in provenance selection or as a source of genetic variation is a useful initial stage of breeding and deserves detailed analysis. However, we shall consider the directive forces of selection within any given population as the basis for understanding selective breeding effects.

In simple models of selection where the effect of a gene is easily recognized, the breeder either simply fixes the good homozygote by crossing only among the good genotypes or breeds the heterozygote by crossing the different homozygotes. While dominance effects may mask the heterozygote, the breeding procedures are simple and the genetic problem is solved in one or relatively few breeding generations. In natural selection for reproductive fitness, selection operates by eliminating defective genotypes. However, environmental effects or genes at other loci cause some errors in artificial or natural selection. These errors occur because the phenotypic expression is different from the average genotypic expression of the locus, because the selection process is not deterministically exact, or because of both factors. In any case, a slower process of allelic substitution occurs, and the average changes in progress to higher fitnesses, or more economically valuable trees, occur in smaller steps each generation.

Considering a single locus with two alleles A and A' and its three genotypes AA, $A'A$, and $A'A'$, the change in value from one generation to the next depends on having more of the preferred genotypes present. If AA is preferred over $A'A'$, or has a higher probability of being selected, then the contribution of parental trees with AA to the progeny generation will be higher, the A allele will be more frequent, and hence AA genotypes will often be more heavily represented in the next generation.

Two genetic factors influence the rate of progress, the relative probabilities of selection or fitness of the genotypes, and the gene frequencies. The greater the differences between genotypic fit-

nesses, or in precision and intensity of selection, the greater the change will be in any one generation. The only complicating factor would be the existence of dominance which might mask the effect of an otherwise unfavorable allele. In the case of overdominance, the best genotype is the heterozygote and the general tendency will be for the population to stabilize gene frequency at some intermediate level. Otherwise, selection in a consistent direction will tend to eventually fix the favored allele in the population, and, in the absence of mutation or immigration, eliminate the other allele.

The change in gene frequency (q) in response to selection pressures also affects the rate of change. The change in gene frequency is a function of the change in fitness and a factor of $q(1-q)$. This is a quadratic function with a maximum at $q=\frac{1}{2}$, and zero value at $q=0$ or $q=1$. Thus, the most rapid changes in gene frequency, and hence the most rapid changes in population fitness, occur in the intermediate ranges of q. Since the actual response depends on the fitness levels, dominance, etc., the rate of change may not be symmetrical with respect to gene frequency, but only when the frequency is intermediate can rapid response to selection be expected. We can further imply that genes involved in selection will exhibit most rapid frequency change when frequency is intermediate and therefore will not usually be found in the intermediate frequency range unless strong dominance to overdominance exists, or unless selection is in a transient state.

These simple models have served as good first approximations but they have some obvious shortcomings. Often, as in competitive situations, a genotype's fitness depends on its own relative frequency and hence frequency-dependent selection models require examination. Discussion of this problem is postponed to chapter 2. A further obvious complicating factor is that genes rarely act alone, and in almost all investigations highly intricate developmental pathways exist and require that gene actions be coordinated. Even if the linear gene-action models are accepted within small changes in gene frequencies in a single physiological system, interactions among the genes of the multiple systems must exist.

When considering even just two loci, the obvious results of the one-locus case cannot be generally extended. Not only does physical linkage between genetic loci affect selection, but the dual factors of epistasis and linkage can form several intermediate frequency equilibria when an analysis of the individual loci would not reveal that possibility. It is also possible that selection would not maximize fitness as in the single-locus case, and hence that intermediate frequencies for the loci may be stabilized at less than optimum frequencies. Hence, in the natural evolution of populations, one-locus analyses may not reveal the reasons for the existence of stable, intermediate gene frequencies maintained by selection. Thus, not only can selection cause stable equilibria, but directional selection as practiced by man may be adversely affected.

If populations have many mechanisms for continually generating variations, they also have others by which genetic variations are lost. In addition to directional selection, the accidental loss of genes from small populations leads to a reduction of variation at least in the local population. The smaller the population, the greater the chance that an allele or a genotypic combination of alleles can be lost. If there is a 10-percent chance of a gene being represented, and only a few trees are sampled, there is a reasonable finite probability that the gene will be lost in one or a few generations. Since many investigations on trees have indicated that small population subdivisions exist, even in continuous stands (Sarvas 1963; Sakai 1971), it is possible that sampling variations have affected the evolution of variation patterns in many forests.

In natural selection, average selective values may indicate the probabilities of a tree's surviving and reproducing "on the average." However, any one tree either reproduces or it does not, and indeed any group of trees with the same selective values may totally fail or succeed. Thus, the average statistics are accurate only for large populations or for many repeated trials of small groups. If the relative selective values of AA, $A'A$, and $A'A'$, for example, are 1:1.5:1, we can expect that an average gene frequency of ½ would exist, and that AA and $A'A'$ would exist in equal frequency. However, if only a single small population was reproduced, it would eventually be either all AA or all $A'A'$, with no $A'A$ heterozygotes, due to natural inbreeding. AA and $A'A'$ would not coexist in the small population. If many such small groups were isolated, each would be either AA or $A'A'$, and though they might have the same frequency if all groups were counted, no $A'A$ would exist. Thus, any group of trees classified by variety, age class, genotype, or alleles may be lost even though selection favors their survival.

The accidents of sampling in small populations can therefore cause more rapid fixation of an allele than might be expected from selection effects alone. In fact, even if an allele is favored by selection, it can be lost by accident, especially if it initially occurs at low frequency. Similarly, the effects of mutation, migration, dominance, and epistatic gene actions can be modified by sampling variations in small populations. In general, more extreme allelic frequencies, fixations of favored or unfavored alleles, and less stable frequencies over populations or generations can be expected.

A balance among the simultaneous effects of selection, migration, mutation, and sampling error is struck in the natural evolution of populations, and the gene system itself may slowly respond to any changes in selective pressures. For breeding purposes, the gene frequencies made available by the natural processes are the raw materials for manipulating future evolution. The limitations on selective breeding imposed by sampling errors are important to consider in deciding how intensive selection should be.

CHAPTER 2
SELECTION THEORY

Since selection has affected evolution and can be used to direct future evolution of populations, the study of selection and its effects has absorbed more interest and effort than any other genetic force. Still, the relationship between the choice of a subset of all potential parents for regenerating future populations and its actual effect on changing genotypic frequencies and on eventually changing a population's phenotypic distribution is a complex of interacting factors that remains poorly defined. In this chapter, we shall investigate the theories of how selection affects populations and the various parameterizations that have been useful in studying the effects of selection. Simple one- and two-locus models of classical types of gene actions are very simply modeled for cases where such simple actions and environmental factors affect phenotypic performance. Since average phenotypic performance, which is genotypic potential, is rarely exactly achieved, variability causes some difficulty in determining the genotype from the phenotype. The effects of selection on the basis of phenotypic measures are therefore modeled as a probabilistic process which, while inexact, would have an expected change on the gene frequency of the selected versus the unselected population. The consequent effect on population mean improvements in the short and long runs is then examined in terms of the effects of N_e (effective population size), heritability, and selection intensity on the improvement. In addition, the general breeding methods which have been developed in light of their relation to selection theories are briefly examined.

SINGLE-LOCUS MODELS

We can work most simply with a one-locus genetic model. In classical genetic theory, the only problem in selection forcing the population into homozygosis for the favored allele or some preferred intermediate frequency is the time it takes to arrive at the stable state. In the simplest case in which genotypes can be phenotypically recognized and easily distinguished, selection for the best homozygote or for an overdominant heterozygote condition is direct and immediately produces the desired population. Only under complete dominance would an "undesirable" allele remain in the population but that can also be eliminated by simple test crossing and selection. To more exactly determine the progress

that can come from selection, consider that generally, each genotype may have some possibility of reproducing in the next generation. From a single genetic locus with two alleles A and A', the three genotypes would then have selective values of $r_{AA}:r_{AA'}:r_{A'A'}$, respectively. We can then define an average selective value for the whole population according to the r values and their respective frequencies as $\bar{r}=p^2r_{AA}+2p(1-p)r_{AA'}+(1-p)^2r_{A'A'}$ where p is frequency of the A allele. We might also define an average selective value of an allele according to the frequency and the average effect that it has in the zygotes as:

$$r_A=pr_{AA}+(1-p)r_{AA'} \text{ and } r_{A'}=pr_{AA'}+(1-p)\,r_{A'A'}.$$

Then $\bar{r}=pr_A-(1-p)r_A=p^2r_{AA}+2p(1-p)\,r_{AA}+(1-p)^2r_{AA}$.

Also, the variance among the average effects r_A *and* r_A is

$$pr^2_A+(1-p)r^2_A-\bar{r}^2, \text{ which equals } p(1-p)\,(r_A-r_{A'})^2.$$

We can now analyze the changes in selective values by noting that $\bar{r}$ is a function of gene frequency and

$$\frac{d\bar{r}}{dp}=(r_A-r_{A'})+\frac{pdr_A}{dp}+\frac{(1-p)\,dr_{A'}}{dp}.$$

Since $\frac{dr_A}{dp}=r_{AA}-r_{AA'}$, and $\frac{dr_{A'}}{dp}=r_{AA'}-r_{A'A'}$,

then $\frac{d\bar{r}}{dp}=(r_A-r_{A'})+p(r_{AA}-r_{AA'})+(1-p)\,(r_{AA'}-r_{A\,A'})$.

Then $\frac{d\bar{r}}{dp}=2(r_A-r_{A'})$.

Since it is also true that,

$$\frac{dp}{dt}=p(r_A-\bar{r})=p(1-p)\,(r_A-r_{A'}),$$

we can see that,

$$\frac{d\bar{r}}{dt}=\frac{d\bar{r}}{dp}\,\frac{dp}{dt}=2p(1-p)\,(r_A-r_{A'})^2$$

which is simply twice the variance in average selective effects.

It is particularly interesting to examine the $\frac{dp}{dt}$ function, since it would indicate the location of potential stationary points where p does not change with advancing t. It also indicates that the rate of change in frequency and fitness with respect to time is partly controlled by the factor $p(1-p)$, which is a symmetrical quadratic function of p with a maximum at p close to ½. Hence, intermediate values of p will always force $\frac{d\bar{r}}{dt}$ to be high, and $\frac{dp}{dt}$ to also be

high relative to the extreme values of p. Hence, p and r change most rapidly when p is intermediate and slower when p is close to zero or one. Furthermore, if a profile of gene frequencies is made among loci which have been subject to selection at one time or another, the great majority of loci will have moved their gene frequencies through the middle ranges and would now be at low or high frequency. This implies that selection is most effective on genes of intermediate frequency and that we cannot ordinarily expect to find many loci kept at these frequencies by directional selection.

However, even the $\frac{dp}{dt}$ function can be described in terms of $\frac{d\bar{r}}{dp}$ as:

$$\frac{dp}{dt}=\frac{p(1-p)}{2}\frac{d\bar{r}}{dp},$$

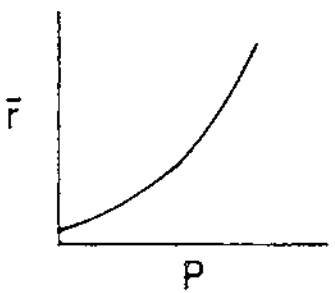
and hence the movement of p can also be analyzed in terms of the relationship between selective value and gene frequency. Since $\bar{r}$, r_A, and $r_{A'}$, are all functions of p, and the three genotypic values r_{AA}, $r_{AA'}$, and $r_{A'A'}$, which we assume are fixed, we can describe $\bar{r}$ in terms of variations in p for given relative values of the three zygotic r's.

To see the effects of selective values on changing gene frequencies, we can follow several sets of relations among the r's, for $\bar{r}$, and $\frac{d\bar{r}}{dp}$, since

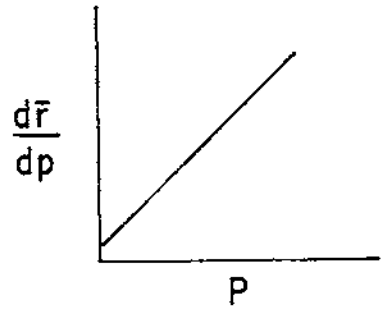
$$\bar{r}=pr_{AA}{}^{2}+2p(1-p)r_{AA'}+(1-p)r_{A'A'}{}^{2}, \text{ and}$$

$$\frac{d\bar{r}}{dp}=2p(r_{AA}-2r_{AA'}+r_{A'A'})+2(r_{AA'}-r_{A'A'}).$$

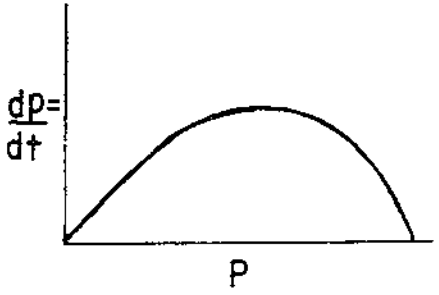
If, $r_{AA}>r_{AA'}>r_{A'A'}$, then r increases monotonically with p in a form like



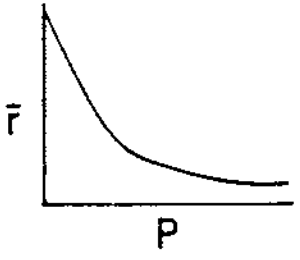
and $\frac{d\bar{r}}{dp}$ is a linear function of p in a form like

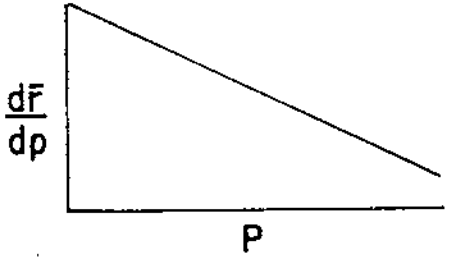


and $\frac{d\bar{p}}{dt} = \frac{p(1-p)}{2}\frac{d\bar{r}}{dp}$ is

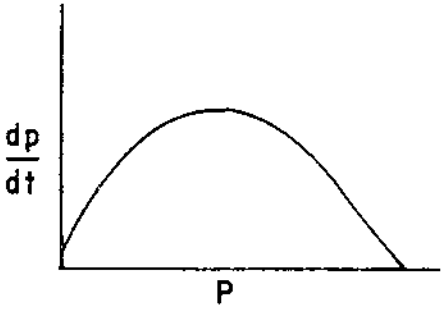


If, $r_{AA} < r_{AA'} < r_{A'A'}$, the reverse relationships hold for similarly scaled r values:

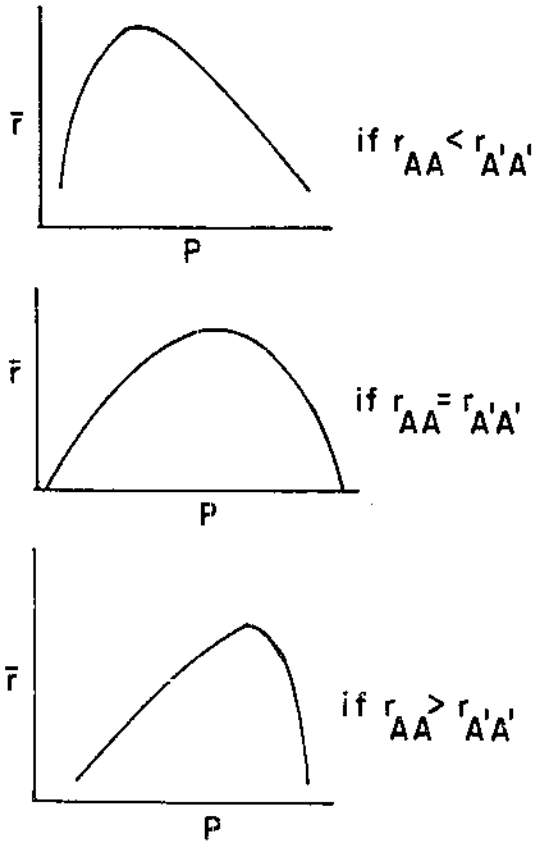




and

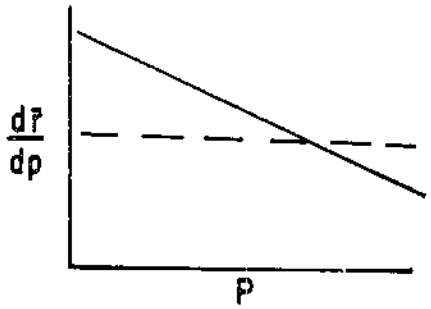


If, $r_{AA} < r_{AA'} > r_{A'A'}$, the $\bar{r}$ has a stable peak at an intermediate p with a maximum to the left or right of $p=0.5$ according to whether $r_{A'A'}$ is greater or less than r_{AA}:

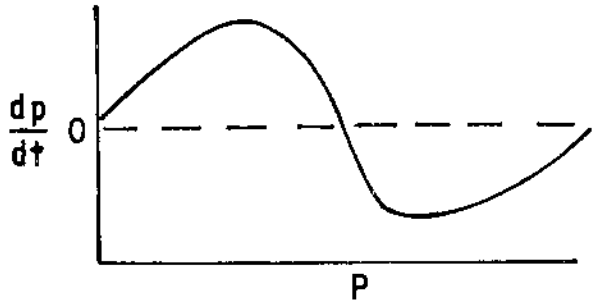


$\frac{d\bar{r}}{dp}$ remains a linear function of p but now must be scaled to cross zero to the left or right of $p=0.5$ according to whether $r_{A'A'}$ is greater or less than r_{AA}.

Generally,

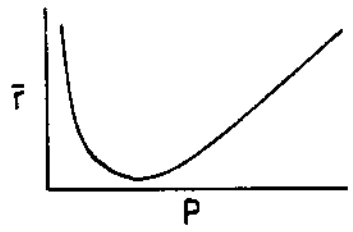


and hence

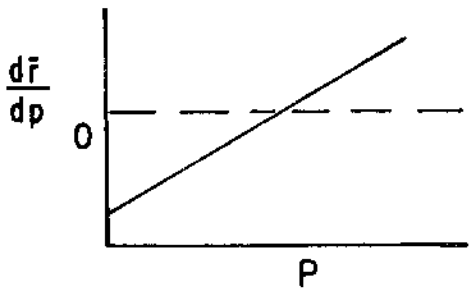


In this case we can also notice that if $\frac{d\bar{r}}{dp}=0$, and if we use $r_{AA}=1-s$, $r_{AA'}=1$, $r_{A'A'}=1-t$; that $p=\frac{t}{s+t}$ represents the equilibrium point for p.

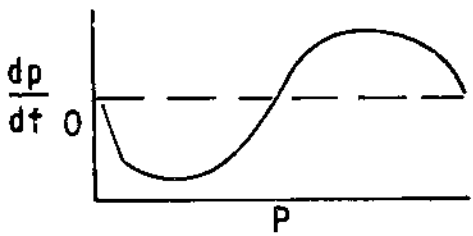
If $r_{AA}>r_{AA'}<r_{A'A'}$, the reverse relationships exist and generally,



and



and



which has an unstable equilibrium at intermediate p.

In all of these cases, the only stable equilibrium, except at $p=0$, exists when $\overline{r}$ is at a maximum. The last case is the only one in which an intermediate p value exists where $\frac{dp}{dt}=0$, but in which it is also clear that small displacements of p from that point cause the p to go either to 0 or 1. At the equilibrium p, p_e, and a small change to the right makes $\frac{dp}{dt}>0$ and hence forces p to go further towards 1, and a small leftward change from p_e makes $\frac{dp}{dt}<0$ and hence forces p further towards 0. In the immediately preceding case of overdominance, small changes from p_e can be seen to have the opposite effect on $\frac{dp}{dt}$ and hence to force p back to p_e.

It can also be seen at the equilibrium points of frequency except at $p=0$, and $\frac{dp}{dt}=0$, that if $\frac{dp}{dt}=p(r_A-\overline{r})=p(1-p)(r_A-r_{A'})$ that $\overline{r}=r_A=r_{A'}$.

If particular values for gene frequency and the r's are known in populations with discrete generations, more exact analyses of changes in gene frequency can be made simply by following the selective process, one generation at a time. The process involved is to find the gene frequency of the generation following selection, in terms of the selection and gene frequency prior to selection, and to then write the relationship in the form of a difference, or recursion equation. Thus, as in the third case as examined above, if the heterozygote is favored and $r_{AA}=1-s$, $r_{AA'}=1$, $r_{A'A'}=1-t$, then:

Zygote	Initial proportions	Selection proportions	Proportions after selection
AA	p_o^2	$1-s$	$p_o^2(1-s)$
$A'A$	$2p_o(1-p_o)$	1	$2p_o(1-p_o)$
$A'A'$	$(1-p_o)^2$	$1-t$	$(1-p_o)^2(1-t)$

The A alleles come from the AA parents with frequency $p_o^2(1-s)$ and from half of the AA' for a total new relative frequency of

$$p_1=\frac{p_o^2(1-s)+p_o(1-p_o)}{p_o^2(1-s)+2p_o(1-p_o)+(1-p_o^2)(1-t)}$$

The A' allele's frequency can be derived similarly as:

$$1-p_1=\frac{p_o(1-p_o)+(1-p_o)^2(1-t)}{p_o^2(1-s)+2p_o(1-p_o)+(1-p_o^2)(1-t)}$$

These formulas can be simplified to:

$$p_1=\frac{p_o(1-sp_o)}{1-sp_o^2-t(1-p_o)^2},$$

$$1-p_1=\frac{1-p_o-t(1-p_o)^2}{1-sp_o^2-t(1-p_o)^2}$$

We might note that the change in frequency is:

$$p_1 - p_o = \frac{p_o(1-sp_o)}{1-sp_o^2-t(1-p_o)^2} - p_o = \frac{-p_o(1-p_o)\,[sp_o - t(1-p_o)]}{1-sp_o^2-t(1-p_o)^2},$$

$$(1-p_1) - (1-p_o) = \frac{p_o(1-p_o)\,[sp_o - t(1-p_o)]}{1-sp_o^2-t(1-p_o)^2}$$

For any generation of selection, the change is similarly formulated and the A or A' allele can gain or lose in frequency according to the sign of $sp_o - t(1-p_o)$. If sp_o is greater than $t(1-p_o)$, the A' allele gains in frequency. If sp_o is less than $t\ (1-p_o)$, the A allele gains. And, if $sp_o = t(1-p_o)$ the change is zero, and from this condition,

$$sp_o + tp_o = t$$

$$p_o = \frac{t}{s+t}, \text{ and } 1-p_o = \frac{s}{s+t},$$

as previously derived. Other equations for other gene action models are detailed in several texts (Li 1955).

The foregoing selection models assume that each genotype has properties which predispose it to given selection frequencies. This is a kind of "soft" selection among genotypes in which selection is in proportion to genotypic propensities for success. A different model of selection is a kind of "hard" selection in which individuals are selected if they perform over a minimal level regardless of how many may be so selected. In breeding practice, a level of phenotypic performance is often determined when genotypes cannot be easily distinguished, and any tree exceeding the specifications is accepted for further breeding. On the other hand, if a certain proportion of selection is fixed, the breeder is implicitly following a "soft" selection procedure.

If selection thus actually operates on the phenotypic level, as is most often the case, then other parameterizations of selection probabilities can be made in terms of phenotypic distributions. Thus, a commonly used model of gene effects would specify a mean effect for a genotype and some distribution of phenotypes expressed for that genotype with a variance σ^2. The probability of selection will differ among genotypes according to the differences among the means as well as the relative size of σ^2 with respect to the mean differences. If σ^2 is relatively large, the selective probabilities will be similar, regardless of genotype, while if σ^2 is relatively small, there might be little error in assuming that specific genotypes are recognizable and are being selected. If we cannot attach high probabilities of selection to genotypic differences, then we admit a certain degree of error in choosing optimum genotypes. Consider, for example, a genotype with mean productivity value of 1,000 units and a variance (σ_e^2) of 1,000 due to various internal and external environmental variations in expressing its average productive capacity. If this variance of 1,000 has no genetic basis, then, of course, selection of the higher

yielding trees will yield no genetic gain. If three genotypes existed, AA, AA', and $A'A'$, with the same variance but with average production capacities of 1,005, 1,000, and 995, respectively, then the variation among genotypic effects can be used. Assuming that gene frequency $q=0.5$ and that there is random mating in a large population, then the genetic variance is all additive and equals $\sigma_A^2=12.5$. For the total population which has a mean of 1,000 and a total variance of 1,012.5, selection of all trees above say 1,050 would be expected to truncate the population as in figure 1.

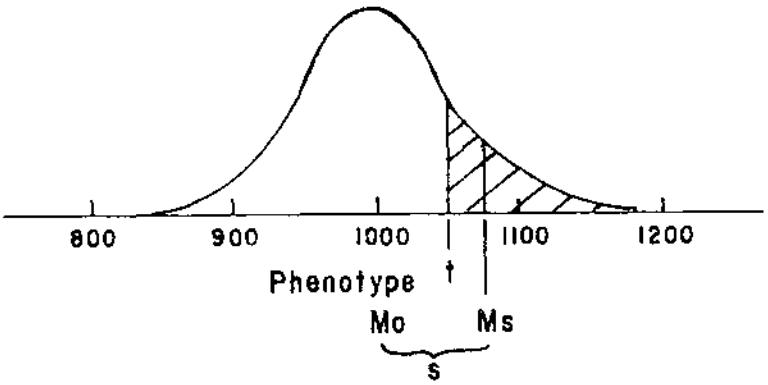


Figure 1.—A normal distribution of tree values around a mean of 1,000 and variance of 1,000, with truncation selection above phenotypic value t.

Since the three genotypes differ in average effect, however, the expected truncation includes different proportions of the expected genotypic distributions as shown in figure 2. It can be seen that

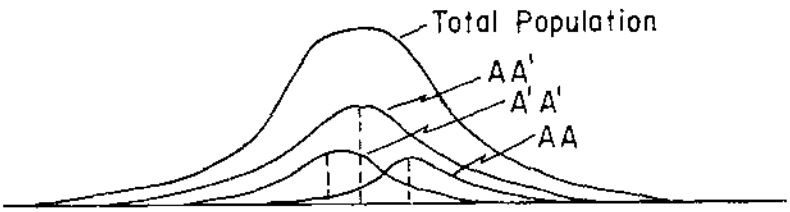


Figure 2.—Relative numbers of three genotypes from a population with a normal frequency distribution of random variations around genotypic means generated by additive gene action and gene frequency 0.5.

while the heterozygote still is relatively heavily represented in the selected portion, the favorable homozygote is more heavily represented than the unfavorable homozygote. If wider mean differences among the genotypes existed relative to the error variance, then the proportions expected in the selected populations

would even more heavily favor the AA genotypes. If the frequency of the A allele were higher, then proportionately more of the AA would be selected over the AA' and $A'A'$ genotypes, but the change in relative gene frequencies may be slower. If only the most extreme phenotypes were selected, then the relative gain in gene frequency would be further increased. Thus, the frequency of the A allele is expected to increase according to the mean differences among genotypes, their error variances, and the selection intensity; and the selection effect on the locus is a function of all three factors. Thus, selection has less immediate effect when the genetic variance is low with respect to the error variance, and progress can be slow even when selection is consistently in the same direction.

Other gene models may be similarly viewed, including dominance and extending the models to include cumulative action of several loci. For example, if the three genotypes of locus A had means of: 1,003.5 for AA; 1,001.5 for AA'; and 993.5 for $A'A'$, and the frequencies were $\frac{1}{4}:\frac{1}{2}:\frac{1}{4}$, the total population mean would be 1,000, but the variance would be 1,014.75, including $\sigma_A^2=12.5$, $\sigma_D^2=2.25$, and the error variance around each genotype would be $\sigma_e^2=1{,}000$. The effect of selection can be seen in figure 3

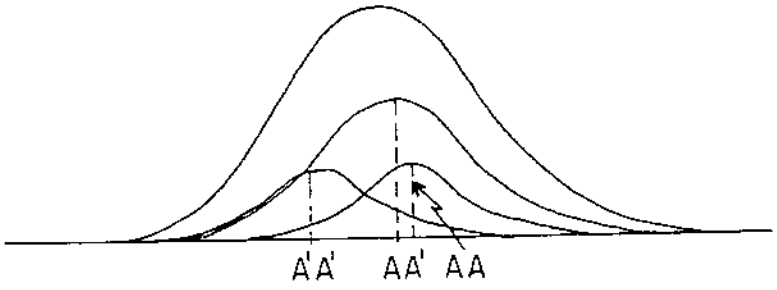


Figure 3.—Relative numbers of three genotypes from a population with a normal frequency distribution around genotypic means generated by partial dominance gene action and gene frequency 0.5.

to be less discriminating among the alleles than under pure additivity since the relative proportions of AA and AA' in the selected group are more nearly equal.

TWO-LOCUS MODELS

Expanding consideration to two loci, a simple additivity of alleles within loci and among equally effective loci would give average genotypic means of:

	AA	AA'	$A'A'$
BB	1,010	1,005	1,000
BB'	1,005	1,000	995
$B'B'$	1,000	995	990

If $q_A = q_B = 0.5$ and no linkage and random mating existed, figure 4 shows how the genotypes may be distributed. If equivalent levels of dominance existed in both independent loci, the following genotypic values would yield the same σ_A^2 and σ_D^2 at each locus as for the single-locus case with dominance as given above:

	AA	*AA'*	*A'A'*
BB	1,007	1,005	997
BB'	1,005	1,003	995
B'B'	997	995	987

Again, selection can be seen to have similar effect on both loci simultaneously, but for the same total selection intensity, there is less effect on each locus' gene frequency than for the single-locus case.

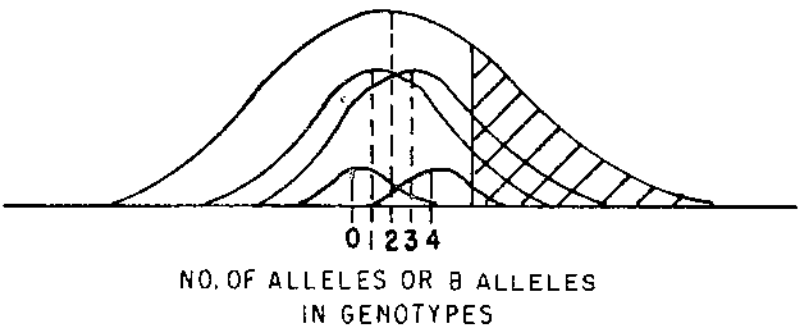


Figure 4.—Relative numbers of five genotypic means generated by two loci, each with additive gene action and gene frequency 0.5.

Various kinds of epistasis may now be included in these models of mean effects such as complementary dominance:

	AA	*AA'*	*A'A'*
BB	1,010	1,010	995
BB'	1,010	1,010	995
B'B'	995	995	985

In fact, any kind of mixed dominance conditions which change according to the allelic combinations of the other locus may be included:

	AA	*AA'*	*A'A'*
BB	1,010	1,000	995
BB'	1,010	1,005	995
B'B'	995	990	985

The effect of epistatic actions on changing gene frequency now becomes very complicated, since the effect on one locus will depend on the changing frequencies of the genotypic state at the alternate locus. Furthermore, linkage can cause the frequencies of the various genotypes to be nonindependent and variable and, therefore, would make selection prediction more complicated and intuitively more difficult to visualize. Even without epistasis, however, the addition of genetic loci can be seen to increase the genetic extremes and variances and hence can contribute a larger portion to the total variance even if individual gene actions have small mean effects.

NONSELECTIVE FACTORS*

Before continuing with more genetic models and how the effects of selection are translated into changes in gene frequencies and hence into population means, a few other complicating effects should be considered which further inhibit the direct response of alleles to selection. One factor is the nature of the breeding system with respect to inbreeding. For example, if selection is not precise and only a few individuals are chosen or if those chosen are related, then there is some chance that the wrong allele will increase in frequency or even be fixed by accident in the breeding population. Since inbreeding would tend to fix homozygotes in the absence of selection, then selection has to be relatively effective, or the genetic differences must be large relative to the error variance, to assure that the correct allele is going to be fixed. Even if selection is for the heterozygote, the pressure of inbreeding towards homozygosis can fix an allele by limiting free recombination of all alleles.

The problem of inbreeding and selection in regular mating systems (as distinct from completely random mating) may be analyzed in the form that Fisher (1965) derived for the long-run behavior of inbreeding systems. The analysis carries the probability distribution of zygotes, gametes, or mating types from one generation to the next which can be found for any regular mating system. The transition probabilities or the probabilities of genotypes or mating types to generate a new array of genotypes or mating types in the next generation are influenced by the mating system and selection effects or any other factors which may be included in the model. These effects can be traced in the eigenvalues and eigenvectors of the matrix. In Mather and Hayman's (1952) analysis of full-sib mating, for example, if selection was for heterozygotes such that homozygote survival was a fraction, $1-s$, of the heterozygotes, the transition probabilities for each of

*Graduate-level statistical training required for thorough understanding.

the mating types on the left for the next generation arrayed at the top would be:

Generation 0	Generation 1					
	$AA \times AA$:	$AA \times Aa$:	$Aa \times Aa$:	$Aa \times aa$:	$aa \times aa$:	$AA \times aa$
$AA \times AA$	$(1-s)$					
$AA \times Aa$	$\frac{(1-s)^2}{4}$	$\frac{(1-s)}{2}$	$\frac{1}{4}$			
$Aa \times Aa$	$\frac{(1-s)^2}{16}$	$\frac{(1-s)}{4}$	$\frac{1}{4}$	$\frac{(1-s)}{4}$	$\frac{(1-s)^2}{16}$	$\frac{(1-s)^2}{8}$
$Aa \times aa$			$\frac{1}{4}$	$\frac{(1-s)}{2}$	$\frac{(1-s)^2}{4}$	
$aa \times aa$					$(1-s)$	
$AA \times aa$			1			

An analysis of the major roots of such a matrix would then reveal the eventual stabilities among mating types and hence the persistence of heterozygosity. The eigenvectors would reveal the expected changes in frequencies of the mating types from generation to generation for any given starting frequencies. Alternatively, we may treat the progress of matings as a general stochastic process with a fixed Markov matrix and can determine for any time value the probabilities that some of the heterozygote (nonabsorbing) states may exist (Feller 1951). In a similar analysis, Hill (1969) traced the progress of changes in gene frequency using transition matrices and determined the probabilities of change by assuming a normal error distribution and given levels of selection intensity. In one-locus models, he confirmed Kojima's (1959a) finding that strong overdominance is required to maintain genetic variability at a locus under selection.

For single loci, an alternative mechanism for maintaining intermediate gene frequencies is variation in the environments which cause genotypic selection probabilities to change over generations. If the environmental variations are uniform over the population but affect selection over time within generations and are repeated each generation, then the net effect of geneotypic differences may be determined in a more complex multivariate form, but would nevertheless be translated into constant probabilities of selection. However, any variations over generations in the life cycle would induce variations in the transition probabilities and may affect the existence of genetic variations. Even such changes as earliness or duration of reproduction, as well as any changes in survival probabilities, would affect the relative fitness of genotypes. Then, even without dominance in any single environment, it is possible that intermediate gene frequency equilibria would be optimal. If

environmental variations exist among population subdivisions, then genetic polymorphisms may also exist. As discussed in chapter 9, populations may evolve stable equilibria under such conditions. In terms of single-locus selection in breeding populations, there would be little problem if the genotypes could be selected for specific ecological or economic environments in each generation, but if there is error in selecting genotypes and error in knowing the environments which will be faced, then more difficulties exist. If environments cannot be subdivided for more uniform treatment and single populations must be bred for mixed environments, then intermediate gene frequency optima may well exist. For multiple-locus traits, selection can have effects which cannot be predicted by simply extending the results of single-locus theory. As previously discussed, epistasis can generate several local optimum points and, with linkage, force populations into permanent disequilibria. Even without epistasis, certain unexpected stable equilibria can exist. For example, Wright (1935b) investigated multiple-locus selection for both additive and complete dominance gene actions and concluded that all loci would move toward fixation. Even selecting for an intermediate optimum would lead to a mixture of homozygous loci with the average gene frequency at an optimum mean frequency. However, Kojima (1959b) showed by using a quadratic fitness model that intermediate levels of dominance could lead to stable equilibria. Lewontin (1964) later extended these analyses to many loci and also found that several loci can be kept in intermediate frequencies with only partial dominance operating on a quadratic fitness model. Hence, many more complex polymorphisms may exist even under constant selection pressures when multiple loci are involved. The analysis of epistatic models in chapter 9 have direct implications for breeding theory with multiple loci.

SINGLE-LOCUS SELECTION WITH PHENOTYPIC VARIANCE AROUND GENOTYPIC MEANS

When error is involved in observing and selecting phenotypes, some additional complications to the immediate effectiveness of selection occur, depending on the distribution of the errors. Genetic effects can be modeled in much the same way as the effects of soil fertilizers or other site factors on tree yields. In a soil fertility experiment, variations in the yield (Y) of the kth tree (Y_k) might be ascribed to, say, potassium X_1 or nitrogen X_2, and the interaction X_{12}. In a linear or additive effects model, yield would depend on the summation of all effects which operate on the tree, including an error term for uncontrolled deviations, e_k:

$$Y_k = \mu + X_1 + X_2 + X_{12} + e_k.$$

In such a model, the effects determine the direction of the tree's performance from the mean. Similarly, for the alleles at a single locus, the variations in yield can be ascribed to effect of each allele

(α_1 and α_2), and the interaction δ_{12}:

$$Y_k=\mu+\alpha_1+\alpha_2+\delta_{12}+e_k.$$

Since effective selection requires the existence of variations in yield and since yield depends on site or genetic factors, it is the variations in site or genetic effects which determine the potential gains. Variation in yield due to variations in fertilizer effects can be measured as a variance and is designated as a σ_x^2 even though it is a variance in Y due to X. Thus,

$$\sigma_x^2=\Sigma\ f_i\ X_i^2\ -\mu^2,$$

as given in chapter 1, where the X is actually the effect of X on the Y measure of yield. Similarly, the variance due to genetic effects is designated by σ_a^2 and σ_δ^2 and depends on the frequency with which those allelic combinations occur as well as on the size of the effects.

Using the definitions of gene effects given in chapter 1 where the effect of the genotypes $AA:AA':A'A'$ was measured in terms of u (the difference between AA and $A'A'$) and au (the deviation of AA' from the midpoint between AA and $A'A'$), the additive genetic variance was given as:

$$\sigma_A^2=2q(1-q)u^2\ [1+(1-2q)a]^2.$$

Using the average effect of the alleles as the a's given above, the average effect of an A allele (α_A) is:

$$\alpha_A=u(1-q)\ [1+a(1-2q)],$$

and the average effect of an A' allele is:

$$\alpha_{A'}=\ -uq\ [1+a(1-2q)].$$

Then, since the frequency of A is q, and the frequency of A' is $1-q$, the variance of the average effects is:

$$q\alpha^2+(1-q)\alpha_a^2=q(1-q)u^2[1+a(1-2q)]^2.$$

This is exactly ½ of the additive genetic variance, σ_A^2.

From this simple linear model of gene effects and environmental variations, the genotypic mean is defined in terms of the α and δ effects. Then, the probability of the genotype being selected is defined in terms of its having those alleles and the phenotype such that it is included in the selected population. From the array of probabilities of each genotype belonging to the selected population, the expected distribution of selected genotypes is derived in terms of the genetic variances. From this same array, random mating among those selected is then derived and the mean gain of the progeny is shown to be well approximated by the familiar $s \times$ heritability formula. The assumptions involved in the derivation are noteworthy.

GRIFFING'S EXPECTED-GAIN FORMULA

Using the above model of gene effects, assuming that the allelic frequencies are p_i^o in the initial generation and using

$$d_{ij}^o = \alpha_i^o + \alpha_j^o + \delta_{ij}^o,$$

the population mean at the original time (o) is:

$$\mu^o = \sum_{ij} p_i^o p_j^o d_{ij}^o,$$

and $\alpha_i^o = \Sigma_j p_j^o d_{ij}^o$.

In the standard definitions of linear effects, the additive genetic variance is:

$$\sigma_A^2 = 2\Sigma p_i^o \alpha_i^2,$$

and the dominance genetic variance is:

$$\sigma_D^2 = \sum_{i \neq j} p_i^o p_j^o \delta_{ij}^2.$$

As previously described, α_i is the average effect of allele i, α_j is the average effect of allele j, and δ_{ij} is the effect of the dominance deviation due to the interaction of the i and the j alleles. The effect of selection can be described in terms of the probability that a particular $i \times j$ genotype will be included in that part of the population which is selected to be the parents of the next generation. Once the probability is determined for each genotype, the probability distribution can be determined, and from any such distribution the mean can be computed. This is the analytical strategy we follow.

If the genotypes are not directly observable, then selection would phenotypically resemble the truncation type shown in figure 1, and the effect on the three genotypes generated by one-locus variations would resemble the type shown in figure 2. The probability of selection would be proportional to the value of d_{ij}, increasing for high values and diminishing for low, and would be inversely related to the total phenotypic variance σ^2 which is the sum of σ_e^2 and genetic variations and hence includes all genetic and environmental sources of variation. The probability of selection is approximately:

$$Pr \text{ (select } i,j) = v\left(1 + \frac{d_{ij}}{\sigma^2} \times s\right)$$

where v is the proportion selected, and s is the difference between the mean of the original population and the mean of the selected population. Since v is a constant for the population, the relative selective value of the i and j genotypes is:

$$1 + \frac{d_{ij}}{\sigma^2} \times s.$$

Therefore, the expected relative frequency with which the particular genotype occurs in the selected population is:

$$p_i^o p_j^o \left(1+\frac{d_{ij}}{\sigma^2}s\right).$$

Given this relative frequency of genotypes in the truncated portion, the weighted mean value of the selected, truncated population is:

$$\begin{aligned}&\sum_{ij} p_i^o p_j^o \ \left(1+\frac{d_{ij}}{\sigma^2}s\right)\ d_{ij}\\ &=\sum_{ij}\ p_i^o p_j^o d_{ij}+\frac{s}{\sigma^2}\sum_{ij} p_i^o p_j^o d_{ij}^2\\ &=0+\frac{s}{\sigma^2}\sigma_G^2\end{aligned}$$

where σ_G^2 is the total genetic variance.

Hence, the expected mean genetic value of the truncated population before any mating or recombination of these potential parents is $s \times$ broad-sense heritability, because σ_G^2 is the total genetic variance and σ^2 is the total phenotypic variance, including all genetic and nongenetic sources of variation.

If mating is now made among the selected parents, randomly with respect to genotype, these parents will leave progeny in frequencies determined by their own altered genotypic frequencies. Assortative mating within the selected group invalidates this assumption. The new expected gene frequency p_i^1 for allele i is determined by the probability that the different carriers of the i allele are included in the selected parental group and would be:

$$\begin{aligned}p_i^1&=\Sigma_j p_i^o p_j^o \ \left(1+\frac{d_{ij}}{\sigma^2}s\right)\\ &=p_i^o \ \left(1+\frac{s}{\sigma^2}\alpha_i\right)\\ &=p_i^o+\frac{s}{\sigma^2}p_i^o\alpha_i.\end{aligned}$$

We now have a difference equation relating gene frequencies for two generations. If mating is at random with these new gene frequencies and the number of selected parents is reasonably high, the progeny generation will have genotypes i and j according to the Hardy-Weinberg frequencies:

$$\sum_{ij} p_i^1 p_j^1.$$

For just two alleles, this is:

$$(p_i^1)^2 : 2p_i^1\ (1-p_i^1) : (1-p_i^1)^2$$

and hence the mean of the progeny population is:

$$(p_i^1)^2 d_{ii} + 2p_i^1 (1-p_i^1) d_{ij} + (1-p_i^1)^2 d_{ij}.$$

In terms of the allelic frequencies in the original population, the frequencies for this population can be generated by the combinations:

$$\left(p_i^o + \frac{S}{\sigma^2} p_i^o \alpha_i\right) \left(p_j^o + \frac{S}{\sigma^2} p_j^o \alpha_j\right)$$

and can be grouped as:

$$(p_i^o)^2 + \left(\frac{S}{\sigma^2} p_i^o \alpha_i\right)^2 + 2\frac{S}{\sigma^2} (p_i^o)^2 \alpha_i,$$

for the ii genotypes,

$$2\left[p_i^o p_j^o + \frac{S}{\sigma^2} p_i^o p_j^o (\alpha_i + \alpha_j) + \left(\frac{S}{\sigma^2}\right)^2 p_i^o p_j^o \alpha_i \alpha_j\right],$$

for the ij genotypes, and

$$(p_j^o)^2 + \left(\frac{S}{\sigma^2} p_j^o \alpha_j\right)^2 + 2\frac{S}{\sigma^2}\left(p_j^o\right)^2 \alpha_j,$$

for the jj genotypes. Multiplying these frequencies by their d_{ij} values provides a progeny population mean:

$$\mu(\text{progeny}) = \sum_{i,j} p_i^o p_j^o d_{ij} + \frac{S}{\sigma^2} \sum_{i,j} p_i^o p_j^o (\alpha_i \alpha_j) d_{ij}$$
$$+ \left(\frac{S}{\sigma^2}\right)^2 2\sum_{i,j} p_i^o p_j^o (\alpha_i \alpha_j) d_{ij}$$

Substituting $\alpha_i + \alpha_j + \delta_{ij}$ for d_{ij},

summing as indicated, and using

$$2\sum p_i \alpha_i^2 = \sigma_A^2,$$

we derive

$$\mu(\text{progeny}) = o + \frac{S}{\sigma^2} \sigma_A^2 + \left(\frac{S}{\sigma^2}\right)^2 \sum_{i,j} p_i^o p_j^o \left(\frac{\alpha_i \alpha_j}{\sigma^2}\right) d_{ij}.$$

Then, if the last term's products are small, a good approximation to the progeny mean is:

$$\mu(\text{progeny}) = \frac{S}{\sigma^2} \sigma_A^2 = s \times (\text{narrow-sense heritability}).$$

HERITABILITY

Griffing's derivation, as outlined above, gives flesh to the relationship between the genetic and phenotypic variances and the

progress in the population mean from selection. The change in the genotypic frequency array is the direct effect of selection which, in turn, affects gene frequencies of the parents and the genotypic frequencies of the progeny population, and, consequently, the new population mean. Since all of these changes can be written in terms of the gene-model effects (α_i) and gene frequencies, the change in population mean is a product function of the gene effects and frequencies. In addition to the genetic variance, there is a selection differential multiplier and a total variance divisor. The ratio of the additive genetic variance to the total variance

$$\frac{\sigma_A^2}{\sigma_A^2+\sigma_e^2}$$

is called the narrow-sense heritability and is a useful statistic to describe relative amounts of additive genetic and nongenetic sources of variance as well as to predict gain from simple selection procedures.

The selection model thus far considered is a simple method of recurrent selection in which individuals are selected without regard to the existence of information on relatives or coancestry, and are simply random mated. We develop more complicated models in chapter 3. The genetic model is for one locus; however, if the trait under selection is affected by several independent loci without epistasis and without linkage, each of small effect, the selection effects may be summed over loci and the same formula would predict one-generation gains for the accumulated action of all loci. As long as the genes operate in approximately the same manner and the individual gene frequencies do not change drastically for several generations, the predictions will hold for each new generation cumulatively. With many loci of small effect, it is reasonable to expect that the total variance may be quite large due to the accumulated genetic variances at each locus. If the selective action at each locus is such that only small changes in frequencies occur on each of many loci, however, the net gain in effect can be large. Hence, continued gain can be obtained in sequential breeding generations as long as some loci continue to contribute useful genetic variance. In this sense, substantial gains can be accumulated and the genetic sources of improvement hence can represent something of a renewable resource for gain if managed in such a way as to preserve variation while still accumulating gain.

SELECTION DIFFERENTIAL

In populations with traits which have a normal distribution, the mean difference between the original and selected parents, s,

can be computed in terms of the proportion selected very easily. The mean of the new parental selection is:

$$\int_t^\infty xf(x)\,dx \div \int_t^\infty f(x)\,dx$$

where x is the phenotypic scale, t is the truncation point, and $f(x)$ is the probability density function, and in the normal distribution,

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

for a population scaled to a mean of zero and variance of one. Integrating the numerator by using the substitution $u = x^2/2$ gives us:

$$\frac{e^{-t^2/2}}{\sqrt{2\pi}}$$

which is the height of the ordinate of the normal curve at the point of truncation. Distribution functions other than the normal can be directly evaluated or approximated on computers to give the relationship between the truncation point, proportion selected, and the selection differential. The denominator is merely the proportion selected and therefore the mean, taken as a deviation from the original mean, is z/p, where z is height of the ordinate of the truncation point, and p is the proportion selected, for a standardized phenotypic variance. If the phenotypic variance is not standardized, then $s = (z/p)\sigma$, where the phenotypic variance $= \sigma^2$. It is sometimes useful to distinguish between the standardized selection differential, which is often called the selection intensity $i = z/p$, and the nonstandardized selection differential $s = i\sigma$, which is the difference as measured in the scale of the original units of measurement. This selection differential, s, is thus the difference in means between what we started with and what we have chosen as parents of the new generation and represents the amount of change we "reach" to achieve.

GAIN

Because the phenotypic variance σ^2 includes nonadditive genetic and error sources of variation, however, only a fraction of this "reach" is actually achieved. The fractional achievement expected under the simple breeding scheme given can now be seen to be the proportion of the total phenotypic variance which is due to additive genetic variances $\sigma_A^2 \div \sigma^2$, which is otherwise known as the narrow-sense heritability. This stands in contrast to the gain achieved in actual parental genotypic mean values for which the

fractional achievement rate is $\sigma_G^2 \div \sigma^2$ and includes all of the non-additive genetic effects in the numerator.

These concepts of gain have also served as the basis for much plant and animal breeding theory for large-population sizes. For traits affected by large reservoirs of additive types of genetic variance, they have served very well. For many species of plants and animals, the various modifications of the theories have reliably predicted genetic gains (Sprague 1966; Allard 1960). However, the models are extremely naive in their assumptions of steady gene frequencies and genetic and phenotypic variances and in their exclusion of obviously important genetic effects. For example, genes do not all act in small increments. Some must change frequency as selection progresses, they do occur in linkage groups, and they undoubtedly have some forms of epistatic interactions. In addition, dominance effects can lead to inbreeding depression (Kojima 1961) and asymmetrical responses to selection (Curnow and Baker 1968). While some experiments may tend to confirm the general adequacy of Griffing's (1960) theoretical estimates, the asymmetry of response to selection and lack of continued gain in other experiments could be due to any of several factors. If the genetic variance and gain from selection are due to few alleles of large effects, the above approximations can be quite inaccurate as these major loci become fixed (Latter 1965).

It is also clear that any one selection trial samples different sets of individuals and may therefore start with a distribution of genotypes, other than what may be expected on the average. In small populations, the genotypic distribution and its concomitant mean and variance measures may therefore vary from trial to trial. Also, since genotypes are observed with some error, the actual selection differential can vary widely for any given genotypic distribution. The above measures are therefore good only for large population sizes but serve as predictors of average results.

POPULATION SIZE

Among the more serious difficulties in accurately predicting gains from selection are the effects of population-size restrictions on changing gene frequencies. Whenever selected populations are restricted in size, there is some chance of losing an allele otherwise favored by natural or artificial selection, even with simple additive gene effects. When consideration is extended to several loci, the chance loss of potentially valuable alleles can severely restrict the size of the potential gain. Since it is generally assumed that no single tree possesses all of the desirable alleles for all traits simultaneously, ultimate progress requires that several genotypes be used in the breeding population to assure the presence of at least most of the useful alleles in the breed. Thus, if we had six equally effective and independent loci with simple additive effects, the following array of 10 genotypes may exist as randomly drawn

with gene frequencies 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6:

Tree	Locus						Total plus alleles
	1	2	3	4	5	6	
A	$\frac{-}{+}$	$\frac{-}{+}$	$\frac{+}{-}$	$\frac{+}{+}$	$\frac{-}{+}$	$\frac{-}{-}$	6
B	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{-}$	$\frac{-}{-}$	$\frac{+}{-}$	$\frac{-}{+}$	3
C	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{-}$	1
D	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{+}$	$\frac{-}{+}$	$\frac{+}{+}$	4
E	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{-}$	$\frac{+}{+}$	$\frac{+}{-}$	4
F	$\frac{-}{+}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{-}$	$\frac{-}{-}$	$\frac{-}{+}$	3
G	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{-}{+}$	$\frac{+}{-}$	$\frac{+}{+}$	4
H	$\frac{-}{-}$	$\frac{+}{+}$	$\frac{-}{-}$	$\frac{+}{+}$	$\frac{+}{+}$	$\frac{-}{-}$	6
I	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{+}$	$\frac{-}{-}$	$\frac{-}{-}$	$\frac{+}{+}$	4
J	$\frac{-}{-}$	$\frac{-}{+}$	$\frac{+}{+}$	$\frac{-}{-}$	$\frac{+}{+}$	$\frac{+}{+}$	7
Total plus alleles	2	4	6	8	10	12	42

In this sample of trees, the best genotype is *J* and its selection would assure a good chance to eventually get an all-plus breed but would not, by itself, give us all of the best alleles. In addition, with some error in observing true genotypic values, there could be only a slightly higher probability that tree *J* is chosen, and not tree *H*, *A*, or *G*. If the random error has a large variance relative to the average genotypic differences, the difference in probability of selection between trees *J* and *C* may be quite small and hence *C* may be picked over *J* amost as often as *J* over *C*. In that case, or in case some mixture of trees is chosen, some good alleles will be lost in spite of the gain in frequency of good alleles which might generally be expected. Therefore, limiting the breeding population can limit progress even without dominance or inbreeding depression. Many forest tree breeding operations appear to have population sizes that are too small to permit continued breeding progress for more than a few generations. With few parents, the subsequent generations will be generated from relatives with increasingly similar ancestral lineages. The number of independent genotypes among the parents must then decrease. Thus, the probability of accidental gene loss would increase even if the physical number of parents remained the same. Relationships among the

parents decreased the effective population size (N_e) as further detailed in chapter 3. The population size useful for computing probabilities of accidental loss of alleles is N_e, which is usually smaller than the number of parents crossed. On the other hand, it is intuitively obvious that greatest progress is expected by the most intensive selection and the consequently greatest reduction of breeding population size to only the very best parents. The dilemma, therefore, is how to maintain both a large population size and a large selection differential. The problem is most easily stated in terms of the special effects of stochastic variation in small populations as discussed in chapter 9. For a large number of independent genetic loci affecting a trait under selection with simple types of gene actions, we could first determine the probabilities of loss of favorable alleles and then consider more complicated models incorporating migration, nonadditive gene action, etc. We return to the applied breeding implications of this dilemma in chapter 3.

DIFFUSION MODELS FOR SELECTION

First, considering a simple diffusion process, the effects of selection are assumed to be such that a constant pressure for a directed change in gene frequency exists. We can easily conceive of gene-action models where this is not so, such as if dominance exists, or even as we developed for Griffing's approximations, the change in gene frequency:

$$p_i^1 - p_i^0 = p_i^0 \frac{S}{\sigma^2} \alpha_i,$$

is a function of α_i and p_i^0. Nevertheless, for small changes in gene frequency and effects, and without dominance, a selection pressure on an average change in gene frequency of x may be a reasonably good approximation. In Kimura's (1964) notation, an additive gene-action model entailing the following probabilities of selection would produce an average change in gene frequency of $\zeta p(1-p)$, where ζ is the difference in the probability or expected frequency of selection against $A'A'$ and for AA. The effect on the zygotes is expected to be:

$$\left(1 - \frac{\zeta}{2}\right) A'A' : (1) AA' : \left(1 + \frac{\zeta}{2}\right) AA.$$

That is, from the expected change in gene frequency on a continuous time scale:

$$\frac{dp}{dt} = p(1-p) \quad (r_A - r_{A'}),$$

where r is the relative fitnesses of the alleles, the mean change in gene frequency, p, is $\zeta p(1-p)$. If the variance in gene frequency is affected only by binomial sampling error $\frac{p(1-p)}{2N_e}$, then the

probability of fixing the favored allele A is:

$$\text{UPF}=\frac{1-e^{-4N_e\zeta p_o}}{1-e^{-4N_e\zeta}},$$

where p_o is the initial gene frequency, and where N_e is the effective population size. As previously suggested, several simplifications on the formulas can elucidate some relationships among the variables, p_o, N_e, and ζ. For example, if $N_e\zeta=0$, then the limiting value of UPF is p_o, which is intuitively satisfactory for the case when selection is not practiced and gene frequency is allowed to drift at random. In the case of selection, however, with $N_e\zeta>0$, then UPF becomes a function of p_o and the product $2N_e\zeta p_o(1-p_o)$, as well as other terms of smaller size.

The distribution function for the whole range of gene frequencies under additive gene action reduces to:

$$\frac{ke^{-4N_e\zeta p}}{p(1-p)}$$

which has peaks at high and low gene frequencies and can be skewed to either end by the effects of selection. This results in a J-shaped curve increasing the frequency of the favored allele and its probabilities of fixation over the alternate allele.

Using a model with overdominance as: $(1)A'A':(1+\zeta)A'A:(1)AA$, the frequency distribution function can be derived to be:

$$\frac{ke^{4N_e\zeta p(1-p)}}{p(1-p)}\text{ (Li 1955)}.$$

In this form of the gene frequencies, the intermediate frequencies enjoy some greater weight; but the extremes still occur to provide a profile as:

In addition, the joint effects of selection and the various effects of population size, migration, and mutation rates can be jointly determined for some simple genetic models. As previously outlined, migrations or mutations may introduce genes into a population at a rate which may either reinforce or act against the effects of selection. In small populations, all effects are further modified by the tendency of genes to become randomly fixed simply by

sampling error. Variations in selection coefficients can also produce a tendency for fixation as analyzed by Levins (1968). If stable gene frequency distributions exist for genetic models with constant selection coefficients, then variances in the coefficients tend to destabilize them. If the variations in the selection coefficients exist in correlated series, however, the variance effects would be ameliorated and populations may behave in cyclic patterns. On the other hand, with moderate directional selection and additivity, variance in selection can induce more stability in intermediate gene frequencies than if selection was consistently in the same direction.

The general difficulty that restrictions on population size may impose on selection advance is the random fixation of alleles which may not be the favorable ones. Thus, even without considering epistasis or inbreeding depression, the loss of good alleles can be a serious problem, especially for long-term prospects of accumulating maximum improvement, on the basis of cumulatively improved breeding population. Indeed, in the long run the breeder will always face the possibility that by restricting population size, he will not have the kinds of genetic variations available for further improvement that he would like. When economic, ecological, or environmental changes occur, he would either have to develop at least some new unselected genotypes with an otherwise less favorable collection of alleles in order to introduce new variants for recombination and selection or else he would have to be content with his limited gains. Thus, the immediate breeding problem is how to compromise his selection program between the maximization of immediate gain by the highest selection intensity and lowest N_e as against the maximization of long-run gains by some partial relaxation of selection in the breeding population. The long-term problem for the breeder is to develop population mixtures which will permit him to continually develop variations without excessively sacrificing general fitness or economic value of the breeding population. The additional problem of developing populations for short- and long-run objectives when the physical and economic environments are changing in uncertain ways is a further problem we postpone to chapter 4. The problem considered here is the effects of selection on populations assuming some known direction. It is, therefore, the genetic problem of response of a population of organisms and not the economic one of the value of the response.

The application of diffusion-process approximations to the effects of selection, as proposed by Kimura, was significantly advanced by Robertson (1960), who considered the ultimate probability of fixation to be a good criterion for judging the long-term effects of selection. Only the simplest genetic models of additive gene action, no migration or mutation, and independent loci were initially considered, though subsequent research has amplified the effects of those forces. The distributions are derived for either a large sample of genetic loci which together affect a trait in a

single population or for a large sample of populations with a single locus displaying the expected distribution of allelic frequencies among the population. It seems clear that other measures of goodness might also serve particular needs including measures such as skewness, degree of heterozygosity, duration of allelic variations, etc., which may give additional information on rates of selection advance. Nevertheless, the probability of fixation is a useful measure which contains much of what breeders are interested in. We shall consider the probability of fixation, $u(q)$, as the expected proportion of equivalent loci which would be fixed in a single population or as the proportion of sampling populations which would have the favorable allele fixed.

As previously noted, without selection, $N_e\xi=0$, and the solution of the $u(q)$ equation is a function only of q_0, the initial gene frequency. Any low initial gene frequency thus has a proportionately low probability of fixation, a 0.5 initial frequency may go either way, and a high q_0 will have a high probability of fixation by random events. With positive selection for the allele, q_0, the $u(q)$ function increases approximately as a function of $q(1-q)N_e\xi$. Therefore, $u(q)$ is dependent on the quadratic function $q(1-q)$ for any given $N_e\xi$ level, and the change is most rapid in the intermediate gene-frequency ranges. The relationship between $u(q)$ and $N_e\xi$ is charted in figure 5 for seven levels of q_0. If dominance exists, somewhat differently shaped curves result as the change $u(q)-q$ function is approximately $(2\ 3)N_e\xi(1-q^2)$ when selection is for the recessive allele. From these figures it is clear that high initial frequencies of favorable alleles present little problem in maintaining them in the selected population and of ultimately eliminating the alternate alleles. Those alleles which start at low frequencies are difficult to advance and are easily lost, especially

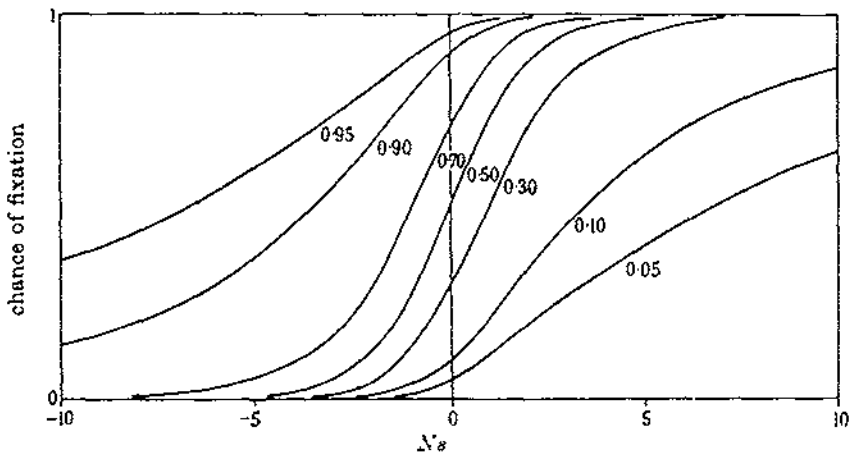


Figure 5.—The chance of fixation of a gene acting additively. The curves are drawn for different initial gene frequencies. (Robertson 1960)

at low $N_e\zeta$. Both N_e and ζ at high levels are therefore necessary to ensure against the loss of useful alleles. The speed with which the frequency of favorable alleles is fixed is also a function of N_e and ζ. The time required for ½ of the total gain to be achieved is approximately $N_e\zeta q(1-q)$, and therefore maintaining large N_e and ζ will also assure rapid progress.

If favored alleles exist at low initial frequencies, however, it is clear that periods of inbreeding or any reduction of the effective population size in the early generations can strongly reduce long-term gain potentials by eliminating alleles before selection has increased their frequency. Thus, in figure 6, the restriction of population size is shown to be always debilatory. However, if

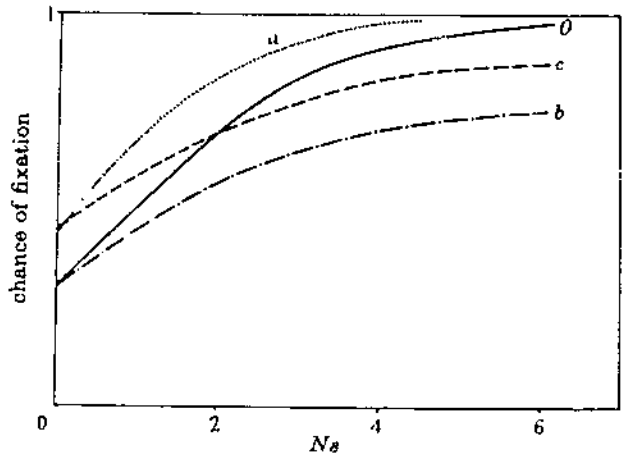


Figure 6.—The effect of various treatments on the curve of chance of fixation against $N\zeta$ for a gene with initial frequency of 0.3. The treatments are three generations of (a) selection with $\zeta = 0.4$ in a large population, (b) restriction of effective population size to 5, and (c) selection with $\zeta = 0.4$ and effective population size 5.0=original. (Robertson 1960)

alleles are suspected of initially being at low frequencies and a high $N_e\zeta$ can be initially maintained, then the initial frequencies can be advanced and less restrictive breeding procedures would be allowable in future generations. Thus, if initial selection can advance low gene frequencies into the intermediate range, considerable safety against accidental loss of alleles is assured. Nevertheless, as seen in figure 7, if the initial selection requires a loss of N_e, those early restrictions in N_e are always detrimental, especially for alleles at lower initial frequencies. Thus, previously unselected populations require large initial efforts to attain large $N_e\zeta$ more so than previously selected or partially improved breeds, although all populations respond better to selection with high N_e. In tree breeding, high N_e is likely to be required to compensate for low and possibly variable ζ.

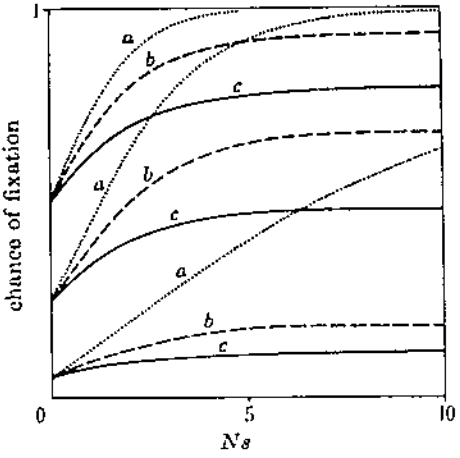


Figure 7.—The effects of "bottlenecks" in population size on the curve of chance of fixation against calculated for initial gene frequencies of 0.1, 0.3, and 0.5: (a) initial population, (b) restriction to a single mating for one generation only, and (c) restriction to a single mating for three consecutive generations. (Robertson 1960)

In terms of phenotypic gains, the ζ used here is equivalent to Griffing's $\frac{S\alpha_i}{\sigma^2}$ and we can compare the relative effect of having a few alleles of large α_i effect versus more alleles with small α_i effects if the total effect of the genes is the same for both systems. Robertson (fig. 8) finds that for initial frequencies of 0.5, the

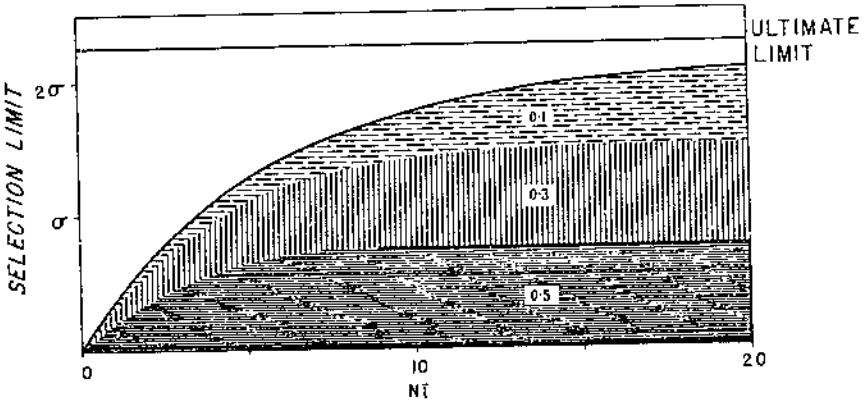


Figure 8.—The expected limits to artificial selection in a population in which all genes have initial frequency 0.5 and in which the possible advance is contributed equally by genes with α/σ = 0.1, 0.2, and 0.5, respectively. (Robertson 1960)

larger effect (larger selection coefficient) genes, although fewer in number, contribute most heavily at lower values of $N_e\frac{z}{\mu}$.

Only at high $N_e \frac{z}{p}$ values do the less heavily selected alleles, but at more numerous genetic loci, contribute as much. Similarly, comparing low versus intermediate initial gene frequency loci of equal effect, the lower frequency loci contribute significantly only at the higher $N_e \frac{z}{p}$ levels since the lower frequency alleles are more easily lost. Thus, N_e affects the relative importance of factors which otherwise would be considered of equivalent merit.

These basic concepts of course involve many simplifications, but they have provided a basis for considering the effects that such factors as linkage, dominance, changing genetic variances, and genetic backgrounds can have on the general progress from selection predicted in these models. In addition, many approximations involved in the derivations are not justified, as noted by Robertson (1960). Hence, alternate derivations and independent tests of the results have been used to determine the adequacy of the models and to propose new, more comprehensive and more exact models. While Ewens (1963) found little error in the diffusion-equation approximations for an additive gene-action model at very low N_e, other effects may distort the expected results. More exact analysis of selection effects, such as derived by Hill (1969), is instructive to describe for its explicit statement of assumptions.

OTHER PROBABILITY MODELS FOR SELECTION *

The first major objective of Hill's (1969) analysis is to derive the probability that each of the $A'A'$, AA', and AA genotypes is represented by exactly n_1, n_2, and n_3 individuals ($n_1+n_2+n_3=N$) in a new population when they were originally represented by m_1, m_2, and m_3 individuals ($m_1+m_2+m_3=M$) in the parental population. We do this for all possible combinations of n_1, n_2, and n_3. From these probabilities, we can then compute the changing genotypic and allelic frequencies for the three types, or more generally for any of g different genotypes in any multiple allelic series. We do this for one generation at a time; assuming that the error variances do not change, and further assuming that these probabilities are independent of n_i and m_i levels, we can use matrix methods to project future population behavior.

Since the truncation point is determined by selecting the top N out of the M potential parents, we cannot be sure where the truncation point will come in the rankings of each genotype, nor where in the entire phenotypic range it may fall. One way to compute those probabilities is to determine the exact probability that the point of truncation will produce n_1 of the first genotype and n_2 of

*Graduate-level statistical training required for thorough understanding.

the second, etc. Since each genotype would have a slightly different distribution (fig. 2), the relative probabilities of representation in the selected population change according to where the truncation point falls and the various distributional differences. The probabilities of obtaining n_1, n_2, and n_3, given m_1, m_2, and m_3, can be obtained by looking at the mutually exclusive events; that the lowest ranked selection is of genotype 1, and n_1–1 are higher, and n_2 and n_3 of the other types also rank higher; that the lowest selected is of genotype 2, and n_2–1 are higher, and n_1 and n_3 also rank higher; and that the lowest selected is of genotype 3, and n_3–1 are higher, and n_1 and n_2 also rank higher. These probabilities can be derived from order statistics and for the first kind of event are:

Pr(n_1 of genotype 1 are selected, and the lowest ranked selection is of genotype 1 given m_1 to choose from)

$$= \frac{m_1!}{(n_1-1)!\,(m_1-n_1)!}\left[F_1(x)\right]^{m_1-n_1}\left[1-F_1(x)\right]^{n_1-1} f_1(x)\,dx$$

where $f_1(x)$ is the probability density function for genotype 1 and $F_1(x)$ is the integrated form of $f_1(x)$ or the cumulative distribution function for genotype 1.

Pr(n_2 of genotype 2 are greater than truncation individual m_2)

$$= \frac{m_2!}{n_2!\,(m_2-n_2)!}\left[F_2(x)\right]^{m_2-n_2}\left[1-F_2(x)\right]^{n_2}$$

Pr(n_3 of genotype 3 are greater than truncation individual m_3)

$$= \frac{m_3!}{n_3!\,(m_3-n_3)!}\left[F_3(x)\right]^{m_3-n_3}\left[1-F_3(x)\right]^{n_3}$$

Since these are order statistics and are all independent, the probability for the first kind of event is the product of these probabilities.

The second kind of event puts genotype 2 in the position of having its lowest ranked selected tree also representing the lowest ranked selected tree for all genotypes, and this joint probability requires only switching notation between 1 and 2 in the above equations.

The third kind of event is similarly treated, and if a multiple allelic series exists, any other genotypes can be similarly handled.

Since each of the kinds of events are mutually exclusive, they may be summed for all types of events, each event's probability being the product of g terms. For three genotypes (g=3), the probability of obtaining n_1, n_2, and n_3 from a given set of m_1, m_2, and m_3 trees when the truncation point is at x is:

$$\frac{m_1!}{(n_1-1)!(m_1-n_1)!}\left[F_1(x)\right]^{m_1-n_1}\left[1-F_1(x)\right]^{n_1-1} f_1(x)dx$$

$$\times \frac{m_2!}{n_2!(m_2-n_2)!}\left[F_2(x)\right]^{m_2-n_2}\left[1-F_2(x)\right]^{n_2}$$

$$\times \frac{m_3!}{n_3!(m_3-n_3)!}\left[F_3(x)\right]^{m_3-n_3}\left[1-F_3(x)\right]^{n_3}$$

$$+ \frac{m_2!}{(n_2-1)!(m_2-n_2)!}\left[F_2(x)\right]^{m_2-n_2}\left[1-F_2(x)\right]^{n_2-1} f_2(x)dx$$

$$\times \frac{m_1!}{n_1!(m_1-n_1)!}\left[F_1(x)\right]^{m_1-n_1}\left[1-F_1(x)\right]^{n_1}$$

$$\times \frac{m_3!}{n_3!(m_3-n_3)!}\left[F_3(x)\right]^{m_3-n_3}\left[1-F_3(x)\right]^{n_3}$$

$$+ \frac{m_3!}{(n_3-1)!(m_3-n_3)!}\left[F_3(x)\right]^{m_3-n_3}\left[1-F_3(x)\right]^{n_3-1} f_3(x)dx$$

$$\times \frac{m_1!}{n_1!(m_1-n_1)!}\left[F_1(x)\right]^{m_1-n_1}\left[1-F_1(x)\right]^{n_1}$$

$$\times \frac{m_2!}{n_2!(m_2-n_2)!}\left[F_2(x)\right]^{m_2-n_2}\left[1-F_2(x)\right]^{n_2}$$

Since the truncation point x may actually occur over the whole range of x, we can sum or integrate over all x values to determine the transition probability of going from m_1, m_2, and m_3, to n_1, n_2, and n_3, and by the gathering of appropriate terms we can in general write this $Pr(n_1, n_2, \ldots n_g, m_1, m_2, \ldots m_g)$ as:

$$\int_x \prod_{i=1}^{g}\binom{m_i}{n_i}\left[F_i(x)\right]^{m_i-n_i}\left[1-F_i(x)\right]^{n_i}\cdot\left[1-F_j(x)\right]^{-\sum_{j=1}^{g} n_j} f_i(x)\,dx$$

The equations are exact for the transition from m to n, and the only remaining problem is how to get the probabilities of transition from n in a population of trees to a new n in the new popula-

tion of trees. That is, starting with a parental array $n^{[0]}$, they mate to produce a new and presumably larger progeny population of $m^{[0]}$ from which we select a new parental population of $n^{[1]}$. The above transition probabilities give us the probabilities of $n^{[1]}$, given the $m^{[0]}$ or $Pr(n^{[1]} \mid m^{[0]})$. We wish to determine $Pr(n^{[1]} \mid n^{[0]})$ and this can be determined by $Pr(n^{[1]} \mid m^{[0]}) \times Pr(m^{[0]} \mid n^{[0]})$. The last probability distributions to consider are therefore the $Pr(m^{[0]} \mid n^{[0]})$, which are determined by the mating patterns used among those selected. Assuming random mating of the $n^{[0]}$ trees simplifies the determination of $Pr(m^{[0]} \mid n^{[0]})$, but any mating pattern may be pursued to give the probabilities. The product of $Pr(m^{[0]} \mid n^{[0]})$ with $Pr(n^{[1]} \mid m^{[0]})$ can then be formed to give $Pr(n^{[1]} \mid n^{[0]})$. The simplification that the random-mating assumption affords is that $m^{[0]}$ is completely determined by the gene frequency (under the Hardy-Weinberg law). If the gene frequency in the $n^{[0]}$ population is p_o, we can determine $m^{[0]}$ as:

$$p_o^2 : 2p_o(1-p_o)^2 : (1-p_o)^2$$

or in terms of numbers of alleles:

$$p_o = \frac{2n_1^{[0]} + n_2^{[0]}}{N^{[0]}} = \frac{i}{N^{[0]}}.$$

Given these frequencies, the probabilities for the generation of $m^{[0]}$ are multinomially distributed:

$$Pr(m^{[0]} \mid n^{[0]}) = \binom{M}{m_1 m_2 m_3} \left(p_o\right)^{2m_1} \left[2p_o(1-p_o)\right]^{m_2} \left[1-p_o\right]^{2m_3},$$

$$\text{or} = \binom{M}{m_1 m_2 m_3} \left(\frac{i}{2N}\right)^{2m_1} \left[\frac{i}{N}\left(1-\frac{i}{2N}\right)\right]^{m_2} \left[1-\frac{i}{2N}\right]^{2m_3}$$

Since this is a function only of i for any given constant population size N, we can determine $Pr(n^{[1]} \mid n^{[0]})$ for all combinations of n vectors and for the complete transition matrix for a given N. This matrix, P, can then be iteratively applied for any consistent mating system and the nature of the ultimate results can be determined in terms of its roots and eigenvectors.

The method is generally applicable, and some simplifications are possible with further assumptions on the form of the different $F_i(x)$. Hill's (1969) results indicate that for alleles of small effect, and independent loci, N may be as small as 8 before the diffusion-equation approximations are bad. The larger the average effect of an allele is with respect to the variance (α_i/σ close to 1), the worse the approximation can be. However, the diffusion approximations cannot be considered poor for these limited models.

The other infinite model approximations as used by Griffing (1960) to predict response require that gene effects, as a ratio of σ^2, be of small magnitude, and Latter (1965) has shown that genes of large effect can lead to much larger or much smaller gains than predicted as well as to differential amounts of change according

to the direction of selection. Linkage can further restrict the useful genetic variations and reduce total response to selection, but the effects of restricting population size may outweigh those of linkage (Latter 1966). Similarly, an additive gene-action model on linked loci showed relatively little effect of linkage on response to selection except when differences in gene effect are large and when one locus may affect the probability of fixation at the other (Hill and Robertson 1966). However, some genetic variances do change as gene frequencies change in some predictable ways (Nei 1963), and epistatic effects do occur, influencing selection response as allelic frequencies change. If such conditions as epistasis do exist, then linkage effects can become important and gain estimates by Griffing's (1960) approximations can be very poor, especially in small populations (Gill 1965b).

Thus, predictions based on esoteric formulas must be examined very closely before too much reliance is placed on specific results. Most of the difficulties and disagreements among the various analytical and computer simulation studies, however, occur when the effective population size is very small, less than 8. In comparisons of the various approximate gain estimation procedures, including dominance-effect models with a normal error distribution (Kojima 1961) and iterative transition matrix models based on them (Curnow and Baker 1968), little bias is found as contrasted with Robertson's predictions when $N_e > 8$ (Pike 1969).

Under somewhat more complicated genetic models, including epistasis or overdominance at some loci, the requirements for reasonable robustness of the various estimators of gain increase the recommended population size that must be carried (Gill 1965a). When the more involved genetic models are used, genetic loci do not behave linearly and additively, and alleles which might be required for ultimate progress are more easily lost. The genetic variances themselves change during a selection program, and predictions based on assumptions of constant variance are unreliable. The effects of inbreeding depression under various dominance and epistatic conditions further complicate the response predictions. Gill's (1965c) computer simulation studies indicate that effective population sizes should be kept above 30 to avoid excessive loss of otherwise favorable alleles. The trends in linkage effects and selection on means and variances for 40 loci on 8 chromosomes clearly indicate that the effects of small N_e are rapidly felt and that alleles are easily lost through the joint action of selective breeding and drift.

SELECTION MODELS

The results of these theoretical analyses and computer simulation studies may be summarized as suggesting that only for the simplest gene-action models, and then only for reasonably large population sizes, do the simple models of Griffing (1960) and Robertson (1960) apply as they recognized themselves. The more exact analyses indicate that for simple gene models, and N_e greater

than 8 to 10, the approximations used in their derivations are not bad and the general results can be reasonably accurate. However, extensions of the model to include large allelic effects, strong dominance to overdominance effects, and epistasis, make the effects of population size less predictable for means and variances of quantitative traits and for probabilities of fixing the desired allelic combinations (Latter 1966). In addition, the utility of Robertson's analyses were specifically investigated by Rawlings (1970) for plant breeding programs and several of the assumptions and derivations were found wanting. For example, one of the derivations used by Robertson requires that $N_e\zeta$ be small in order that the approximations used be accurate, but if $N_e\zeta$ is around 0.6 as required, then for reasonable levels of heritability and selection intensity, N_e must also be less than 8. Since most tree breeding will involve $N_e > 8$, the predictions of selection limits may be quite imprecise. In attempting to account for these errors of approximation, the factor $N_e\frac{z}{p}$ can be translated into the multiple-locus case by dividing the total selection differential effect into as many loci as desired using the approximation for each locus of:

$$\frac{\alpha_i}{\sigma^2} \approx \frac{2(h^2/m)}{q(1-q)}$$

where h^2/m is heritability divided by number of loci affecting the trait. For simple additive effect models, and to provide a high probability of fixing the favored alleles, Rawlings finds the required minimal $N_e\frac{z}{p}$ values given in table 1. Thus, at highly intense selections for alleles at low initial frequencies, a quite large N_e is required. When selection intensities are low or when many traits are simultaneously selected for, the requisite N_e increases rapidly, especially for those low-initial-frequency alleles which can easily be accidentally lost. The biases are most seriously felt in predictions of long-term progress and ultimate probabilities of fixation and less so in gain estimates for a few generations of selection.

Table 1.—Minimum values of $N_e i$ to give $\mu(q) > 0.95$

Initial gene frequency (q_0)	Heritability/loci (h^2/m)				
	1/40	1/200	1/1,000	1/2,000	1/10,000
¾	4	9	20	28	63
½	7	15	33	42	104
¼	12	26	58	84	188
1/10	21	45	100	139	313

It may be further remembered that N_e refers to the effective population size, which can be considerably smaller than the actual number of genotypes used. As relationships among the genotypes increase due to disproportionately high representation of some families, the coancestry within breed populations may not be controllable (Burrows 1970), and the actual numbers required may be much larger than the N_e figures given. By controlling allowable levels of inbreeding and making family sizes more equal than would occur in random mating, some of the expected decreases in N_e can be avoided. By controlled intermating of selected parents such that each parent is equally represented in the progeny population, N_e can be larger than the number of parents at lone heritabilities and only slightly lower at moderate heritabilities (Rawlings 1970).

An alternative method for obtaining purebreeding populations with somewhat lower probabilities of loss was suggested by Baker and Curnow (1969). They suggest splitting the single population into smaller sets and breeding within each for several generations and then selecting in only the best subpopulations. While the smaller subpopulations will lose favorable but low-frequency alleles more quickly, the more immediate 5- to 10-generation gains are made with intermediate-frequency alleles of large and moderate effect anyway. As long as subpopulations are kept at $N_e > 16$, not many of those alleles will be lost though some variations among the subpopulations can be expected. Thus, the average of all subpopulations will be slightly lower than the expected gain in a single large population, but the best one among the several subpopulations is expected to be substantially higher. Furthermore, if several replicate subpopulations can be developed, the best among these may then be intercrossed to produce a new base population for advanced sequences of population improvement, taking advantage of the variations in the loci fixed for alternate alleles and any formerly low-frequency alleles maintained at higher frequencies in any of the replicates chosen. However, the advantages may often be quite minimal (Madalena and Hill 1972) and would certainly involve more complex breeding programs.

For hybrid breeding programs, selection within populations which provide the parents for hybrid seed also requires the advancement of gene frequencies, and the only major difference in developing the recurrent selection population is that the gene frequencies are moved to diverge as much as possible between the two populations. Otherwise, the cumulative improvement of the recurrent selection populations is under the same restrictions of selection differential and N_e as for purebred populations.

For these simple gene-action models and for all breeding systems, it seems desirable to keep a high selection intensity by generating large populations from which to select a minimal number of parents. Intensive selection may thus be coupled with a sufficiently large N_e that immediate gains can be achieved without greatly sacrificing future gains. At the higher levels of selection

intensity, vast increases in numbers examined may be required to significantly increase the selection differential. On the other hand, increases in selection intensity when the number of parents is fixed are most easily achieved by increasing population size when population size is initially relatively small. Hence, subdivision can yield substantial advantages to breeders. In addition, if time can be afforded, subdividing selection into generational sequences may yield savings in the sizes of populations necessary to carry in each generation. For tree breeders the time costs may be excessive, but for short-generation species like cottonwood, the advantages may be significant.

SELECTION EXPERIMENTS

Theoretical investigations such as we have been reviewing, even for simple models, lead to some imprecision in predicting long-term results, and many variations in gene action, frequencies, dominance, epistasis, linkage, etc., can occur in actual populations. Tests of selective predictions with real organisms are therefore needed to indicate how well the reaction of some population systems is approximated by the theories. Most of the population testing has been done with animal populations in which family sizes and selection for many generations could be controlled. One of the primary difficulties in both directly testing theories and in applying theoretically advantageous breeding methods is a correlated decline in reproductive fitness as size or other economic measures are increased. Tree breeders often can partially overcome this problem through extensive cloning. By developing large numbers of fruiting branches, he can often obtain sufficient numbers of viable seed, even though the genotype is a relatively poor seed producer. Linkage, epistasis, dominance, and relations with the fitness factors bias the results of selection. Even without direct effects of the trait selected on fitness, an interaction between them can exist. There seems little doubt that for normally cross-bred organisms, restriction of population size leads to fixation of deleterious alleles by either random or directed, correlated selection effects (Latter and Robertson 1962). Under selection, there is also a tendency to create more relatedness among parents than if random mating occurred unless coancestry is strictly controlled. In addition, by controlling reproductive rates to equalize population sizes instead of allowing random selection and mating to occur, the hidden effects of natural selection against reproductively deleterious alleles can be ameliorated in the selected group. Thus, different traits even with the same heritabilities may exhibit different responses to selection according to their allelic relations with fitness, linkage, etc. Certainly different species will respond differently to selection and restrictions of population size, numbers of alleles, etc.

For hybrid breeding programs in which the product is a cross between selection populations, the selection populations themselves

may be inbred with little consequence of the inbreeding depression except on seed-production capacity. The deleterious effects of loss of alleles would be present, but inbreeding depression would not affect the theory of selection advance.

It is, therefore, clear that wide testing of selection theories is required to determine any generalities about natural populations from which guidelines may be drawn for untested populations.

Among agronomic crops, direct indications of response to selection and random mating among the parents are available from the long-term selections in such species as alfalfa, sugar beets, corn, wheat, and barley. (See Allard 1960; Penny and others 1962, 1966; Sprague 1966, 1967; Smith 1966, for review.) In tests with relatively mild selection intensities and most often with large populations, long-term response has been steady. Even after 100 generations, the populations can respond to mass or bulk selection in which phenotypic selection and random mating are performed. While the long-term experiments are not conducted in strictly controlled environments, and some environmental variations must have occurred over the years, the direction of selection has been persistent and the response always positive. Furthermore, most shorter duration tests also show substantial responses to selection for additively inherited quantitative traits even for small initial-population sizes.

However, in the breeding process, possibilities for further genetic advance can be eliminated, as was particularly evident in the lack of response in sugar beets to continued intensive selection for sugar content, root form, and other traits. While 100 years of mild selection increased sugar content by over 100 percent, from 7.5 to 16 percent, advanced intensive selection has netted relatively little advance. Failure may have been caused by changes in the gene effects themselves as major physiologically limiting factors were met. Perhaps new combinations of genes and traits are required for any new advances. Severe inbreeding has persistently led to loss of genetic variance and inbreeding depression in all cross-pollinated crops, even when efforts were made to select for inbreeding ability and to save the lines. The loss in fitness is partially due to directed selection for traits which directly affect survival in noncultivated environments. This conclusion was demonstrated in selection experiments in which the selected types were placed under no selection for a few generations or were actually placed in direct competition with bulk varieties. Either the trait suffered selection towards the nonselected mean, the variety displayed a relative loss of competitive ability, or both. However, there are also debilities from inbreeding depression in which even the most strenuous efforts fail to carry lines to survival without competition. There is usually some variation among individual lines with some surviving with vigor equivalent to the bulk variety, but fitness loss is the common expectation (Laude and Swanson 1942; Bal and others 1959; Allard 1960). The loss

in fitness may be related to competitive ability (Harlan and Martini 1938; Lerner 1954; Finlay 1963), susceptibility to predators, weak reproductive mechanisms, or some combination of these. In species in which vegetative vigor or stem mass is selected for, debilities of the reproductive organs down to some limit may not be significant, but any loss of vigor would directly affect the efficacy of selection.

The lack of response to selection may sometimes be due to a loss of alleles which might otherwise provide a basis for continued response. If traits were affected by a few alleles of large effect, then fixation at small population sizes can quickly exhaust the available variability. Crumpacker and Allard (1962), for example, estimated that heading date in wheat was controlled by only three major genes but that many minor genes also affected its inheritance. When gene frequencies for the major genes can be advanced close to one, then progress with the action of the minor genes may then be effective, if somewhat slower.

In tree species, the early indications of inbreeding indicate that vegetative vigor and survival traits are directly affected by inbreeding depression (Franklin 1968) and are therefore subject to both of the detrimental effects of limited population size.

While many organisms have been studied in long-term selection experiments, mice and *Drosophila* have been intensively worked organisms and provide some illuminating experiences in selection experiments. As in plants, the common experience has been that with reasonably large populations ($N_e \geq 50$) and moderate selection intensities, response for the first 10 or 20 generations is quite uniform and of a size according to the heritability. Thus, Kojima and Kelleher (1963a) state: "From these findings it may be concluded that the total response in the mean of the population continues to change, on the average, linearly in the direction of selection during the early period of selection, regardless of the kinds of organisms and traits and of the methods of selection."

However, there are limits to the generality of the results in both the physiological gene-action effects and in the loss of fitness by inbreeding depression and the loss of usable genetic variations. Thus, in mice, both upward and downward selection for body weight reach limits in 17 to 22 generations (Falconer 1955) and in *Drosophila* a plateau in response also occurs in 20 to 30 generations. In many such experiments the population carried is reasonably large, and at the stage when plateaus occur, genetic variations still exist (Falconer 1960), sometimes with even higher heritabilities than were originally present (Robertson and Reeve 1952). Thus, limits to selection which cannot be caused by exhaustion of genetic variation, inbreeding depression, or linkage with deleterious fitness factors also exist. These may possibly be the effect of a changed physiological or genetic milieu and hence different gene effects, or due to the existence of complicated epistasis. If gene actions are so complicated or if natural selection opposes any particularly directed selection, then the relaxation of

selection should cause a decline in mean response. In these cases, as well as in the case when different alleles may have been fixed in subdivided populations, a breeding procedure which can utilize gene differences among subpopulations is required.

A series of studies by Frankham and others (1968a, 1968b, and 1968c) is particularly enlightening. Those authors conducted replicated tests of a two-treatment factorial combination of population size (10, 20, and 40 pair matings per generation) and selection intensity (selection proportions of 10, 20, 40, 80, and 100 percent) over 50 generations. Over the first 12 generations, the strongest effect was that of selection intensity. Gain was clearly linearly related to selection percentage, except that the 80-percent selection level (four-fifths saved) was almost indistinguishable from the control population (100 percent saved). These results agree well with Kojima and Kelleher's (1963b) observations. In contrast, the effect of population size is not as strong as that of selection level in the short run, but the larger population tends to attain a higher total gain at the same selection intensities. In fact, the 40-percent selection population with 40 pair matings did exceed the more intensive 20-percent selection carried with 10 pair matings. Also, in terms of the ratio of achieved gain to selection intensity (the realized heritability), the milder selections tended to exceed the more intensive percentages and indicated that a longer period of response, and eventually a greater total response, may be obtained at the 40-percent selection level than at the 20 or 10 percent. It would thus appear that even in the short run, the increased gain by increasing selection intensity can be detrimentally affected if the populations maintained are small. While milder selection around 50 percent may provide a long, slow gain, at least the number of pair matings should be kept large.

In general, however, the response is a linear function of the initially estimated heritability and selection intensity and somewhat less of N_e. While the effects of selection intensity on realized heritability (i.e., achieved gain ÷ reach) are not clear, the early responses suggest that both small N_e and high selection intensities tend to reduce total gain. The variation among replicates of the populations was so high, however, that no one population could be expected to follow these average trends very closely. The smaller population sizes in particular exhibited great variations in response, indicating that at least sampling variations affect the replicate variance in the stochastic processes involved in the generational sequences. Even under these controlled environments, and with an organism adapted to those controls, the gene effect, selection, and mating processes generate substantial variations among identically treated (with respect to population size and selection intensity) population trials.

In the longer run, of 50 generations, most populations still appear to be responding to selection. The higher levels of selection intensity also still produce higher responses per generation.

However, the effects of population size, which earlier were not clearly established, became a major factor in determining response. By the 20th generation, there is a clearly established effect of larger population sizes on increasing response as lower selection intensities begin to exceed higher selection intensities if they also have larger population sizes. By the 50th generation, there is a rough equivalence in response of the main effects of increasing population size and of increasing selection intensity, and hence there is a much greater realized heritability for the larger populations. In addition, while all populations show some reduction in genetic variance, the larger populations continue to display a higher response rate than the smaller ones. It was clear that even with 16 percent heritabilities and using simple mass selection, the population responses of the larger populations under 10 percent selection exceeded the original mean by 1 standard deviation in two generations and by 2 standard deviations in five generations, with continued response after that. Thus, in relatively few generations, the population means far exceeded the original extremes.

In additional tests of Robertson's (1960) suggestions that early selection might advance gene frequencies into a safer intermediate range, smaller population replicates were split off from the 10-, 20-, and 40-percent selection populations with 40 pair matings each after 16 generations, by sampling 10 pair lines and breeding in those at their same selection intensities. All subpopulation splits of 10 pair lines immediately fell behind the larger population and the lag accumulated. This is a clear experimental counter evidence to the concept that it is generally safe to restrict population size after an initial period of selection in a large population.

The variations among replicate populations, especially smaller populations, remained very large, tending to increase as a function of the mean response and hence increasing as selection intensity increased. The larger populations continued to exhibit less variation than the small ones. In addition, the variation among populations was exhibited when temporary plateaus and rapid responses alternated. While the average response for the replicates at each of the selection intensities was reasonably smooth, individual replicates varied widely in size and period of response. The average declines in fitness, as tested in lines drawn from the selected lines and placed under relaxed selection, were moderate and lasted only a few generations. Therefore, there was only a moderate amount of natural selection opposing the directed selection. Some individual lines, however, did regress strongly due to recessive lethals still being carried and possibly also due to strong epistasis and linkages.

These long-term results indicate that epistatic interactions and the formation and destruction of linkage blocks can be important in holding genetic variations in populations, at times impeding, then aiding response to selection. The populations continue to respond to selection and though there is some moderate decline in

heritability, large variations among replicates also exist. In checking the state of the populations, it was found that some lethals were still present and affected selection response, but that the genetic variance was by no means exhausted. Several loci with large-effect alleles were still present at intermediate to high frequencies and some at low frequencies. Thus, it was concluded that several large-effect genes at low initial frequencies can continue to affect selection response long after one might otherwise assume their fixation. In addition, the presence of complicated linkage and epistatic effects can so confound the response to selection that useful genetic variance can also persist for many generations, especially at the larger population sizes and lower selection intensities.

From these various theoretical and long-term experimental studies, the possibilities of progress from selection and simple mating schemes among selected parents can be broadly sketched. For traits which are inherited in a truly quantitative manner, the response to selection is a reasonably linear function of the narrow-sense heritability or selection intensity, at least in the short run. If heritability is well estimated, population sizes are kept high, and truncation selection applied with accuracy, the average linear estimates of gain, such as by Griffing's (1960) formula, should be reasonably close. The effect of severe restrictions on population size, however, is felt even before the 10th generation, and can have major early effects if there are large-effect loci at low frequency in the population or if epistasis and linkage are strong. Thus, even in the early selection generations, a large N_e is required. Furthermore, since large-effect loci may possess the favored allele at low frequencies for many generations, a continuously large N_e is required for continued selection gain. Since N_e in such sequential breeding populations is sensitive to occasional bottlenecks, a continued monitoring of N_e is required.

Since gain is therefore affected both by the selection differential and N_e and since the two are somewhat antagonistic objectives, some compromise is required. That is, the more intensively selection is applied, the smaller the number of selected parents will be. The problem is easily avoided by producing and examining large numbers in the intervening progeny generations between selected parental generations. Since the selection differential is a function of the proportion selected while N_e is largely a function of the numbers selected, the obvious solution is to increase the base population from which the parents may be selected. This may be done to maintain a minimum acceptable effective number of parents so that the expansion of the population examined increases the selection differential or may be carried out by proportionately increasing both N_e and z/p. The cost of increasing the selection differential by increasing the numbers of trees tested can be very high, as noted by Shelbourne (1973), where the increase in z/p by a factor of 2 requires vast increases in test size (figs. 9 and 10) at higher selection intensities (Namkoong and Snyder 1969).

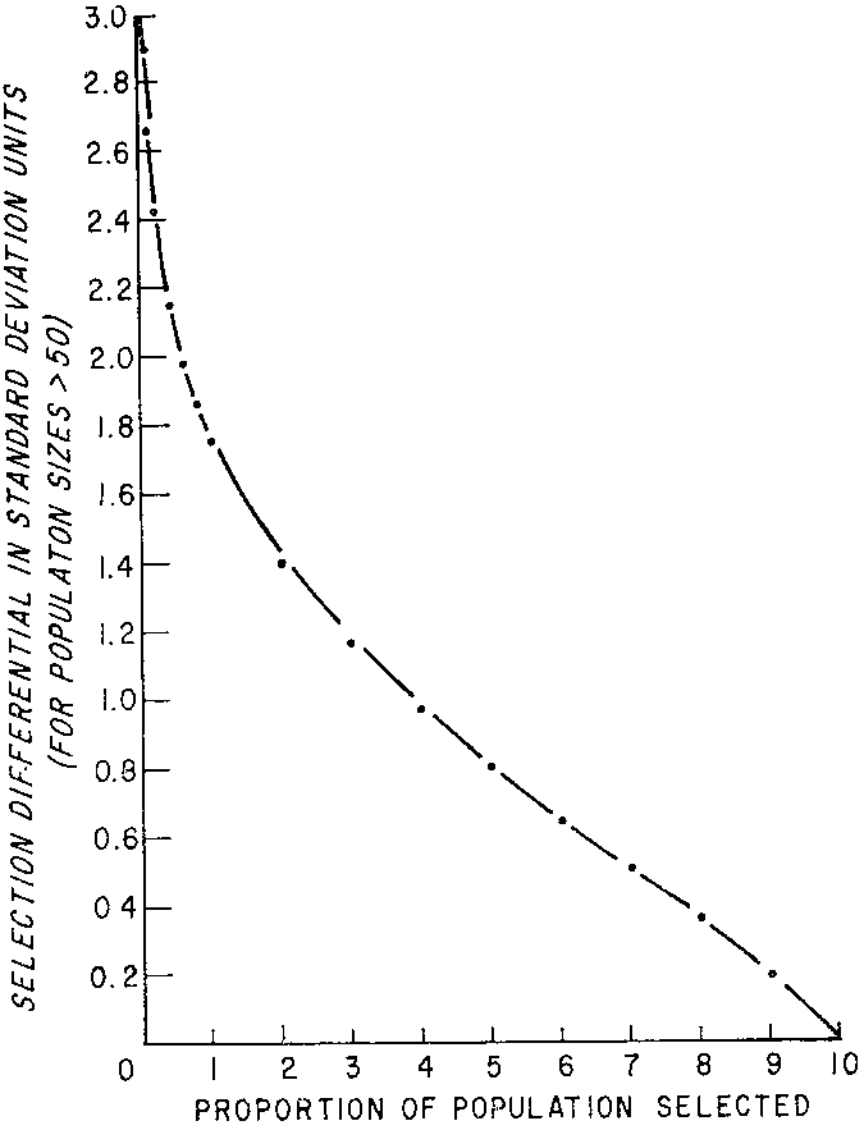


Figure 9.—Relationship of selection differential to proportion of examined population which is selected. (Shelbourne 1973)

If the cost of increasing the test population is high relative to the cost of the generation time, one may save by selecting in tandem sequences at lower selection proportions (Rawlings 1970). Nevertheless, there is a constant requirement for keeping large N_e and as high a selection intensity as compatible for short-run gains per generation. In the long run, less intensive selections would give greater total gains, but such plans would require careful examination for economic evaluation.

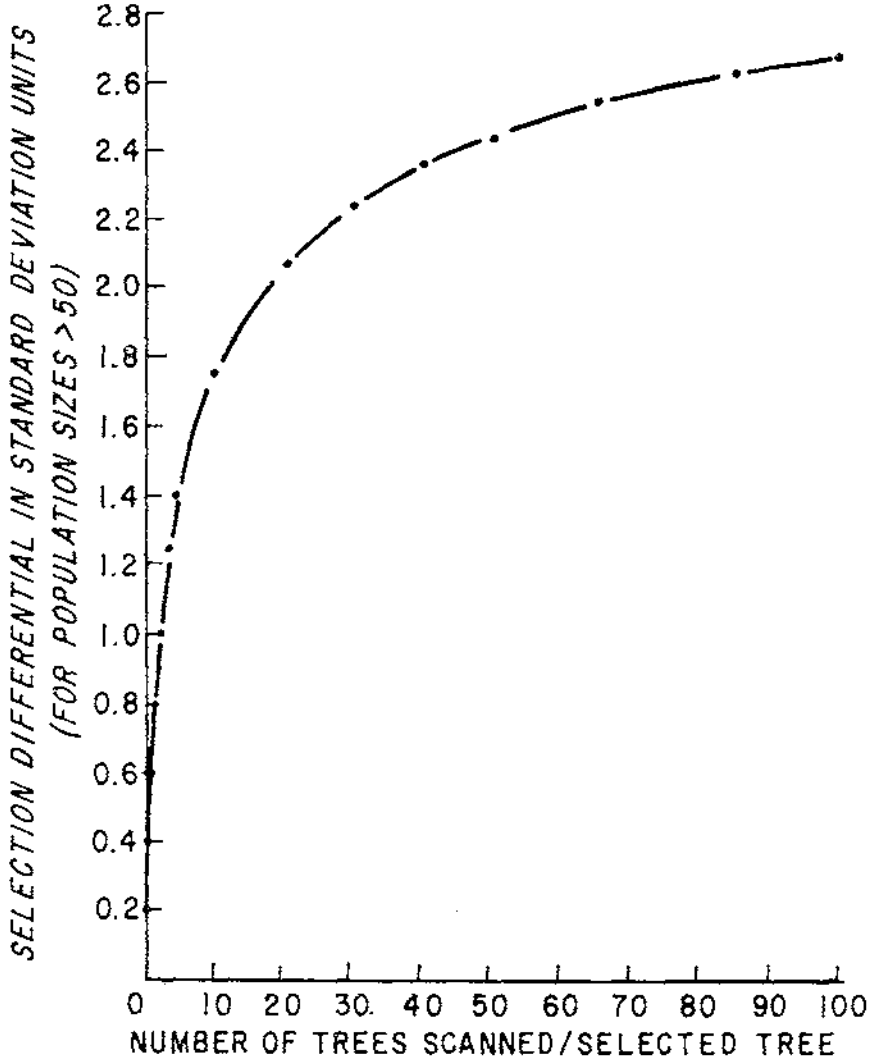


Figure 10.—Relationship of selection differential to number of trees examined per selected tree (inverse of proportion selected). (Shelbourne 1973)

Continued response from selection due to genes at initially low frequencies may also be expected regardless of the size of their effect. While large-effect alleles at intermediate frequencies and with small N_e may be quickly fixed, maintaining larger N_e can keep favorable, low-frequency alleles in the population for many generations while substantially increasing their frequency. Therefore, large population sizes, especially in the founder populations, can significantly affect progress for many generations. The effect of such low-frequency alleles will also be felt in crosses among any subdivisions of a larger population as may be developed. This is especially important if few alleles are expected

to exist in natural populations at frequencies close to $\frac{1}{2}$.

It may be common for frequencies of favorable alleles to form a bimodal or skewed unimodal curve. Traits presently or recently under selection would force gene frequencies to extremes unless overdominance is strong, which is unlikely to occur for very many loci. Such traits would tend to exhibit unimodal distributions with the allele favored by natural selection in relatively high frequency but not necessarily fixed due to the effects of slow response at high frequencies, migrations, etc. Loci with little present selective pressures might be more uniformly distributed except as drift would cause extreme distributions or as past selections would cause skewed distributions which have not been homogenized by gene migrations. It can thus be conjectured that the maintenance of genetic variance, which requires stable gene frequencies, and continued response to selection, which requires changing gene frequencies, are simultaneously possible to achieve. Both genetic variance and initial responses to selection will depend on intermediate-frequency genes or large-effect genes at low frequency. As these can change very rapidly, they soon will join the pool of high-frequency genes with little further effect on either variance or mean. The pool of genes at low frequency of favorable alleles must then be moved into the effective frequency range and can continue to slowly feed genes into positions to affect means and variances. Thus, the initial profile of gene frequencies can affect the continuity of response without linkage, epistasis, or other effects which further complicate the response patterns over generations. Thus, also correlations can change rapidly over generations according to which loci are changing frequency, while the total genetic variance itself remains stable.

In general, it might be concluded that over the generations of a selection program, the correlations among traits are highly susceptible to change as well as to very poor initial estimation (Bohren and others 1966). To effectively select for many traits simultaneously, therefore, requires an understanding of the mechanisms invoking the correlations to predict their changes as well as to modify them by selection. This further places a premium on keeping a high N, so that the opportunities for special selection of recombinations might be effective. The special association of traits under selection with reproductive fitness traits requires special attention for its modifying effects on selection.

INITIAL SELECTION CONSIDERATION IN FORESTRY

For most forest tree species which have not been heavily selected for fitness in plantation environments, the prospects for selection gains by almost any method for many traits of commercial importance seem temptingly unexploited. As long as founder population and subsequent effective population sizes are kept large, well-estimated heritability or selection differential can

be expected to reliably estimate average gain for quantitatively inherited traits. If large populations exist from which to select new generations, then intensive selection from many trees can still provide for a relatively large number of selected parents. Then, at a constant heritability, gain is maximized by maximizing the size of the populations in which selection is practical. For minimal breed population sizes, however, increasing the selection intensity by expanding the progeny populations or the wild, original population from which the breed parents are selected may often cause associated problems in adequately judging the selective value of candidate trees. Not only is it more difficult and costly to examine or test large numbers, but the tests cannot be held to standard conditions very well. Thus, as more numerous or more massive tests are run, the error in estimating selective value increases and the heritability of the value measure decreases. In addition, more variable environments are encountered and less control of age, spacing, or other significant variables will also decrease the heritability of the traits. Thus, even if costs of expanding the test populations can be afforded, there is no advantage to increasing the differential when the decrease in the heritability exceeds the gain in the differential. A further difficulty in measuring the selection differential occurs when large numbers of trees are examined and more errors are made in determining the trees which actually rank highest phenotypically. Tree breeders may therefore be faced with economic and physical limits due to the size and time requirements of trees not encountered to the same degree in other organisms.

The requirement for maintaining large, effective population sizes, however, is not diminished especially if traits can commonly be expected to have loci with favorable alleles at low frequencies and selection pressure per trait, per locus, is not very intense. Then, the need for large N_e for even short-run gain maximization is acute. Even if breed population crossing programs are considered for future breed development, the low-frequency alleles are still required for new recombinations to provide advanced gains. This is particularly acute for those species and means where the main natural populations are being replaced by the new breeds. However, even when large, unselected populations may remain, the new breeds can be expected to be such improved forms that any reselections in the original populations will be costly. Furthermore, even without the expected existence of inbreeding depression, the need for ancestral data and control of ancestral relatedness in future breeding populations is necessary. Since inbreeding depression does occur so commonly among forest tree species (Franklin 1968), and epistasis must eventually affect selection response, forest tree breeders will be continually selecting under some adverse effects of natural selection, regardless of any direct associations of the selection goals and reproductive fitness. Any loss of reproductive fitness may be overcome at a cost, by special treatment to increase fruiting branch tips as by

cloning, or by enhancing natural reproduction under controlle environments. Nevertheless, linkage, pleiotropy, and epistasis cai confound natural and artificial selection effects.

The tree breeder does have some considerable advantages in controlling selection progress which have not been considered in these simple programs. Breeding programs can be devised that offer many alternatives to the simple mass selection and random mating of selected parents which can aid ancestral control. Furthermore, the breeder can test and retest an individual and its relatives in controlled environments and hence increase the heritability of the value measure. By controlled mating, controlling family size and designing periods of relaxed selection or crossing among subdivided populations, or deliberately avoiding or minimizing coancestries, he can inhibit the reduction of N_e. Thus, breeding programs tailored to the organism and traits can have wide latitude in making the general selection program efficient and effective. Tree species present particular operational problems, and certainly each species and trait selection has unique problems in applying the general procedural principles of optimum selection and breeding methods. Nevertheless, within the limits of the reproductive mode and of extrapolating results too extensively, the general principles of plant breeding can be developed from the theories of selection outlined in this chapter and can be applied as outlined in chapter 3.

CHAPTER 3
BREEDING THEORY

In most breeding programs with large numbers of parents, inbreeding and coancestry can be maintained at low levels for many generations. Some mating plans for selected parents can rapidly lead to high levels of inbreeding or coancestry, however. Crossing all parents to a single male or female would induce high inbreeding and quickly create high ancestral relatedness among all members of the breed population. The effective population size would be rapidly diminished and selection progress reduced. Only by introducing new materials would new genetic variants be available for continued selection. If the new genotypes, however, were from an unselected or unadapted population, the breeding population would immediately suffer some loss in mean value and might take several generations to recover its former gains. For crops in which several cycles of intensive breeding can alter varieties in a few years, this course of action may be feasible. Until that is possible with forest trees, however, it is far more efficient to avoid the necessity. Reasonable gains can be achieved in selection systems that maintain large populations and in mating systems that minimize inbreeding.

Discussion of breeding methods aimed at fixing optimum combinations of major genes is beyond the scope of this book. Major genes, as contrasted with polygenes, have such a large average effect relative to variations in phenotypic expression that the genotypes can be identified with little error except for dominance or other masking types of gene effects. Relatively few loci would affect any single trait, and breeding methods to fix optimum genotypes would be simple and are well described in classical genetic texts. At this time, few economically important traits in forest trees are known to involve major gene effects but more will undoubtedly be found as data and measurement techniques develop. It is assumed that as detection of such genes progresses, they will be fixed by well-established procedures within breed populations that are also being developed for the totality of traits requiring genetic improvement. It is expected that many traits will be improved through a combination of major and polygene breeding methods, but that in forestry, primary emphasis will remain on polygene improvements for many generations. Regardless of the emphasis placed on one type of gene action or another, a continually improving base population is needed from which

subpopulations can be drawn for any special breeding procedures.

Our concern here lies with selection and mating programs designed for sequential development of breeding populations. These programs are distinguished from those for the expansion of genotypes into seed orchards for the production of seedlings, cuttings, etc., for commercial forests. We shall postpone for later consideration the problems of measuring economic and ecological value and temporarily assume that tests and measurements on trees have been made and that phenotypes have been accurately observed and evaluated. We shall also assume that all problems of planting, cloning, pollinating, etc., present no restrictions on our choice of either selection or mating design.

Regardless of the breeding method, each generation is expected to produce genotypes with cumulatively better collections of alleles. Multiple production of sibs by crossing among large numbers of parents, or among single pairs of genotypes, or by selfing especially good lines for commercial seed production is a technical problem for the breeder but is not treated here. Similarly, the best genotypes can be periodically chosen for vegetation reproduction, as in poplars, but again, we shall have to consider production problems as ones of technique in handling and distributing what the breeder produces. The concern of this chapter is on iteratively improving the breed population from which the best propagules can be drawn for commercial use. The next chapter focuses on the problems of identifying the best genotypes within any generation. We would generally expect that only the extremely best propagules would be used in reforestation in any one breed generation. Some ecological balancing to avoid the dangers of monoculture has to be considered as well as the optimum mixture of growth forms to satisfy the multiple use requirements of the forest land. In general, we do not expect to decrease the genetic sources of variance in the basic breeding populations. Therefore, breeding programs will not generally reduce the tree-to-tree variances as has sometimes been claimed. Only if a restricted subset of the breed is used in plantations can the genetic sources of variance in plantations be reduced. Otherwise, genetic uniformity can come only at the expense of the breeding program.

Three kinds of populations can be envisioned in each cycle of selection and breeding: (1) the selected parents which are mated in certain designs to produce the next generation; (2) the next progeny generation so produced, which serves as the base population for the next cycle of selection; and (3) the population of genotypes used in the production of propagules for commercial use. The last may be a subsample of the selected parents set aside for production matings or vegetative propagation. It may be more intensively tested and selected to a smaller set of parents or ortets for immediate propagation. It may then be used as a base population in short-term breeding programs for pure-line or single-cross production. On the other hand, if seed or ramet pro-

duction needs are pressing, the population for that production may include all of the parents that are used for the breeding population or an even wider sample of trees as may be required for commercial propagation. We shall be concerned with the first two kinds of populations and how they are generated by a selective reduction to some minimal set of parents, how the parents are mated to produce a larger and improved base population, and how the reselection is to start again. Methods for selection and breeding are given first for single populations. Distinction is made between breeding procedures used for the long-term breed development and breeding within any single generation. Hybrid breeding methods are then examined as is provenace selection, before integrated breeding program organizations are reviewed.

SINGLE-POPULATION BREEDING

In general, the breeding system used is highly dependent on the normal mode of reproduction exercised by the organism and its native sources of genetic variation. For normally crossbred organisms such as most forest trees, the maintenance of crossing among a large sample of genotypes can be achieved within a single population, even though the eventual population composition may include few genetic variants. Single-population breeding methods may be maintained for hundreds of generations, and may even then contain sufficient genetic variations to respond to changing ecological or economic objectives. We first examine the systems involving intensive inbreeding and then other systems which allow for less inbreeding and more control of coancestries.

In organisms such as corn and wheat, which can be adapted to selfing or high degrees of inbreeding, pure lines or pedigrees are commonly developed, though they may be outcrossed for commercial production. Much of the success of these methods lies in a good selection of the original parents and in the ability of breeders to advance many lines for many generations to derive the final, limited, selected set of genotypes for commercial seed production. A single inbred is usually grown for production, but the lines may be crossed for the seed released as well as for the establishment of new segregating populations from which new selections for pure-line developments can begin. If inbreeding depression or survival and reproduction are not too severe, and the genotypes selected can be accurately observed in spite of the opposing depressive effects of homozygosis, then pure-line breeding can be useful. For limited objectives on a few loci where homozygosis is beneficial, such pure lines may be profitable. If a line is already developed and a few loci or chromosome segments are to be substituted, then various backcrossing schemes may be useful. Such systems, which rely on the development of invariant genotypes, are most easily carried out with natural selfers apomicts, or those that rely relatively little on genetic recombination for reproduction. Most such species—tobacco, oats, peanuts, etc.—

exhibit some degree of outcrossing but often have relatively low chromosome numbers and recombination frequencies.

INBREEDING SYSTEMS

Systems of selfing and partial combinations of half-sib and progeny testing systems may be constructed to fit the requirements of the organism being bred. In forest trees, however, most such family selection systems are not required since the original, relatively unselected parents can be saved. Furthermore, application of such systems to trees would often entail severe inbreeding depression. Thus, selfing systems and pure-line breeding, in which single genotypes are sought for development, are not often practical for trees. The half- and full-sib family selection systems are usually not followed by sequential inbreeding within selected families but rather are used in recurrent selection schemes for either general or specific combining ability. Such systems, however, are possible to develop with forest trees as exemplified by the Douglas-fir selfing system used by Orr-Ewing (1965).

In contrast to recurrent selection systems in which intermating among selected parents is sequentially used to generate new, generally cross-pollinated populations in which variability is maintained, line breeding extracts more purebreeding homozygous genotypes either for self-propagation or for use in specific crosses or in synthetic varieties. Forest geneticists will generally start with unselected crossbreeding populations, which will generally resemble the F_2 populations used by pure-line plant breeders. The emphasis in such breeding systems generally lies in maintaining line samples from as large a proportion of the base population as possible and not on selection among lines until the final generation is reached. However, one may select among lines according to pedigrees or may segregate especially good bulked populations to concentrate effort on more promising lines even during the early generations. The balance achieved would presumably depend on the need for commercial breeding lines during the intermediate generations, and the trustworthiness of early-generation selection. The heritability appropriate for computing gain from selections includes the total genetic variance among lines in the numerator and the phenotypic variance of line means in the denominator. Since the genetic variance changes with inbreeding and gene effects and since responses to environments may affect the phenotypic variance differently from generation to generation, it will likely be necessary to reestimate the appropriate components of variance more often than in the recurrent selection systems.

HYBRIDIZING INBREDS

Since severe depression is an expected consequence of inbreeding forest tree species, few tree breeders expect to use inbreds directly as the commercial material. Instead, inbred lines may be

used as parents in some form of crossing system among lines to create a relatively vigorous hybrid. Single crosses take but one generation to develop in contrast to triple, double, and higher order configurations and might, therefore, be expected to be more commonly used than others. A single cross may be widely planted, but such a practice would be subject to the same dangers as any other monocultural system. Instead, sets of single crosses may be used in commercial plantings. It is also possible to test-cross among many possible line combinations and use only crosses among those which combine particularly well in a synthetic variety. Such synthetic varieties are simply the product of intercrossing among a small number of appropriately selected parental lines and resemble recurrent selection populations except that the source of material is generally more uniform genetically within each parental unit. Specific test crosses and plantings for selection of good hybrid vigor are made and entries into the synthetic are determined on that basis.

At present, there is little information on forest trees to indicate that pure breed-hybrid systems can overcome the difficulties in maintaining such lines (Franklin 1970a) or that selection among selfs using additive gene effects can be effectively used for developing synthetic varieties. However, little effort has been expended in these directions.

MASS AND SIMPLE RECURRENT SELECTION

Various programs which eventually develop a uniform breed or variety, but without selfing, are possible to carry out with forest trees. Mass selection with a limited number of parents selected each generation is one example. While these methods eventually also rely on additive types of gene action and in the very long run would eventually lead to pure breeding varieties, inbreeding can be controlled and the normally outcrossing behavior of most tree species can be maintained. Thus, in mass selection, the open, randomly pollinated seeds from selected parents constitute the progeny generation from which the next generation of parents is selected. Simply maintaining a large N in the parental populations assures a reasonably large N_e and hence some continual variation in the breed. Under recurrent selection, the matings would always be among the selected parents. With perennial organisms that fruit repeatedly, there is no need to self the selected parents to keep their genotype intact and hence the only major difference between mass and simple recurrent selection is that recurrent selection requires the systematic intermating of all possible selected parents instead of simple random mating with only female parent identification. Thus, the characteristics of mass and simple recurrent selection in trees are the lack of test crossing and the more or less complete intermating of selected parents with uncontrolled versus controlled pollen parentage.

Mass selection is the simplest system to operate and requires little time or effort. It may, therefore, be the most common breed-

ing system. However, simple recurrent selection such as practiced in clonal seed orchards can also be very economical. Agencies that must provide seed for a planting program with many species may find that establishment and seed costs are only negligibly increased, if at all, by establishing clonal seed orchards (Perry and Wang 1958). Hence, minimal breeding programs with either of these two methods are readily justified economically.

At a slightly increased cost, pollination can be partially controlled and parental sources can at least be identified in the field if planted in blocks or rows and identified for selection. Some coancestry control and data for testing parents can therefore be obtained relatively cheaply. Large N is thus generated at a small loss in future N_e and precision of selection.

There are alternatives to testing and to the patterns chosen for controlled matings. For instance, determining general combining ability of individuals may require test matings on the trees in the base population to some general tester set of trees. The best performers can then be completely intermated as in simple recurrent selection. Because the base population genotypes of crop plants often cannot be directly observed or saved, some form of family selection is practiced. Since the relatives would not have the same genotypic composition as the original plants but may be more precisely tested, the expected gain from selection is not easily derived. For example, if the original plants are lost and only open-pollinated seed are available for testing and subsequent use as selected parents, then the selection of such a half-sib family on the basis of performance in a replicated test can be very precise. However, the individual plant(s) chosen to represent the family as a new parent is only a half-sib of the plants tested. While such measures are rarely necessary in tree breeding, the results are quite similar to the practice of collecting open-pollinated seed and selecting among these half-sib families on the basis of the family performance in test plantations.

To estimate the gains from such procedures, we can develop the concept of heritability in a somewhat different way than Griffing (1960) did but with essentially the same approximations and limitations. The result is easier to apply to plant breeding situations (Empig and others 1971). If we again consider a simple genetic model with many independent loci, each affecting the trait in a similar, small, and cumulative way, then the effect of selection can be estimated for one locus and added over all effective loci. Using the model of gene effects as in chapter 7, the genotypic and phenotypic mean of the population is:

$$\mu = u_i[q_i^2 + (1-q_i)^2 + 2q_i(1-q_i)a_i] = u_i[q_i + (1-q_i) + 2q_i(1-q_i)a_i].$$

The total genetic variance computed from

$$\sum_i \text{frequency} \times (\text{genetic value})^2 - \mu^2 \text{ is:}$$

$$\sigma_G^2 = q_i^2 u_i^2 + 2q_i(1-q_i)a_i^2 u_i^2 + (1-q_i)^2 a_i^2 - \mu^2.$$

The additive genetic variance is:

$$\sigma_A^2 = 2q_i(1-q_i)a_i[1+(1-2q_i)a_i]^2.$$

The remaining dominance variance is:

$$\sigma_D^2 = 4[q_i(1-q_i)a_i a_i]^2.$$

The total phenotypic variance (σ_P^2) includes a purely environmental variance component σ_E^2, and assuming no correlations nor interactions of genotypes with environments,

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

Then the population change under a selection differential of s is a function of the change in gene frequency and the change which those frequencies would have on genotypic or phenotypic means. In a normally distributed population, $E(s) = i\sigma$, or $(z/q)\sigma$ as previously derived in chapter 2. In the following discussions, we shall assume that the s is appropriately determined for any given distribution and is determined by considerations of minimum N_e, total numbers to be examined, and cost factors independent of heritability. We also assume that the populations are large enough or are replicated sufficiently that variations in s and h^2 are not important. In forestry, the assumption of independence of s and h^2 is questionable, but discussion of this is temporarily postponed.

The change in gene frequency can be approximated by a linear regression of the frequency on genotypic value which would be:

$$\frac{\text{Cov (gene freq., phenotypic value)}}{\text{Var (phenotypic value)}}.$$

In mass selection, the allelic frequencies for each genotype and the selection values are:

Item	Genotype		
	$A'A'$	AA'	AA
Genotype frequency (f)	$(1-q)^2$	$2q(1-q)$	q^2
Mean phenotypic or genotypic value (x)	$-u$	au	u
Allele A frequency (y)	0	½	1

The mean frequency of allele A is:

$$\bar{y} = (\tfrac{1}{2})\ (2q(1-q)) + q^2 = q.$$

The mean phenotypic value is:

$$\bar{x} = [q^2 - (1-q)^2]u + 2q(1-q)au.$$

The sum of cross products between gene frequency and phenotypic value is:

$$\Sigma fxy = q^2u + q(1-q)au,$$

and therefore the covariance is:

$$\Sigma fxy - \bar{x}\bar{y} = q(1-q)u[1+(1-2q)a].$$

The effect of such a gene-frequency change on phenotypic mean ($\bar{x}$) can be approximated by:

$$\frac{d\bar{x}}{dq} = \frac{d[u[q^2-(1-q)^2+2q(1-q)a]]}{dq} = 2u[1+(1-2q)a].$$

Therefore, the gain is:

$$\Delta G \approx s \cdot \frac{\text{Cov (gene freq., phenotypic value)}}{\sigma_p^2} \cdot \frac{dx}{dq}$$

$$= \frac{s}{\sigma_p^2} 2q(1-q)u^2[1+(1-2q)a]^2,$$

and from the definition of σ_A^2,

$$= s \cdot \frac{\sigma_A^2}{\sigma_p^2}$$

summed over all loci.

Finally, we may observe that the $\frac{dx}{dq}$ is in effect approximated by the covariance between the new genotypic values and the gene frequencies when divided by $q(1-q)$. Therefore, the product of the two, as used to approximate gain, is:

$$\text{Cov} \cdot \frac{dx}{dq} = (\text{Cov})^2 \div q(1-q).$$

Since $(\text{Cov})^2$ is a function of the additive genetic variance, σ_A^2 is a function of the covariance between the phenotypes selected and the expected phenotypes or genotypes of the selectively regenerated population. This assumes that the selected parents are randomly or completely intermated. Thus, the concept of heritability as a regression coefficient with Cov (phenotypes selected, genotypes generated) ÷ phenotypic variance is a useful approximation to actual expected changes in allelic and genotypic frequencies.

FAMILY SELECTION

In half-sib family selection, as detailed by Empig and others (1971), the family means are estimated using random matings with the unselected population, and the best families are intermated to produce the next generation. We ignore individual selection within families until later in this chapter. In family selection with crops, the common parents of the families usually are not available, and the actual units selected are the open-pollinated or

randomly mated progeny which are half-sibs of the test materials. Thus, parents of the AA genotype occur with frequency q^2 and have randomly mated offspring of types AA and AA' at expected frequencies q and $1-q$, respectively. Expected values of u and au, respectively, give a weighted mean value of AA parents of q^2 $[qu+(1-q)au]$, since the AA occurs with frequency q^2. The frequency of the A allele in the progeny of this family is expected to be $\frac{1+q}{2}$. Computing the similar expectations for the AA' and $A'A'$ genotypes gives the following set of values:

Item	Parent genotype		
	$A'A'$	AA'	AA
Genotype frequency (f)	$(1-q)^2$	$2q(1-q)$	q^2
Offspring mean value (x)	$qau-(1-q)u$	$\frac{(2q-1)u+au}{2}$	$qu+(1-q)au$
Allele A frequency (y)	$\frac{q}{2}$	$\frac{1+2q}{4}$	$\frac{1+q}{2}$

The covariance of gene frequency and phenotypic value is:

$$\frac{q(1-q)u}{4}\left[1+(1-2q)a\right].$$

The $\frac{dx}{dq}$ function remains the same as before:

$$\frac{d\bar{x}}{dq}=2u\left[1+(1-2q)a\right]$$

and hence:

$$E(\Delta G)=\frac{s}{\sigma^2_{\text{half-sib }(HS)}}\cdot \text{Cov}(q,\bar{x})\cdot\frac{d\bar{x}}{dq}$$

$$=\frac{s}{\sigma_{HS}{}^2}\frac{q(1-q)}{2}u^2\left[1+(1-2q)a\right]^2$$

$$=\frac{s}{\sigma_{HS}{}^2}(1/4)\sigma_A{}^2.$$

This value can be summed over all independent loci, since the $\sigma_A{}^2$ is that which is summed over all loci and σ^2_{HS} is the common expected denominator. The gain expectation could alternatively have been derived by simply noting that the covariance between the genetic value of the individual units selected and the test materials is that of half-sibs. The ratio of this covariance to the phenotypic variance is a heritability appropriate for half-sib

family selection:

$$h_{HS}^2 = \frac{\frac{1}{4}\sigma_A^2}{\sigma_{HS}^2}$$

and expected gain of:

$$E(\Delta G) = s h_{HS}^2.$$

In general, σ_{HS}^2 is more easily controlled and hence can be much smaller than σ_p^2 if uncontrollable environmental variances in σ_p^2 are large.

If the original parents can be saved for breeding by some form of self-propagation, then the matings for commercial production can be controlled to include only selected parents. This occurs, for example, when open-pollinated or bulk pollen may be used to produce test seedlings while the original parents' genotypes are grafted or otherwise preserved. Then, the parents are rogued according to the half-sib family tests and the selected clones allowed to intermate. In this case, the gene-frequency gain of the favorable allele is essentially doubed, and hence doubles the covariance between value and favorable allele frequency. Alternatively, the covariance among half-sibs may be viewed as being essentially constant, but the effective selection differential doubled. In either way of deriving the expected gain, the preservation of the original, undiluted genotypes and their intermating doubles the expected gain. This is actually a form of progeny testing on the basis of half-sib family performance, but the covariance relationships are most easily derived, as we have above.

It should be noted that while only $\frac{1}{4}$ or $\frac{1}{2}$ of the additive genetic variance is effective in the numerator of the gain heritability, the denominator is the variance of the half-sib family means used in the tests. In the nested design, as reviewed in chapter 8, the female half-sib family variance is:

$$\sigma_e^2 + r\sigma_m^2 + rm\sigma_f^2,$$

and the family mean variance with r replications and using a set of m male tester pollen per female is:

$$\frac{\sigma_e^2 + r\sigma_m^2 + rm\sigma_f^2}{rm} = \frac{\sigma_e^2}{rm} + \frac{\sigma_m^2}{m} + \sigma_f^2.$$

Thus, by increasing the r or m factors, this variance of a family mean can be considerably reduced if σ_e^2 is high as it must be if the h_{HS}^2 is low.

In full-sib family selection, the family means are estimated for each pair mating and the best families are propagated. If propagation is by random or complete intermating among the best families and specific dominance effects are not used, as they may be in special pair matings, then only the additive genetic variance is used in the numerator of the gain heritability since all dominance deviations are expected to be randomly distributed. Thus,

the genotypic mean frequencies, values, and the frequencies of favorable alleles would be given in Empig and others (1971) and yield an expected gain of:

$$s \cdot \frac{\frac{1}{2}\sigma_A^2}{\sigma_{FS}^2}.$$

In this case, σ_{FS}^2 is the variance among full-sib family means and would usually be slightly larger than σ_{HS}^2, since it uses just one other parent as a tester source of pollen. In terms of the nested analysis used for the σ_{HS}^2 composition, the σ_{HS}^2 would include:

$$\frac{\sigma_e^2}{r} + \sigma_m^2 + \sigma_f^2.$$

The numerator again is twice the covariance of half-sibs since the mating is done independently of any dominance or specific cross combinations which may give high yields due to specific gene interactions in the progeny.

If the original parents could be saved, or somehow the full-sib families could be reconstructed by selfed seed or vegetative propagation, the immediate gain could be enhanced with specific crosses which may have combined especially well. Such interactions, as measured by the deviation of the specific cross from what may be expected as the average value of the parents in other random crosses, is called the specific combining ability. It is a deviation from the average of the parental performances, which are their general combining abilities. In this case, the full covariance of full-sibs would be used in the numerator of the gain heritability and would include:

$$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2.$$

However, if cumulative gains are desired from such initial selections, the specific crosses have to be reconstructed to select for heightened specific combining ability. Otherwise, if a general crossing system is followed and new selections are made from a random or complete crossing scheme, only the $\frac{1}{2}\sigma_A^2$ can be used to predict cumulative gains.

HERITABILITY CONCEPTS

Since many types of breeding programs are available, forest geneticists are sometimes confused over the appropriate definition and use of heritability for each program. Animal breeders and geneticists originally defined heritability as either the total genetic variance ÷ total phenotypic variance (broad-sense heritability), or total additive genetic variance ÷ total phenotypic variance (narrow-sense heritability). Since plant geneticists apply different forms of selection and breeding, the proportion of the genetic variance that can be translated into gain is different for them. They can also change phenotypic variance at will with plots, environmental replicates, etc. Therefore, the ratio of genetic

and phenotypic variances requires further specification in plant genetics.

In the approximations of Griffing (1960) and Robertson (1960), simple genetic models were used and allelic effects were summed over independent loci. The assumed conditions were a form of mass selection or simple recurrent selection with random mating among the selected parents and no change in variances over many selection cycles. In the regression concept of heritability, as described by Empig and others (1971), and more explicitly developed by Hanson (1963) and Falconer (1960), continuous genotypic distributions are assumed; it is further assumed that recurrent selection procedures will regenerate all variances in each generation. Since the different derivations vary only in the contribution of N, and in some epistatic components, all provide the same predictive quality and all are susceptible to most of the same limitations and have the same model deficiencies.

THE NUMERATOR

In the regression concept, the numerator of the gain heritability is the covariance between the genetic value of the plants produced for ultimate utilization and the phenotypic measures used to estimate that genetic value. As noted above, a breeder may be interested in one of two types of produced materials: (1) the exact reproductions of the families or clones tested or (2) the randomly mated or completely intercrossed new population in which genetic recombination is expected to reduce the effects of specific combining abilities. When tested families can be reproduced by saving parents and remating according to test values, then the full genetic variances indicated by the covariance among family members constitute the numerator. If selfs or clones are produced, total genetic variance is included. If full-sibs are produced, the array of genetic variances should include:

$$\tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \tfrac{1}{8}\sigma_{AA}^2 \ldots,$$

because all these elements enter into the variance among full-sib families. Similarly, for half-sibs or any kinds of families produced, the genetic variance among the selection units that should be entered in the numerator of the gain heritability is the repeatable part of the variation. On the other hand, if selected family representatives are mated or if parents are completely intermated to generate a population for the next cycle of recurrent selection, then the full contributions of epistatic and dominance effects will be reduced by the extent to which the intralocus and interlocus allelic correlations are lost in the recombinations and matings. Then, regardless of whether the parents were selected on the basis of clonal, full-sib, half-sib, or other family mean values, most of the nonadditive genetic contributions to the differences among selection units will be lost in random or complete intermatings. Only by selfing or selective crossing, as in single crosses or

synthetic variety construction, or by recurrent selection for some specific combining ability, can the nonadditive variances be effectively used in a cumulative sequence of generational gains.

One factor that can influence the genetic variances in the numerator is the presence of cryptic differences between the measured traits and the traits desired for actual improvement. Since the performance of an individual or family under forest conditions may be only partially correlated with performance under test conditions, the appropriate heritability numerator should be the genetic covariance and not the genetic variance. Some compensation for differences between age, condition of test, etc., should be made when possibilities of measuring correlated instead of directly measured traits can be estimated. Further treatment of selection with correlated traits is postponed to chapter 4.

One other element may enter the numerator of the gain heritability due to nonlinear of nonadditive relations between genotypic and environmental effects—the genotype-by-environment interaction. If testing and evaluation are performed over a sample of environments, and genotypic value is determined as the average of each genotype over the various environments, then the genetic variances can be defined in terms of the plants' reactions to these environments. However, particular genotypes may perform especially well or poorly in certain conditions. If so, their potential yield on those sites will not be well predicted by an average yield over all environments, an average of the environments, or the mean contribution of both genotype and environment. Consider, for example, two families on two sites. If genetic variance exists on each site and family A scores 10 and family B scores 5 on site I; but A scores 15 and B scores 20 on site II, then the relative rankings change, average site difference from 7.5 to 17.5 would be observed, but no overall genetic variance would exist. Since there is no average difference among genotypes, selection for average performance would be futile. However, if the sites are classified and planting zones are distinguished for separate breeding efforts, then the genetic variances within sites can be used to predict gain and the genotype-environment interaction no longer is defined. Conversely, if general performance over all sites is desired but testing can be made only on a few sites, then genotype-environment interaction may cause bias in estimating general performance. Nevertheless, the interaction can be useful if it is recognized and if environments and genotypes can be altered to take advantage of especially favorable combinations of special trees on special sites. On the other hand, if the extra variations caused by these interactions are large, small environmental samples will not provide good estimates of true average genetic differences. Hence, the covariance between test performance and average genetic value is reduced by the extent to which the interaction adds to the family differences in the site(s) tested. Thus,

estimates of genetic variance taken on one site may apply only to similar sites. Some value for an interaction variance must be subtracted to predict the gain from selection for planting on many sites.

If fertility, stand density, or other controllable site factors have strong interaction effects such that some genotypes do exceptionally well under certain regimes, the interaction effects can be used to enhance average genetic values and gains. It may then be possible to select trees with good response to intensive culture, for example, and to combine the improved seed with a cultural regime recommendation. If so, the interaction effects should remain in the numerator.

In a closely related sense, the economic conditions in which the forests must have adaptive value simply represent another class of environments or performance requirements. However, it may be easier to project estimated economic values according to models of forest uses than to predict environmental variations and frequencies. Multiple traits that directly and indirectly influence some value parameters are easy enough to measure, but breeders need to know the correlations between such traits and between juvenile and mature tree characteristics. Hence, multivariate analyses of genetic variances and covariances should be planned along with measurements of performance in multiple ecological environments.

THE DENOMINATOR

The denominator of the gain heritability is the variance of the estimated mean values for the selection unit. Under mass, or simple recurrent selection, the individuals are usually assumed to be randomly located with respect to all environmental factors; hence the variance among individual units is simply the sum of all contributing variance components, genetic and nongenetic alike. The sampling errors are both genetic and nongenetic and can be reduced by extensive sampling to properly rate a selection unit with respect to other units. Since trees can often be replicated in plots and over environments, several components can be recognized. We will usually assume that gross macrosite effects can be recognized and variations in these adjusted for before estimating relative values. If adjustments cannot be made or can be made only with some error, the error variance of the adjustment or lack of it would have to be included (σ_E^2). In addition, any interactions of macrosite effects and genotypes (σ_{GE}^2) would contribute to the variance among units. Both of these components would be reduced by a good sampling of several environments and, if e environments were randomly sampled, would contribute:

$$\frac{\sigma_E^2 + \sigma_{GE}^2}{e}$$

to the variance among selection unit means. In replicated tree

plots, the error variation due to families not behaving the same way in all replications within macrosites is a plot error variance (σ_p^2), which can be reduced by the number of replicates in each macrosite sampled (r) and the number of macrosites $\left(\frac{\sigma_p^2}{re}\right)$. An additional source of variation among units lies in microsite or otherwise uncontrolled measurement or sampling error among trees even in the same plot (σ_w^2). This remaining error and the residual genetic variation among trees within family plots (σ_{gw}^2) can be reduced by the number of trees per plot (n) and hence would affect the phenotypic variance by:

$$\frac{\sigma_w^2+\sigma_{gw}^2}{nre}.$$

A small portion of the within-family genetic variance ($1/n$) is carried in σ_p^2. The final commonly designated source of variance among units is the variance component due to genetic differences among the units themselves (σ_g^2). Together with σ_{gw}^2 and the small portion of the plot error due to genetic sampling, the total genetic variance is approximately:

$$\sigma_G^2 \approx \sigma_g^2+\sigma_{gw}^2.$$

Thus, for single-tree plots, unreplicated but with adjustments made for macrosite variations, the phenotypic variance is:

$$\sigma_T^2=\sigma_w^2+\sigma_{gw}^2+\sigma_p^2+\sigma_{GE}^2+\sigma_g^2 \text{ or } \sigma_w^2+\sigma_p^2+\sigma_{GE}^2+\sigma_G^2.$$

If these are n families per plot, with r blocks or e macrosites, the variance among family means that should be entered as the denominator of the regression gain heritability is:

$$\sigma_{TF}^2=\frac{\sigma_w^2+\sigma_{gw}^2}{nre}+\frac{\sigma_p^2}{re}+\sigma_g^2.$$

Optimum allocations of n, r, and e, for different cost constraints can be developed to maximize efficiency or minimize the variance for given costs of establishing "n" trees, in "r" reps, in "e" sites. However, the problem of estimating the variance of the heritability regression coefficient can be more complicated than the design considerations reviewed in chapter 8 since both numerator and denominator are estimated and we require the variance of the ratio before optimum designs can be defined. Some simpler designs such as the parent-offspring regressions discussed in chapter 8 lead to some easily derived estimates of the error variance in heritability estimates and to easily computed optimum allocations of plant materials to efficiently estimate h^2 (Falconer 1960). Also, when the heritability can be easily constructed as an intra-class correlation or as a simple ratio of two mean squares, the distributions are well known (Hanson 1963) and optimum allocations of materials can be derived by standard calculus procedures. How-

ever, when variance components have to be estimated and several mean squares have to be used in combination to derive the h^2, the variance is more complicated to derive and optimum designs more difficult to design. Some combinations of numerator and denominator for different kinds of heritabilities are listed below.

The forest geneticist may design experiments to estimate the components in a variety of ways and, independently of such estimation experiments, may construct heritabilities appropriate for several types of selection and breeding programs. If the component estimation program is indeed independent of the breeding program, then designs can be constructed to efficiently estimate the components on the heritabilities, as discussed in chapter 1. Once such estimates are obtained, the geneticist may then compare the relative merits of different breeding systems according to their expected gains by constructing the selection differentials and heritabilities that apply to those systems. Some numerators and denominators for different heritabilities for some breeding programs are:

Numerators	Denominators
σ_A^2	σ_T^2
$3/4\sigma_A^2$	σ_{TP}^2
$1/2\sigma_A^2$	$\sigma_{TW}^2=\sigma_T^2-\sigma_D^2$
$1/4_A^2$	
$\sigma_A^2+\sigma_{AE}^2$	
$1/2\sigma_A^2+1/4\sigma_D^2$	

Appropriate types of heritability for certain selection methods are:

Type of heritability	Selection method
σ_A^2 / σ_T^2	Simple recurrent selection
$\sigma_A^2 / \sigma_{TP}^2$	Recurrent selection with clonal or selfed family testing
$3/4\sigma_A^2 / \sigma_{TW}^2$	Individual-tree selection within half-sib families
$1/2\sigma_A^2 / \sigma_T^2$	Mass selection without pollen parent control
$1/2\sigma_A^2 / \sigma_{TP}^2$	Mass selection without pollen parent control, with clonal testing
$1/2\sigma_A^2 / \sigma_{TW}^2$	Cumulative portion of recurrent, within full-sib family selection
$1/4\sigma_A^2 / \sigma_{TP}^2$	Half-sib family selection
$(1/2\sigma_A^2+1/4\sigma_D^2) / \sigma_{TP}^2$	Full-sib family selection

EXPECTED PROGRESS FROM SOME RECURRENT SELECTION AND BREEDING SYSTEMS

If the selection differential and heritability are well estimated, expected gains from various proposed tree-breeding methods can also be well estimated. However, variance in actual levels of either factor leads to variance among sample populations. Here, gains from just one generation of breeding are examined for various breeding and seed-orchard procedures for their average or expected results. Then, we consider breeding methods for repeatedly and cumulatively improving breeding population in particular mating patterns.

Simple mass selection is perhaps the easiest breeding method in forestry; phenotypes are selected according to their individual performance and open-pollinated with unselected pollen. In simple recurrent selection, selected parents are systematically crossed either in the forest or in orchards. In both methods, selection is based on the individual's own phenotype only, and both have been called mass selection in plant and forest breeding. If the pollen parents are unselected, the effective gene frequency of the favorable allele is halved and the gain is half of the usually computed mass selection:

$$\Delta G = s\frac{\frac{1}{2}\sigma_A^2}{\sigma_T^2}$$

This value is also essentially the regression of offspring on parents when the offspring were from unselected, or uncontrolled matings. The numerator covariance is that of offspring-parents, while the denominator is the variance among the parents which were used to estimate the value of their projected offspring.

If the original parents were saved and matings were among selected trees only, then the full narrow-sense heritability would apply:

$$\Delta G = s\frac{\sigma_A^2}{\sigma_T^2},$$

which is double the regression h^2 noted above. In this case, the cost of keeping the same s value increases by the amount required for controlled pollination, establishing orchards, etc., but seed collection costs can actually decrease due to seed production and harvesting efficiencies (Perry and Wang 1958). In some species, it is difficult to cross selected trees except in clonal orchards, and the cost of establishment is small enough that mass-selection clonal seed orchards are standard operations. Other species, however, may be easily crossed onsite or may have propagational problems too difficult or costly to bear in a clonal orchard program. If onsite crossing is not feasible, a form of selfing or other family selection may be required. The cost of seed may actually depend more on vagaries of crop size, species, site, weather, and use of

mechanized equipment, than on manner of crossing or control of pollen parent.

One further procedure similar to mass selection in its full utilization of the additive genetic variance is selection on the basis of clonal performance, as recommended by Libby (1964). The great advantage that vegetative propagule testing has is that for traits with low h^2, replications can reduce σ_{TP}^2 well below σ_T^2 and hence can increase the heritability that can be translated into gain. However, the method requires that the performance of vegetative propagules in tests accurately reflect that of seedlings or ramets as used in commercial production, or at least that the covariance between genotypic value and performance be close to the full genetic variance. Topophysis, age effects, rooting or graft compatibility variation, etc., can all reduce the covariance between performance and breeding value and hence reduce the value of the numerator σ_A^2. In addition, testing costs are high and should be offset by later uses such as for seed-orchard materials. Time delays in generating the new breeding population can also be costly, and any reduction of numbers of genotypes which can be examined by these methods may cause the s to be severely reduced, decreasing the value of the breeding effort. Nevertheless, for species which can be easily and cheaply propagated by cuttings, apomictic seed, etc., and which perform without much c or special and biasing clonal effects, the advantages can be great. The covariance between observed performance and breeding value is not a true genetic variance for clones or different aged materials that are either more juvenile or more senile than desired. Gain is made on the basis of the correlated response of tree value on measured trait, as outlined in chapter 4.

In addition to these forms of mass selection, various types of progeny test selections have been employed, as described by Namkoong and others (1966). In these programs, an initial selection on the basis of individual performance is made, as in mass selection, but more parents are selected than are actually to be used in the final breeding population. A second selection is made on the basis of further family or clonal tests to reduce the population to the minimal N_e desired. In clonal seed-orchard programs, such as described by Zobel and McElwee (1964), the trees originally selected in the first phase are crossed to some designated tester trees, or often, a set of heavy pollen producers serves as pollinators for the other parents. The families thus produced are evaluated in field tests and estimates of the combining ability of the parents are made. The trees which prove to be poor parents are culled from the orchard, or a new orchard is established with ramets only from the best parents. From these reconstructed orchards, the seedlings of the reselected parents can be generally considered as half-sibs of the seedlings produced for the testing, though a slight bias exists if any of the testers are also selected. Therefore, there is a half-sib relationship between the families tested and the seedlings from the reconstructed orchard, and the appropriate

covariance is that of half-sibs. When both parents have been selected, the effect of the selection differential is doubled. Hence, the gain from the second culling is:

$$\Delta G_2 = 2s_2 \cdot \frac{\frac{1}{4}\sigma_{A'}^2}{\sigma_{TP}^2},$$

where σ_{TP}^2 is composed of:

$$\frac{\sigma_w^2 + \sigma_{gw}^2}{nre} + \frac{\sigma_p^2}{re} + \frac{\sigma_{GE}^2}{e} + \frac{\sigma_{mf}^2}{m} + \sigma_y^2.$$

σ_{mf}^2 is the male × female interaction variance and m is the number of male testers. It should be noted that the genetic contribution to the error is somewhat differently allocated between this factorial design and either the hierarchical or diallel designs and is somewhat smaller for the factorial than either of the others. However, the difference is negligible if m is over 3 or 4 in any of the designs. $\sigma_{A'}^2$ is the additive genetic variance in the second-stage materials. Since no genetic recombination occurs between the two selection stages, $\sigma_{A'}^2$ will be less than σ_A^2 by an amount proportional to the initial heritability and initial selection intensity, both of which tend to reduce the amount of genetic variation among the initially selected trees. The reduction in $\sigma_{A'}^2$ is tabulated by Finney (1956) for various levels of initial σ_A^2 / σ_T^2 and initial selection intensity.

The advantage of this method lies in the gains which may be achieved in the second stage for traits of low-mass selection heritability but which can be tested to greatly improve the half-sib heritability. The method is especially useful for traits such as early growth, response to soil fertility or spacing, and some pathogen resistances that may have very low heritability and may be weakly correlated with other breeding values. Testing in controlled environments may then give substantial heritability since σ_{TP}^2 may be very much smaller than σ_T^2 such that the only reasonable gain possible may be in the second stage on those traits. Direct costs and time costs of such testing, however, are substantial. In addition, the costs may inhibit the number of entries accepted into the testing stage so severely that only a very small second-stage selection differential can be afforded. Such a limitation would clearly destroy the value of the testing, because gain is proportional to the product of s and h^2. Thus, only if the differences in h^2 are large enough to offset the cost of the replicated testing can the advantages of second-stage testing be utilized.

Since selections in the two stages are based on performance data which would be correlated to the extent that genetic effects control the phenotypes, it is clear that high heritabilities would mean a high correlation in the performance data between stages 1 and 2. The correlation is increased even further since the stage 2 selection will be based on both initial and replicated-test performance. If the additional testing is needed because of initially low h^2, however, the data will be less well correlated and the additional gain more significant. By examining a wide range of initial h^2

and post-test h^2, it was found by minimization under constraints that additional testing costs were offset only when initial h^2 was less than 3 percent (Namkoong 1970a). Using a linear programming analysis, van Buijtenen and Saitta (1972) similarly found that progeny testing can often be wasteful. When initial gains appear difficult to make, more careful initial examination and selection should generally be attempted first since even heritabilities of 5 percent will usually be enough to make some gain more quickly and cheaply than progeny testing.

There are some operational advantages to the progeny testing which some breeding programs are using. For commercial seed production, the parents can be culled or selectively mated as soon as reliable data begin to indicate quality differences. Progeny testing can give such data relatively early, and elaborate test designs may not be necessary. Such advantages in the commercial value of the seed can be applied in each generation, regardless of the size of the breeding population. In addition, it is possible to begin intercrossing among a wider sample of potential parents, as may be present before progeny testing in an orchard, to generate the breed population. Then, when progeny test data become available, crosses with the culled parents can be discarded from the breeding program while intercrosses among the tested and selected parents are saved. This step is costly and may reduce the progeny population of the next generation which can be carried because of the wasted efforts on crossing among culled parents. Without it, however, the benefits of the progeny test cannot be incorporated into the breeding population until after testing has identified the proper parents, those parents have been intercrossed, and the seedlings have matured. Such delays themselves are costly, and unless the cost of extra crosses and seedlings is worthwhile or the time interval for testing and producing a new generation is small, the breeding population will develop at a faster rate without progeny testing.

One other advantage of progeny testing may exist when selection is made for traits more readily observable or with higher heritability in progeny than in older parents. Thus, traits like rapid early growth may not be observable in older parents, and hence selection in parents is only for a correlated trait, whereas in juvenile progeny, the genetic variance itself is useful (Snyder 1969).

In future generations, both parental selection and progeny test efficiencies are likely to increase. Initial heritabilities will be higher since the material will usually be grown in better-known environments and more measurements will be accurately taken. Testing will likely be easier and better done and at earlier ages, but some traits cannot now be improved without detailed or complicated tests. Some agencies will have other uses for the tests which may be done quickly and cheaply enough to justify the large post-progeny test selection differentials required to achieve reasonable gains. There appears to be little difference in test estimation ef-

ficiency among the mating designs used as long as the number of crosses per entry and the seedlings per cross are reasonably high.

A slightly different form of two-stage selection, using seedling propagules instead of clonal regeneration of the initially selected genotypes, has also been extensively used in tree breeding (Goddard and Brown 1961; Wright 1964a; Stern and Hattemer 1964). In these seedling seed-orchard methods, the initial selections may not be easily propagated other than by seed, and it may be relatively easy to induce early flowering on the seedlings such that seed production from the seedling orchard is at least as good as from any other materials. The seedlings from the initially selected forest trees serve as their own test performance material.

The heritabilities and gains for initial selection are the same as in any simple recurrent selection system except that the male pollen may be unselected. If the initial selections are intercrossed, then the full-mass-selection gain is achieved, but if open-pollinated families are used, the gain is halved. Then crosses among selected seedling family members instead of clones produce the commercial seed.

In the second stage, which follows the initial selection, if many unselected pollens or open-pollinated seeds are used in the orchard-test plantation and then families are selected on the basis of average family performance, gain must be computed from half-sib family covariances, and the variance of the family means is:

$$\frac{\sigma_w^2 + \sigma_{gw}^2}{nre} + \frac{\sigma_p^2}{re} + \frac{\sigma_{GE}^2}{e} + \sigma_g^2,$$

where m, the number of males, is assumed very large. The covariance numerator of the gain h^2 between the commercial and test material is simply the genetic variance among test units, which is the covariance of half-sibs. The half-sib families are the units of test and selection and are to be mated among similarly selected families as in half-sib family selection. As noted by R. D. Burdon (personal communication), the variance among these units is reduced by the initial selections in the same way as for clonal selection. Hence, the second-stage gain or family selection is:

$$(s_2)\frac{\frac{1}{4}\sigma_A^2}{\sigma_{PT}^2},$$

where s_2 is limited by the number of families brought into the family trial and by the number of different families allowed to pass into the breeding populations.

Since the construction of families would also permit selection among individual family members on the basis of their own performance, an additional selection gain is possible in a form of mass selection within families. If selection is made in this tandem fashion—first families, then individuals within families—the gain

in the last stage of individual selection is:

$$(s_3)\frac{\frac{3}{4}\sigma_A^2}{\sigma_T^2},$$

where (s_3) is the differential for selection among family members, $\frac{3}{4}\sigma_A^2$ is the genetic variance within half-sib families, and σ_T^2 is the variance of individuals within families. σ_A^2 should ordinarily be regenerated by matings and hence should be close to the original σ_A^2.

Alternatively, if the families entered the orchard as unrelated full-sibs (or if a full-sib family selection was made), then the variance among families would be $\frac{1}{2}\sigma_A^2$ plus the dominance and epistatic variances, while the useful variance within families would be $\frac{1}{2}\sigma_A^2$ plus the remaining nonadditive genetic variances. Hence, the gain from alternative procedures will vary according to the selection differentials and control of crossing exercised and the relative sizes of the heritabilities. Using a single pollen source would create excessive inbreeding in one generation, but it would generate $\frac{1}{4}\sigma_A^2$ among the full-sib families within the single half-sib family, and $\frac{1}{2}\sigma_A^2$ among the individuals within families.

If the operational advantages of secondary selection and propagation do not dictate the choice of intermediate selection stages, the advantage of the seedling seed-orchard test combinations is in the additional selections that can be made not only among the original families but also among individuals within families. By simply using enough seedlings within each family, the last selection can have a large s_3, and while h^2 is on an individual tree basis, σ_T^2 may be somewhat reduced by the easier environmental control and more detailed observations possible within experimental conditions. While both the clonal and seedling seed orchards share the common problem of having to balance s_2 and σ_{PT} for maximum gain, it seems that in individual selections, s_3 and σ_T^2 can more easily be balanced by reasonably large family sizes. It is also likely for individuals to be selected on the basis of an index of information on family as well as individual performance (Namkoong 1966b). In such cases, it is feasible to construct testing and breeding replicate blocks that contain only one or very few family members at sexual maturity. Then the combined individual and family selections can be made and the expected gain approximated from the average s values per block, or by an s for the index selection. The chances of rapid inbreeding are enhanced by heavy family selection. However, if pollination is controlled and potential inbreeding and spacing are not otherwise serious problems, the gain from individual within-family selection can be emphasized. It can be increased by having many trees per family, and the h^2 for that phase can be maximized by using many seedlings per plot if the plot error is high or many plots otherwise.

The method described therefore provides a two-generation sequence of selection in one step. Substantial quick gains can be had if a large selection differential is generated (Namkoong and

others 1966). Unlike for clonal seed orchards, optimal allocation of selection intensities among stages yields relatively equal selection proportions between the stages of selection for reasonable costs. However, special handling and care are required to combine the objectives of testing and eventual seed production, and more time or cost is required to generate a completely new recurrent selection generation than in other methods. The additional time is required for crossing and establishing a seedling generation. An alternative is to make crosses for the next generation earlier than culling allows and hence to make unnecessary crosses which are subsequently omitted from breeding. However, in addition to the difficulties of establishing seedling seed orchards, the existence of large genotype-by-environment interactions may require that seed-production techniques be postponed until testing is finished. If the interaction is large, poor families in the seed orchard may be genotypes which should be picked for propagation. If the advantages of substantial improvements in h^2 and the additional selection differentials in the seedling generation warrant it, however, the method merits the work needed to overcome the experimental problems of simultaneous testing and seed production. In particular, the use of clonal replicates of seedling entries could substantially improve individual seedling selection heritabilities and make gains on that basis very strong (Libby 1969).

The time and effort of progeny testing in either clonal or seedling seed orchards can clearly be substantial, but with experience and data on juvenile-mature correlations, it should be far easier to handle in future generations. As the breeder develops the capacity to evaluate more juvenile materials, the value of progeny testing increases (Nanson 1967), and the main limitation is to induce sexual reproduction in juvenile stages without mitigating the value of the tests. Clearly, one means would be to have different clonal replicates for testing and for reproduction to treat each ramet appropriately for its purposes and to develop rapid testing and reproductive cycles.

Similar forms of selection among families can be generated from single pair matings instead of the half-sib forms of the intercrossing outlined above (Libby 1969). After the initial selections, possibly at somewhat lower intensities, crosses are made among them, and full-sib family identities maintained in test crosses. Then several optional systems may be followed. The best full-sibs can be identified, and the specific good combinations selected for reproduction by repeated crossing of the same selected parental pairs in special, limited combination orchards. The gain due to additive genetic action is similar to that of the progeny-tested clonal orchard if the s factors are equivalent, but a gain of:

$$s\frac{\frac{1}{4}\sigma_D^2}{\sigma_{PT}^2}$$

due to dominance can be added. This gain, due to dominance, is not cumulative if the next generation will be created through recur-

rent selection of completely intercrossed trees, but it can be at least partially cumulative if the parents are selfed or otherwise regenerated and new selections are based on specific cross-test performance (Namkoong and others 1966). Subsequent generation gain can then be based only on the genetic variances generated among individuals within parental lines.

Selecting within the full-sib families for advanced generation crossing provides the same advantages and problems as when the families were generated as half-sibs, except that the family selection gain is on the basis of:

$$\frac{\frac{1}{2}\sigma_A^2}{\sigma_{PT}^2},$$

while the within-family individual-tree selection gain is on the basis of:

$$\frac{\frac{1}{2}\sigma_A^2}{\sigma_T^2}.$$

Clearly, the benefit of this method over other methods of crossing and selecting depends on the allocation of the selection intensities and on the sizes of σ_A^2, σ_{PT}^2, and σ_T^2 (Squillace 1973). For example, we can contrast selecting, say, 400 trees and making 200 pair crosses of 1,000 seedlings each with making a partial diallel in some partially blocked design such that 5 crosses per entry produce 200 seedlings per cross for the same number of seedlings. Further, supposing that in the first case we pick the best 100 crosses (1:2) and the best single tree in each, and that $\sigma_A^2=1$, $\sigma_T^2=20$, $\sigma_{PT}^2=2$, the additive genetic gain from the progeny test stage is:

$$\begin{aligned}\Delta G_{FS} &= i_{FS}\ (\tfrac{1}{2})\ \sigma_A^2 / \sigma_{PT} + i_I\ (\tfrac{1}{2})\ \sigma_A^2 / \sigma_T \\ &= 0.80\ (0.5) / 1.41 + 3.37\ (0.5) / 4.47 \\ &= 0.283 + 0.377 = 0.66.\end{aligned}$$

Also suppose for the second case that the best 100 entries were picked (1:4) and the best full-sib (1:5) family was picked from those, and the best individual from them chosen (1:200), and that $\sigma_A^2=1$, $\sigma_T^2=20$, $\sigma^2_{PT(HS)}=2$, $\sigma^2_{PT(FS)}=3$. Then the gain is:

$$\begin{aligned}\Delta G_{DIAL} &= i_{HS}\ (\tfrac{1}{4})\ \sigma_A^2 / \sigma_{PT(HS)} + i_{FS}\ (\tfrac{1}{4})\ \sigma_A^2 / \sigma_{PT(FS)} \\ &\quad + i_I\ (\tfrac{1}{2})\ \sigma_A^2 / \sigma_T \\ &= 1.27\ (0.25) / 1.41 + 1.40\ (0.25) / 1.73 \\ &\quad + 2.89\ (0.5) / 4.47 \\ &= 0.225 + 0.202 + 0.323 = 0.750.\end{aligned}$$

However, if $\sigma_T^2=10$ instead of 20, then $\Delta G_{FS}=0.816$ and $\Delta G_{DIAL}=0.884$. On the other hand if we selected down to a population of 50, then for $\sigma_T^2=10$, $\Delta G_{FS}=0.983$ and $\Delta G_{DIAL}=0.952$. If the half-

sib selections cause inbreeding, these reductions would make the full-sib system more attractive (Squillace 1973). Inbreeding, however, is induced in the full-sib phases of both systems.

In mating patterns for either progeny test selection or breed-population generation, partially balanced designs will often be necessary (Snyder 1966). Thus, for diallels any of the partial and blocked partial designs described for estimation of variance components may be used. For selection among blocks, check entries or overlapping blocked subsets can be easily installed. The reduction of error variance in testing among genotypic means by the use of subblocks can be especially valuable in forestry. However, if a choice exists between using varietal checks to allow for interblock selection and increasing the selection differential by allowing more entries in the test, greater expected gains will generally favor the inclusion of more entries.

The various forms of recurrent selection for general combining ability, or in the one special case of full-sib selection for specific combining ability, are reviewed by Namkoong and others (1966). They find that operational and time costs can significantly affect the choice of breeding method, because the forms dictate different lengths of breeding cycles and the gains due to various selection stages occur at different times. Fairly complicated considerations of the relationship between the selection intensity and its effect on the numbers of entries, and hence on h^2, also make it difficult to generate any general statements on choice of method. van Buijtenen and Saitta (1972) concluded for their conditions that heavy progeny testing can be justified only if its primary use is for developing advanced breeding generations.

When it is possible to produce clones instead of seeds for commercial reforestation, additional gains can be achieved from nonadditive genetic variations though these gains are not generally cumulative. The basic breeding population is expected to develop mainly from recurrent selections and general crossing among all selected parents. Only for specific, short-run breeding programs would groups, families, or individual lines be inbred to cumulatively utilize nonadditive gene actions. However, even when a breeding population is improved by simple recurrent selection system, specific clones can be picked and their peculiar gene combinations used within each generation even though that gene combination may be superseded in the developing breed. Thus, for example, the breeder may use the above diallel-crossing procedure and can expect to accumulate gain in the breeding population as computed above. However, in each generation additional gain can be achieved by selecting for improvements due to the nonadditive genetic variance among the selection units. While the selection intensities and the phenotypic variance denominators would remain the same, the genetic variance in the numerator would be increased. In half-sib family selection, the additive-by-additive epistasis and other higher order epistatic variances would be added. In full-sib family selection, $\frac{1}{4}$ of the dominance and

additive-by-additive epistasis, 1/8 of the additive-by-dominance epistasis, 1/16 of the dominance-by-dominance epistasis, etc., would be added. Finally, in individual-within-family selection, 3/4 of the dominance and all remaining epistatic variances within full-sib families would be added. If the nonadditive genetic variances are substantial, the one-generation gains can also be substantial. Gains in future generations would also be substantial but would have to start on the basis of the cumulative gain achieved in the basic reference breeding population.

MATING PATTERNS

Regardless of the method for selecting parents for the next generation and for any progeny testing, the parents may be further used in two distinct ways. The commercial product may be generated from a subset of those parents, all of them, or a wider sample of genotypes than will be retained in the breeding population. On the one hand, a single pair may be chosen to produce all of the commercial seed desired while a separate population is bred by intercrossing among many trees for future selections and breeding. On the other hand, it may be difficult to obtain the seed required even from all of the selected parents, and hence the commercial seed-production orchards may include trees which would have been culled for breeding purposes.

We distinguish between the seed-production and breeding-production operation, but they may sometimes be the same, as in mass selection. In general, however, seed production can be separated, and such separation is generally desirable if the seed product is noninbred while the breeding population may be inbred. Hence, actual commercial production, such as with single full-sib families or with clonal collections in "synthetic multiclonal hybrid varieties" (Schreiner 1968), is considered as an alternative only in its selection and breeding phases and not in production phases of the material released for commercial propagation.

Since crossing can usually be done immediately before commercial seed production, the only limitations on making many crosses to generate a breeding population are the costs in time and effort. The sizes, times, and designs of these crosses are critical for breeding advance. Estimation experiments and test-cross designs may be required to yield data as early as possible. Factorial and diallel designs may be adequate for estimation, and little distinction can be made among them for testing purposes.

Designs must be examined, however, for their efficiency in developing a breeding population from some sets of parents. The major criteria are maintaining large, effective population sizes and achieving rapid selection advance in the base populations. Many breeders will want to combine at least some of their estimation and test designs with breed-population production and sometimes also with their seed-production operation. At this time, we shall consider crossing designs only for purposes of generating

advanced breeding populations and shall generally assume that some limitations on time and effort exist.

The easiest operation, of course, would be akin to a mass or simple recurrent selection system in which only the selected parents are allowed to randomly mate, all seed is used commercially, and the entire production of seedlings then forms the next population for a new generation of selection. However, while the method is easy, some crosses will be very heavily represented in the subsequent generation, while others may be absent. Not only will the expected level of inbreeding increase with increasing departures from uniform representation, but stochastic variations among trials may be large enough to create unacceptable risks that a particular breeding population will either lose favorable alleles, suffer excessive inbreeding depression, or both. Since any recurrent selection program will eventually accumulate high inbreeding, outside sources of genotypic variants may have to be periodically infused. In some systems, materials from outside ancestry are expected to be continually available and can be tested against the breed population. The best of the new introductions can be used with profit if some of the population proves less valuable than the new (Burrows 1967). Introducing new materials would entail some loss in gain of favorably fixed or other high-frequency loci, and this loss can be considerable as generations advance. As the mean is cumulatively improved and more forests are established from the select breed, it becomes less likely that such materials will be found useful. Therefore, in plans for developing the main breed, many generations should be selectively advanced without such recourse. Other techniques, such as replicating breeding populations, should be used to postpone the crisis. Therefore, controlled crossing programs can be useful if plans are made to incorporate all useful alleles in the base breeding population and effective population sizes are kept large enough that they may be expected to advance.

In simple recurrent selection systems, all selected genotypes are completely intercrossed, and the seed is composited for the next generation. That generation is later reduced by selection to about the same number of parents as previously chosen for a new generation of intercrossing. However, through controlled intercrossing and compositing, ancestral controls can be imposed. Various degrees of control are possible. At one extreme, bulking pollen from all male genotypes and bulking all seeds are almost like mass selection. At the other extreme are keeping female parent identities on seed lots and making identifiable crosses with specific males. Since such care is more expensive, fewer individuals may be available for selection and hence the selection differential may be reduced. Within these limitations, various forms of complete crossing and seedling identification have been proposed which are short of complete intercrossing and complete control.

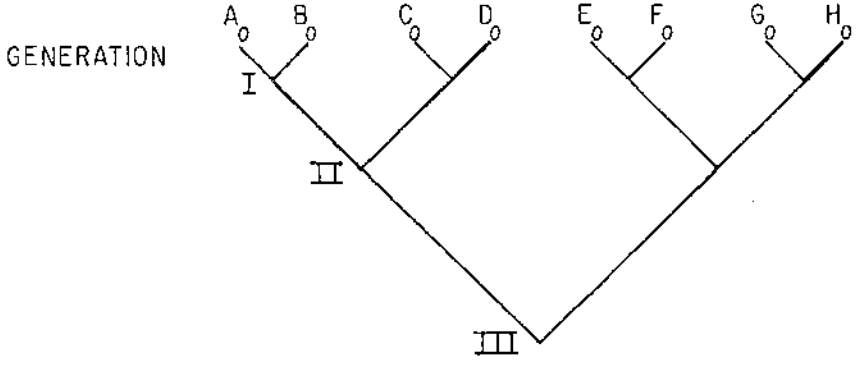
With completely controlled crossing and seedling identification, pairwise mating systems which allow some inbreeding but less than random pairwise matings have also been proposed. Regular mating systems which can be continuously followed in the whole population and repeated every generation offer some insights into how relatedness and inbreeding develop. In such regular systems as have been studied, selections are generally assumed to have been made within families, and the number of selected parents remains constant over generations. Thus, the selection differential is governed by the number of seedlings generated per cross, and family information is not used in selection except for ancestral control.

We can thus examine a variety of mating systems according to the manner in which ancestral control is completely, partially, or not at all maintained. Among the systems with complete control of mating and coancestry, differences exist in the numbers of families generated from each selected parent of the breeding population. The fewer the crosses or families made per parent, the less chance for family selection, but, presumably, the greater the number of individuals to select from within each family.

RECURRENT MATING SYSTEMS WITHOUT FAMILY SELECTION

Complete avoidance of inbreeding by mating only single pairs from unrelated lines has been recommended at least as a temporary measure for forest trees. If complete control is maintained, however, this system requires that the selected population shrink by at least half in each generation or that it be mated in carefully controlled patterns.

Thus, forgoing any interfamily selections, the proposed matings would be:



Family selection at any stage would, of course, more rapidly lead to the final necessity of crossing among individuals of a single

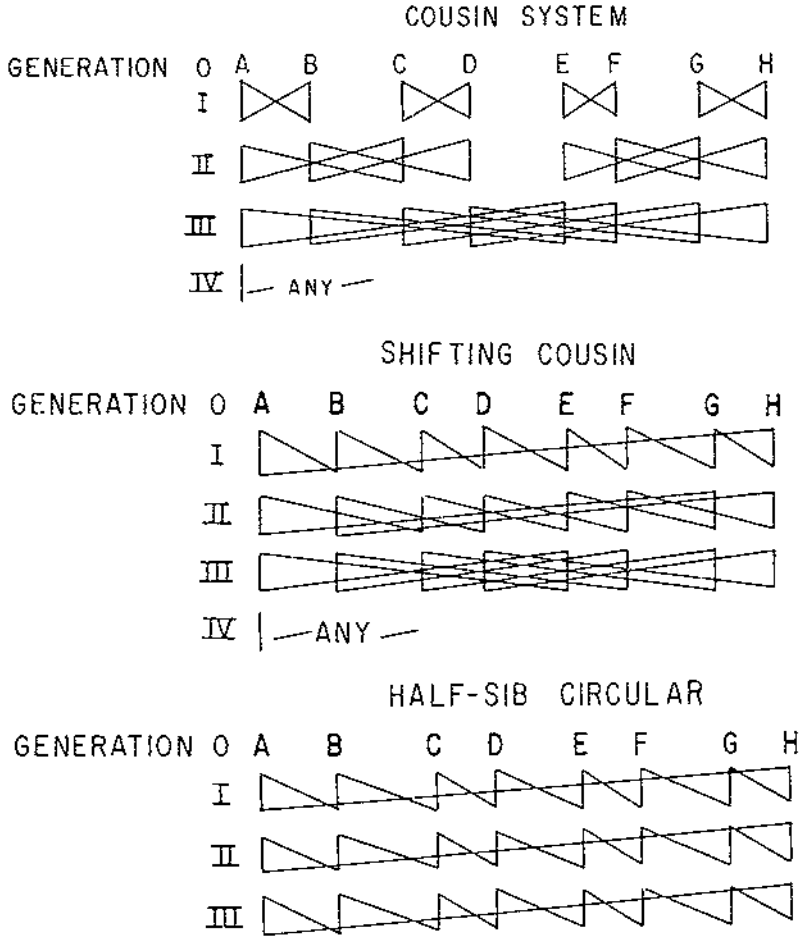
full-sib family. A replicate population from the same original parents would permit cousin matings at lower inbreeding levels, which are discussed below.

Replicate populations, however, do not add to the number of generations which may be developed from a given set of genotypes without any inbreeding. Such systems contain no early inbreeding but greatly increase the common coancestry of breeding population genotypes and lead to the accumulation of very small inbreeding for a few generations, followed by large increases in inbreeding (F) when avoidance is no longer possible. Thus, extreme early avoidance of inbreeding with an initial population of n individuals can be followed without inbreeding for k generations if $n=2^k$, but must thereafter involve high F. Therefore, such extreme systems may be followed to temporarily maintain a given low level of inbreeding but, in recurrent systems, may lead to higher inbreeding in the longer run (Cockerham 1970).

Such early avoidance systems can be designed to permit only mating of distant cousins after matings of unrelated pairs are impossible. Such systems rapidly build up the average coancestry among trees while avoiding inbreeding.[1] However, they do accumulate coancestry more rapidly and for small parental population sizes (around 8) do have higher inbreeding than random pairs after 24 generations.

On the other hand, a regular system of circular half-sib matings, as described by Kimura and Crow (1963), maintains the same number of families in each generation if no family selection is permitted but will immediately lead to higher inbreeding but a slower increase in coancestry. Therefore, the high initial inbreeding is thereafter accumulated more slowly, and with $N=4$, the inbreeding becomes less than the cousin system of mating by the 15th generation at $F=0.68$. With $N=8$, however, it takes until the 35th generation ($F=0.7$) and with $N=16$, the 95th generation ($F=0.78$). Thus, the relationship between early avoidance and eventual inbreeding is clearly an inverse one, but one which takes many generations for the lower rates of increase of the coancestries to overcome the initial levels of inbreeding, and this only occurs at quite high F values. Thus, early avoidance, cousin systems of matings may be quite feasible in forestry. However, regular systems may be required to assure that the pairings are made each generation in the desired patterns and that all families contribute equally to the new generation. For example, in the following diagrams of eight parents per generation with one of several hundred individuals from each mating selected, the cousin systems eliminate inbreeding entirely until the fourth generation, whereas the circular half-sib system initiates inbreeding in the third generation:

[1] Cockerham, C. C. 1969. Notes on quantitative genetics. Unpublished lecture notes, Sect. 5. N.C. State Univ., Raleigh, 29 p.



For any regular systems of inbreeding in which linear recursion relations can be established for coefficients of inbreeding or relationships, asymptotic results can be easily determined by an analysis of the roots of the recursion equation matrix (Crow and Kimura 1970), as shown in chapter 9, for mating frequency recursion equations. Inbreeding coefficients for general patterns, however, may be computed by machine (Cruden 1949).

Other patterns of pairwise mating can be formed by grouping subsets of parents in less rigid hierarchies and mating among groups when within-group inbreeding exceeds predetermined levels. Patterns such as proposed by Aalders (1966) have some merit in compromising between complete avoidance of inbreeding and minimizing coancestry, and they may be easily handled in field operations. However, for current tree breeding with large population sizes, an early avoidance system may be most practical.

In the early stages of breeding, new genotypes can be infused into the populations without much loss in value. Early avoidance systems can give very low inbreeding for at least a dozen generations. For example, a system can be designed to completely avoid early inbreeding with separate replicates of full-sibs in the initial generations. Distant cousin relationships are formed among trees in the replicate sets. When the inbreeding potential within sets becomes excessive, crossing of sets will cause little inbreeding immediately but more rapid increments thereafter. It is then always possible to change mating patterns, but the effect of sequences of patterns may not be advantageous.

RECURRENT MATING SYSTEMS WITH FAMILY SELECTION

Mating and coancestry patterns become far more complicated when multiple crosses are made among the parents selected to regenerate the breeding population. In early generations, family selection is liable to be high as major genes are sorted out and inexperience with useful selection takes its toll. Family selection may also be popular for intensive, short-run breeding. However, the benefits of selection among families must be balanced against the cost of making many families that are not selected. The cost must also include any loss in the within-family selection intensity, which is reduced if limited land and funds are spent on creating many families. The latter loss may not be too debilitating since neither selection intensity nor additional experimental cost is linearly proportional to numbers of families. But a solution that optimizes gain by balancing selection among and within families is desirable. As previously shown for seedling seed orchards or for breeding populations, the balance depends on heritabilities and selection differentials. We can affect the family selection differentials by creating more or less different kinds of families.

The breeder's options for multiple crossing among the parents of the developing breed extend from single pair matings and their reciprocal full-sibs to a full diallel of all possible $\frac{n(n-1)}{2}$ crosses and reciprocal full-sibs plus selfs. Crossing patterns that lead to immediate and extensive inbreeding should be avoided. It is clearly better to choose patterns that reduce early inbreeding.

As pointed out by Libby (1969), the hierarchal mating design that may be useful for other purposes holds no particular advantage over single pair matings for breeding population development, even if inbreeding problems are ignored. Other crossing patterns such as partial, disconnected, or disconnected partial diallels, on the other hand, make possible alternate sets of pairwise matings. The breeder can select among sets to take advantage of variations, both within and among families. In such recurrent selection systems, which develop breeds on the basis of their

intrapopulation general combining ability, little advantage can be taken of dominance genetic effects expressed in specific combining abilities, but selection for additive gene effects can be effective. Where multiple crosses are made, full- and half-sib families can be formed in early generations, and more complex cousin groups can be formed in advanced generations. Some original families may be disproportionately represented in succeeding generations if selection of individuals is based on family and individual performances. Crossing patterns that include some inbreeding may be the result. There is no need to produce exactly the same numbers of families as parents, but fewer families than parents rapidly lead to high coancestries. Any such mixed patterns require the careful tracing of coancestries to minimize inbreeding and the concomitant loss of genetic variance. If coancestry records are maintained, high levels of inbreeding in the breed population can be reduced for commercial seed production by crossing unrelated trees in seed orchards. In addition, the seed production crosses may be designed to utilize nonadditive and otherwise noncumulative gene effects in specific crosses. Partially controlled blocked diallels (Braaten 1965) or blocked factorials (Burdon and Shelbourne 1971) offer many more variations on partially subdivided mating patterns. However, the effective population should not be inadvertently reduced below the desired size by blocking of mating sets in a way that induces positive assortative mating. If controlled crossing systems are feasible for breeding populations and identification of parental sources can be maintained in commercial seed production, the selection differential for within-family selection is maximized by using all seed orchard products in selecting the next generation. When carefully controlled sites are required for accurate selection and controlled crossing with identification is not possible, it is impractical to reselect from the entire population. If costs of such special plantations are low enough and evaluations and seed production can be early enough, seedling seed orchards may have considerable advantages. However, any of the selection-breeding methods can be iteratively applied, and subdivision of the breeding population into breeding units greatly increases the possible variations in crossing patterns.

For breeding populations, control of coancestry is required even though we have very meager knowledge of the effects of inbreeding on both inbreeding depression and loss of genetic variance in selected tree populations. Indications thus far favor minimal inbreeding. In general, since breeding generations are so long and breeders and organizations in forestry do change, it is likely to be even more important for future generations that large populations and strict coancestry control be maintained (Namkoong 1971). It may sometimes be possible to breed with small numbers of parents and to tolerate high levels of inbreeding in some rapid selection and breeding systems. Methods to achieve quick gains with small numbers, such as 5 to 10 clone orchards

with mass or single recurrent selection, may represent viable short-term alternatives, and require experimental testing.

PARTIALLY CONTROLLED MATING SYSTEMS

Operational problems or costs may sometimes prevent a breeder from maintaining the identity of all parent ancestries. The costs of maintaining identities in field plots after controlled crosses can be large. In addition, early selection and early seed production can be forced, then the relative costs in time, effort, and selection differential of making such crosses can be high. However, the benefits of maintaining identities, or conversely, the risk of loss in variation and inbreeding depression through unperceived inbreeding, may well justify thorough identification. If partially controlled breeding programs can partially control such losses, then an optimum intermediate level of control may exist.

Most analyses of breeding systems are based on average levels of inbreeding or coancestry, and the results assume a linear relationship with cost or loss of value. While the relationship between mean inbreeding and loss due to dominance gene effects may be linear, a more meaningful loss function might relate variables in breeding method to risk of achieving homozygosity. While two methods may be similar in average inbreeding ($\overline{F}$) or expected heterozygosity ($1-\overline{F}$), one may generate homozygosity levels with less variation than another and hence may be judged to be a better method if the risk of high homozygosity is costly. Risk analysis is especially important if breeders cannot fully identify parents. Whether or not identification can be maintained, optimizing selection at the level of inbreeding that can be tolerated to achieve selection gain or the functional relationship between the two requires far more information than is now available on inbreeding effects. The critical need is for data on performance at F values below 0.25 (Burrows 1970).

When it is possible to identify at least seed parentage, then it may also be possible to partially control male parentage by using different sets of males for mass pollination each year. Then, identification of seeds by years would identify male sets, at least, and hence probabilities of parentage would be more closely determined. Alternatively, subsets of factorial mating designs could be segregated in which the male entries for each subset are pooled into pollen mixes specific for each subset of females. Variations on this polycross system are described by Burdon and Shelbourne (1971), wherein subsets are completely separated and no genotypes occur in two sets, or where genotypes may overlap among subsets with some males or females present in two or more subsets. Such designs can be varied according to the availability of pollen and female flowers. In these systems, as in controlled crossing, the dangers of rapidly reducing the effective population size by assortative mating should be recognized. While exact relationships are not known, the probabilities of selecting closely

related individuals in subsequent generations are increased by selecting among subsets and hence reducing the base population for further breed development.

When complete records are not feasible, it is profitable to at least identify the seed (♀) parentage of field stands. The cost of such records is minimal. Only nursery and seedling lot need be identified, and sets of females can be confounded with location and year of planting. The problem in analyzing the alternate systems for their cost-benefit functions again lies in determining the loss function at the expected levels of inbreeding. Since variation in homozygosity level is likely to be much higher than where identity is controlled, risk analysis may be a more valuable computation of expected loss.

It is instructive to regard average inbreeding level generated by selection among open-pollinated families as a key criterion. Even with open pollination and up to 1/6 family selection, Burrows (1970) finds that average inbreeding in the third generation still lies between 4 and 12 percent in a *Eucalyptus* seedling seed orchard. Empirical studies are needed on selection methods to determine if this method, simple mass selection, or any other systems can function as expected and with what variation among replicate trials.

It should be remembered that in a seed-production orchard, replacement of clones with advanced generation materials is a continuing operation. New seed-production ramets replace old ones as the benefits of replacement become clear. Some genotypes may remain in the orchard for several generations, because they continue to rank high genetically. Others are replaced as newer material proves better. In this situation, inbreeding must be controlled, because both parents and their offspring may be present. Testing and replacement can be programmed to gradually change the composition of genotypes making up the breed, and testing should include samples of generations other than the current one.

All the above programs use selection among individual trees within families to some extent. Since selection is most accurate among contiguous trees, the allocation of trees within plots and among plots, replicate blocks, or stands should heavily favor trees within plots. To the extent that family selection will be important, however, site replication will be important and may cause some reduction in the optimum number of trees per plot.

HYBRID BREEDING SYSTEMS

The foregoing discussion has been concerned with methods of improving purebred or recurrent-selection populations developed for general combining ability and using the cumulative effects of additive gene actions in single populations. While some methods discussed above may temporarily use any discovered specific combining abilities, they do not cumulatively develop lines or popula-

tions for crossing to obtain specific combining ability effects. For naturally cross-pollinated plants, however, inbreeding is aberrant, and debilities of inbreds may recur for many generations. Many lethals and sublethals found under intensive inbreeding destroy lines despite vigorous efforts to save them. Those lines or populations that do survive, do so with less vigor and size than the normally outcrossed varieties. All methods for developing single populations eventually increase inbreeding and coancestry; the methods discussed only affect the rate of inbreeding. In contrast, hybrid systems develop at least two lines or populations for crossing. These systems use the phenomenon of heterosis, which is often viewed as the direct opposite of inbreeding depression. Outcrossing restores vigor and reproductive fitness, and intervarietal and even interspecies crosses often exceed the development of both the midparent and the highest parent. If the gene action that produces the greater value of the hybrid is not heterosis or if heterozygosity is not necessary, then hybrid superiority may be due to additive gene actions on a combination of traits. In that case, it is usually far simpler to create a single base population of F_1 crosses and to improve that population as any other single population breed for the combined traits. Only if heterosis is useful, will it be generally advisable to enter hybrid programs.

Limits to cross compatibility are often wide enough to allow distant evolutionary relations to cross. Stephens (1961) has classified hybrid breeding programs according to the extent to which incompatibilities restrict the segregation of new genotypes. Within species, genetic divergence also may generate a quadratic response in vigor as more and more divergent sources are crossed. In corn, for example, heterosis is measured by the excess of the hybrid over the median parent or the F_2. In one study, heterosis rose as the varieties which were crossed increased in diversity of their origin up to a peak and declined as the diversity of origin apparently exceeded an optimal level (Moll and others 1962, 1965). Thus, at the varietal level of biological organization, variations in heterotic response may be a predictable quadratic response function of diversity. However, it is also possible for dominance levels among alleles within populations to be maximized by natural selection, and to diminish as more foreign alleles are paired.

At the species level of diversity, less distinct patterns of heterotic responses are visible, even in tree species in which considerable amounts of natural species crossing occurs. Information is confounded because species relationships are constructed partly on the basis of crossability, but there seems to be no strongly defined hierarchy of chromosomal or other incompatibilities (Wright 1962). Species that cross seem to do so without differences with respect to origin of parents. However, heritable differences in morphological traits exist and may sometimes show heterosis.

Selection for hybrid performance can mean selecting among inbreds for inbred-cross performance, among varieties for varietal hybrid performance, or among species combinations and sources within species for specific tree-by-tree crosses. Different degrees of inbreeding with respect to the outcrossing product are therefore tolerated to achieve the relatively outcrossed product, and variations among hybrid breeding methods exist in the purity or allelic homogeneity of the parental populations.

In general, selection units will vary within at least one of the parental populations, and selection is based on tested performance as potential parents of hybrids. Further genetic recombinations are created within the selected parental populations by some breeding method and a new cycle of selection for cross performance is instigated. Thus, cumulatively better hybrids are developed from base breeding populations selected to regenerate variations in hybrid performance. Hybrid systems resemble those for developing populations with high general combining ability in the sequence of developing genetically variable populations and reselecting parents for iterated cycles of improvement. Hybrid systems require designs for testing, selection, and intrapopulation mating to generate the parental population.

In hybrid breeding, production and testing seed are distinct from breed population development. Commercial seed production can similarly come from a subset of either or both parental populations, or a large set of parents if seed demand exceeds production capacity from selected parents. As for purebreds, the sex with larger gametic production will tend to have fewer parental entries in the commercial seed, but for hybrids a choice exists as to which population serves as male or female. While some mixture of sexual role is generally expected, optimum sexual functioning may require that the population be treated to maximize gametic production of the less prolific sex. Progeny testing with all of its attendant costs is always required, but with reasonable efforts towards juvenile testing and early reproduction, the costs can be mitigated as for the recurrent selection programs previously outlined.

Among the various ways to make the test crosses and to carry cross identifications into long-term field plots, the individual tree crosses are most expensive, but they afford greatest possibilities for selecting specific crosses and developing specific combining abilities among crossed parents. Any of the mating designs may be used for any level of hybridization, and since testing is distinct from breed population development, factorials or blocked factorials would not carry the inbreeding problems they do for the single-population recurrent selection programs. Specific crossing combinations may then be identified for use in special seed orchards, but unless those lines are identified and inbred for future selection for specific combining ability, that gain is not cumulative. The cumulative gain is therefore generally based on the cross-general combining ability among generally cross-

compatible populations. Specific combinations of cross parents may be developed in separate subpopulations, requiring that within the alternate populations specific line or subsets be developed which cumulatively cross well with their opposite numbers in the other population. These will not be followed in this discussion, though they represent a viable, short-run alternative in hybrid breeding.

After testing and selection, parental populations in hybrid programs must be further developed through a stage of recombination within the source populations. Thus, some degree of inbreeding will be present in the hybrid seed product when an established variety or seed source exists and cannot be improved, only the new population requires the selection and recombination phases. Since in forestry this is expected to rarely be the case, the following discussion will generally refer to improvement of both populations. If single crosses within an adapted variety are the products, then lines of inbred (for example, selfed) parents would be developed and retested for the next generation, and subsequent selection would be based on lines and individual within lines. If varieties or species are hybridized for the seed product, then each selected population would be intercrossed within themselves to regenerate allelic combinations for advanced generation hybrid development. If a general cross-performance is desired, intercrosses among the selected parents within the population are necessary. If specific single crosses between individual trees in the alternate populations are desired, a greater degree of inbreeding within lines within populations is required.

Inbreeding is useful in a breeding program for a cross-pollinated species only if the hybrid product is better or more uniform or otherwise more easily controlled than what could be developed by normal outcrossing procedures. If the development of pure lines is feasible, then hybrids may still be sought to improve traits affected by dominance or overdominance types of gene action.

HYBRIDS OF INBRED LINES

Selection of inbreds for crossbred performance requires either direct testing in hybrid combinations or a high correlation between inbred and hybrid performances. This correlation can be highly variable (Allard 1960; Allard and others 1966), and even though it may take a long time, direct testing is likely to be best. Special care of inbreds and the possibilities of doubling monoploids may rapidly create relatively homozygous lines for selection in a few years (Stettler and others 1970; Orr-Ewing 1965). In that event, simple selection among lines will be feasible to produce standard inbreds for hybrid seed production. Selection on inbred performance may take the form of mass selection in which the individual's own performance is the basis for selection, or, as in most plants, it may more often take the form of selection on family performance, including sibs as well as parents. Early generation selection for inbred performance in later generations

of crosses, however, is difficult and may take too long to be practical in forestry. Only if crosses can be developed each generation as the inbreds become more selectively homozygous, will long-term inbred development be worthwhile.

Testing the developing inbred lines in all combinations is clearly the most desirable way to select among lines for the best crossers or for the best single-cross combinations. Costs and other physical limitations, however, generally preclude such complete testing. Hence, something less than complete testing but more than selection on individual performance alone is attempted with other crops than trees. Topcrossing has served as both a testing procedure as well as a form of hybrid seed production in which the inbreds are crossed to a mixed source as the alternate parent. The topcross tester may be other standard inbreds, a standard heterogeneous variety, or any stable mixture of other materials which gives a mixed genotypic source against which the inbred's hybrid performance can be observed. In corn breeding, the tester has generally been a standard variety, but any set of lines can be used in testing as well as for seed production. If generation and testing times are short, topcrossing may also be used in preliminary screening for general combining ability of crosses to reduce the number of single crosses for testing and to develop the best single-cross combinations. In forestry, some argument can be made for local or traditional seed sources or identified clonal sets as being useful as a stable, standard variety, or at least as the population to develop in the single-population breeding program. In most cases, however, considerable room for improving even these "varieties" exists, and a dual improvement program will be most appropriate.

Single crosses in advanced generations of inbreeding may not be sufficiently viable for seed production, or they may not contain all of the traits desired for commercial seed production. Triple and double crosses may then be feasible for development through additional testing, but development time may be too costly to support such procedures. If trees can be vegetatively propagated, it would seem better to develop lines for single crosses and to expand the number of fruiting branches by such propagation methods as cloning reproductive tips.

HYBRIDS OF POPULATION

Recurrent selection for specific combining ability is an alternative to pure-line development for hybrid performance that is similar to developing recurrent selection populations. In this case, instead of using a heterogeneous set of testers to select for general combining ability, a particular line or stock is used and the genotypes are selected for cross-performance. The best trees are then completely intercrossed, and the new population is again reduced to a set of selected parents according to test-cross results with the same tester stock. This method thus develops a population that complements a specific tester stock. That stock would

have to be well defined, continuously available, and useful in seed production as well as testing. Otherwise, direct recurrent selection for general combining ability would be easier and just as effective.

Gains from selection for specific combining ability depend on differences in hybrid gene effects being accumulated in the parental populations (Cress 1966b). For most forest tree species which have not developed standard varieties of any purity, a reasonable approach to developing high specific combining ability in crosses would be the mutual development of complementary populations such as by reciprocal recurrent selection. The operations involved in this selection system are identical with recurrent selection for specific combining ability, but instead of using a standard tester stock, the pairs of developing populations are used as reciprocal testers, both of which are mutually improved. This method has become the standard for many cross-pollinated crop plants against which other methods are compared.

Theoretical comparisons of methods for hybrid population development methods are difficult to derive, because the specific kinds of dominance effects required to make hybrid breeding advantageous depend upon gene frequency locus and distribution (Cress 1966a). Therefore, while genetic variances, effects on covariances of relatives, and selection advances within generations can be derived (Stuber and Cockerham 1966), the gene frequencies will presumably be diverging in subsequent generations, relative inbreeding within parental population will become stronger, and translation of gene effects and variances between generations will be less well defined. In fact, in the F_1, special dominance effects exist which may not be seen in the F_2, and variations will appear in the F_2 which were hidden by dominance gene actions in the F_1. Nevertheless, empirical studies on the efficacy of reciprocal recurrent selection indicate its value when dominance gene actions are important. Theoretically, the method should be able to utilize any within-population general combining ability not masked by hybrid effects, as well as the interpopulation specific combining ability between the complementary sets (Kojima and Kelleher 1963a).

In an actual breeding experiment contrasting reciprocal recurrent selection (RRS) with single-population development, Moll and Stuber (1971) found in corn that for roughly comparable selection differentials RRS can utilize both general and specific combining abilities. The hybrid product of RRS was slightly better than the best population bred by full-sib family selection in a recurrent selection system. It was also much better than the hybrid between the parental populations which had been bred for general combining ability. Another population was developed by ordinary full-sib family selection but from an initial population which was composed of F_1 of the original parental varieties. This selected population performed at about the average level of the two full-sib family selection populations carried within each parental variety, and not as well as the RRS hybrids. Nevertheless,

full-sib family selection within each parental line was moderately effective and utilized general combining ability variations in the original parental populations. Instead of using the original parental varieties, the parents developed for RRS were also used to try full-sib family selection from each parental variety. RRS displayed a greater development of heterosis in the hybrids and created a better hybrid population than the recurrent selections within the parental or hybrid variety. However, RRS was not as effective in improving the intravarietal performance as was full-sib selection. Moll and Stuber (1971) also found for their constant and moderate selection intensities that the gains by all breeding methods were reasonably constant and hence may be predicted for at least six selection cycles from first-generation results.

In such hybrid breeding programs, as in all breeding programs in forestry, immediate gains in commercial seed are required because development of trees through multiple generations often requires more time than can be justified. Therefore, intermediate products are always required, and, as suggested by Cress (1967) for other crops, production of synthetic varieties in the intermediate generations of RRS population development is desirable. Such synthetic varieties may be composed of entirely different genotypes for each generation, or may include some particularly good genotypes for several generations, until better ones are developed. Regardless of the origin of the parental genotypes, or their homozygosity, commercial seed-production orchards can be composed of a subset of entries with especially good specific combining abilities among crossed parents. Seed requirements would determine whether fewer or more parental combinations are included in the seed-production phase than in the breed-production populations. In many tree species, it is possible to vegetatively propagate a set of especially good clones instead of using sexual reproduction. Thus, intermediate stages of Schreiner's (1966) "synthetic multiclonal hybrid varieties" may also be produced.

The development of hybrid breeds is clearly dependent on the importance of dominance types of gene actions, which are not easy to estimate in the generally heterozygous populations of cross-pollinated species. Moll and Robinson (1967) clearly show that initial estimates of dominance levels can be substantially affected by linkage, and that favorable epistatic combinations of alleles can be lost during breeding and mating (Gardner and Lonnquist 1959). Also, dominance levels among alleles within populations may exceed those between populations. Therefore, hybrid breeding programs developed on the basis of initial estimates of high dominance effects may not be as beneficial as expected. In a comprehensive review of plant breeding in the United States, Sprague (1966) examined the famous hybrid-corn breeding programs of the past and concluded that, strictly from the viewpoint of genetic progress obtainable, selection programs based on developing general combining ability within single popu-

lations could have been at least as good. Furthermore, since the existence of heterosis does not by itself justify a hybridization program (Stuber 1970), we should not assume that the success which corn hybridization has achieved can be applied to other species.

Hybridization has played an impressive role in many tree breeding programs and will continue to do so in many cases (Wright 1964b). Most such programs were developed to combine traits of different but related species; goals included combining acceptable growth rates with acceptable resistance to a disease or insect, or performance in a particularly harsh environment (van Buijtenen 1970). Many examples of species hybrids created for those purposes exist, but few programs have sought cumulative improvement of the hybrid populations. The creation of a base hybrid population through single-population recurrent selection, as outlined by Stuber (1970), is potentially valuable. Throughout Europe, extensive plans exist to introduce special traits into species through species and varietal hybridization. The species chosen are generally good and may be further bred to produce base populations for recurrent selection, reciprocal recurrent selection, or single-cross types of programs (Nilsson and Andersson 1970). Among the early programs designed to create a hybrid base population for future recurrent selection were those on the *Pinus radiata* × *P. attenuata* hybrids (Righter 1960). More recently, Conkle (1970) suggested that reciprocal recurrent selection can be profitably applied to those two species as the parental populations. The extensive development of *Pinus rigida* × *P. taeda* hybrids in Korea has also recently led to the development of reciprocal recurrent selection plans in addition to a continuing program of selection within the hybrid base population (Hyun 1971).

In all these cases, population development still requires that the one or several parental-source populations be large enough to avoid the loss of favorable alleles. The problems of recurrent selection for general combining ability and the strictures on maintaining large population sizes are as important in hybrid as in single-population programs. The limitations on crossing patterns within parental populations and the requirements for large family sizes and numbers of families are the same as for single-population recurrent selection, as are the limitations placed on the selection differential by the requirements of replicated testing. In fact, these problems are even more acute with hybrid breeding programs since both parental populations will often require separate development and testing, and hence some reduction in capacity to replicate tests. If a minimum of 50 to 100 genotypes is deemed necessary in single-population breeding to maintain genetic variation without significant loss of selection intensity and to preserve genetic variations in traits not presently under breeding pressure, then approximately the same number would have to be maintained in each of the parental populations

for hybrid breeding. An adequate test size, including numbers of trees per plot and crosses per parent, would then require about double the effort of single-population development plus testing time. The only relief found in hybrid programs is that inbreeding depression of intrapopulation crosses does not reduce gain in the seed products.

The possibilities of replicating populations for parallel breeding programs are the same for hybrid as for single-population breeding. The same advantages of safety exist, since each breeding unit can be developed for several generations before the parental populations begin to deplete their genetic variations, and the possibilities of selecting among fortuitously good units also exist. However, the advantages of crossing among the better replicates to regenerate variations will presumably not exist for hybrid programs, since they depend on creating complementary gene arrangements. At this time, no theoretical work has been done on these topics, however, and variations on the replication theme have not been explored enough to dismiss the possibility that some forms of replicated reciprocal recurrent selection may be uniquely advantageous in tree breeding.

Except for pure-line development for parents of single-cross hybrids, the performance of parents as individuals is not nearly as important a basis of selection in hybrid breeding as in single-population breeding. Thus, only direct testing of cross combinations is reliable for hybrids. The gain in each generation can be predicted as the variance among selection units, but prediction for future generations is uncertain. Therefore, only the methods which utilize some form of progeny or sib testing are useful. Unless required for other reasons, mass selection and seedling orchard methods in which the observational or test materials are also used as parents are not suitable for hybrid breeding. Separate seed-production operations will almost always be required. If clonal reproduction of the commercially produced genotypes is desired, then a separate operation for regeneration is clearly required after the best genotypes are chosen. For any program in which separate commercial seed production is required, there is no need to maintain this operation in the same areas as the test sites. In fact, controlled pollination may be easier outside the natural range of the species.

A major problem in hybrid breeding occurs after the initial parents within each population have been intermated. Each population may have several thousand trees available for selection. A tree's phenotypic performance is likely to be poorly correlated with its performance as a hybrid parent. The crossing and evaluation phases, therefore, can be massive unless some schemes for staging sequences of testing can reduce the numbers of entries which require intensive handling. Various methods are possible for sequential testing and selecting to reduce their parental numbers to the minimum sizes required. If sequential data on hybrid performance in reciprocal recurrent selection or recurrent

selection for specific combining ability are available, some form of family selection may be feasible. It may also be desirable or even necessary to select within parental populations for general vigor before making hybrid test crosses and reselecting on progeny performance. In all such cases, gain can be estimated by the ratio of the genetic variances of the hybrid generation, as described by Stuber and Cockerham (1966), and the phenotypic variance of the test materials. The selection differential requires the same considerations as for single populations, and the compromise choices for maximizing $s \cdot h^2$ between the s and h^2 elements are essentially the same and require no development here. In reciprocal recurrent selection, for example, the same s may require that fewer individuals be tested and selected as in single-population recurrent selection. The denominator variance of h^2 should be comparable, and the numerator covariances would be equivalent in the additive genetic components to those for progeny-test or sib selection. The numerator covariances, however, would also include a dominance variance contribution which would vary according to gene-frequency differences among the parental populations.

The development of selections in a single population of hybrids is no different from any other single-population breeding program, except that the F_2 generation must be used to represent the base, noninbred population, even though some linkage effects will linger for several generations.

Otherwise, the same basic considerations of population size and mating patterns remain. Single pair matings in any of the various sib or cousin patterns may be duplicated in either kind of program, and expansions of sets of pair matings in hybrid programs into multiple cross or partially controlled cross systems are identical within parental populations. Test matings are the only distinctive feature.

MIXED BREEDING PROGRAMS

Some traits may be best improved by utilizing heterosis, and others by using additive gene effects. If the forest is to be composed of a mixture of tree types, then breeding populations for different objectives may be separated. However, if a single-breeding population is required to simultaneously improve traits by both hybrid and a single-population breeding, then a mixed program is required.

A mixed program may be done by tandem selection, say first for the additively inherited traits in each of the parental populations, and then for the heterotic traits in hybrids. Seed production is then from the hybrid populations in general or specific crosses, while breed-population regeneration requires intrapopulational mating. Alternatively, additive and heterotic gene actions can be simultaneously selected for if information on all performances is available. The additional information on a tree's own performance and that of other relatives would always be useful for gain in

traits dependent on additive gene actions, while progenies of hybrid crosses are required for evaluation of traits dependent on heterosis.

CONSIDERATIONS IN CHOOSING BREEDING METHODS

Mode of reproduction, operational costs, time costs, and types of gene action will determine the optimal breeding system. While it may be simple to derive the basic concepts for estimating gain for any system of s and h^2, errors in estimation of parameters, costs, and operational interactions are considerable. Thus, testing large numbers of trees increases s but introduces wider environmental and measurement errors and hence decreases h^2. Also, the cost of increasing s by one unit at high levels of selection intensity is very high, because vastly greater numbers and proportions are needed to change s. Changes in operational costs affecting h^2 are seldom linear; the marginal cost of adding replicates, for example, can be low if other tests and experiments are to be conducted anyway. Benefits of small increases in breeding products are also unlikely to be linear functions of gain in physical parameters, and, therefore, relative costs and risks are likely to be nonlinear with respect to experimental size. Hence, small gain differences even at high immediate land or operational costs may justify choosing a more expensive breeding program.

The simplest form of breeding is rapid and cheap mass selection that any intelligent forester can apply to seed collection and forest regeneration. While clearly the simplest, it may not be as efficient or as profitable as the more sophisticated methods already described. There are clear differences in the manner and efficiency in which the various methods accumulate a favorable set of alleles, since information is used differently and matings are made differently. If mass selection is considered as essentially costless or, at least, no more costly than other methods of seed procurement, then the costs of controlled crossing programs are only the marginal costs of making specific crosses and keeping ancestral identities. The benefits of such additional operations lie in the control of effective population size for any given number of trees and in the possible uses of family selection to increase heritabilities. Of the various crossing methods discussed, from partial to complete control, fewer parents would be required for breeding if ancestries are known, because some expected or feared level of additional inbreeding would have to be assumed without control, thereby decreasing the effective population size. To maintain some minimal N_e, more parents would be included in the breeding population, making less intensive selection desirable for methods with less ancestral control. In addition, controlled crosses in the breeding population, especially those of the diallel patterns, allow the breeder to choose among favorable combinations of family and individual selection to maximize the product of s and h^2 and the sum of gains derivable from each stage. There is some trade-off

between s and h^2 and between the efforts of making many crosses with few trees per cross, and few crosses with more trees per cross, but optimal combinations do exist. An additional benefit of controlled crosses is the possible use of specific combining abilities in temporary or short-run breeding operations. Only if families are identified can they be used as recombinants for increasing specific combining ability responses. Indeed, the detection of such efforts themselves requires controlled intercrossing.

In addition to rapidly improving breeding populations, controlled crossing permits somewhat more precise selection if progenies within any generation are tested. General and specific combining abilities can be estimated and used to improve the commercial seed product within any one generation. Also, while slower and more expensive, controlled crossing can achieve more gain per generation if the s and h^2 for progeny tests compensate for the time lost in breed development.

If the tree breeder is responsible for several species, coordination of operations can be overwhelmingly complex. He must determine for each species the unique genetic and phenotypic means and variances, costs of operation, etc., to arrive at optimal operational compromises with respect to selection differentials, heritabilities, population sizes, etc. Then, a desirable extent of controlled crossing desired and its pattern and sequence can be established for the breeding operation. With limited time and resources, various strategies may be followed to maximize total improvement. Some may wish to establish a complete program for each species, taking them in some order of importance, while others may wish to attack all species simultaneously in a single program. In general, however, efforts must be concentrated on those most valuable species which can profit most by intensive breeding programs. Programs for other species usually are limited to minimal mass selection or simple recurrent selection. For such multiple species programs, decisions on such questions as desired selection differential for family and individual selection affect the effort affordable on other species. Hence, efficient breeding of a key species can determine the form of the entire program. Most multiple species programs will probably have three classes of operations. One or two widely planted species with high-genetic-gain potential will receive maximum effort. Several species will receive moderate effort to establish breeding populations with controlled breeding options. And minimal efforts will be directed to species that require some improvement but for which progress is limited by lack of knowledge or planting potential. Allocation of effort among breeding agencies within regional cooperatives or governmental units could assure the long-term development of all potentially useful populations.

All of the above methods depend completely on additive types of gene action and eventually lead to homozygosity within replicate trials. The choices between them rest on testing efficiencies and costs, maximizing both s and h^2 at minimal costs, maintaining

genetic variance, and species characteristics of the plants.

A major difference in concept exists if any of the hybrid systems are pursued. Hybrid selection methods clearly depend on developing complementary sets of genes in the parental populations. If gene actions turn out to be largely additive, hybrid selection may produce the same results as single-population selection at a higher cost. However, if overdominant gene actions exist and can be accentuated by developing complementary allelic frequencies, these hybrids can offer great advantages. Otherwise, less than complete dominance and epistasis may not be any better utilized than in the single population. Various types of semihybrid programs, however, such as developing single-population breeds from a hybrid base population, may offer considerable advantages when species or provenance hybrid combinations bring valuable traits into the population for further concentration. Eventually, the hybrid programs, too, will lead to homozygosity within parental sources. In hybrid programs, however, homozygosity will occur so far in the future that new variants will be generated if populations are kept large.

In forest tree breeding, the present need is clearly for experimental evidence on the biological and economic feasibility of the diverse methods that appear to be available. Responses to moderate selection and inbreeding must be found, and analyses are needed on the effects of different environments on phenotypic and genetic variances. Small replicate populations are particularly well suited to single-population breeding, and the effects of their use require empirical testing. Organization of hierarchies in such replicated breeding populations, as recommended by Namkoong and others (1971), should be explored for single-population breeding and can be adapted to hybrid breeding in which the parental populations are separately developed. Experimental testing of breeding methods on rapid generation sequences is required.

SEED SOURCE SELECTION

The first step in all breeding programs has traditionally been the choice of provenances to utilize available geographic variations. Since forest trees have been relatively unselected, unique opportunities exist for exploiting natural racial variations within species. Regardless of any other patterns of variation that may be discerned, it is only reasonable to examine genetic differences among subpopulations for their possible utility in building breeding populations and for any limitations which may exist in crossing among them or with other potential parents. One objective of selecting trees from the best provenances is to collect the best alleles into the base population and to increase their frequency by breeding without having to return later to unimproved populations for useful alleles. It is most practical to start breeding at as high a level of value as possible. However, proper breeding of an average provenance will soon yield varieties better than any existing unimproved provenance. Therefore, the breeder will have

to choose between seeking improvement either through an initial cycle of provenance tests or through immediate breeding of the best provenance he has available. Variation among provenances is usually exhausted for purposes of selection among them after the best have been chosen.

Significant variations exist among local populations of almost all wide-ranging forest tree species. While there are some notable exceptions to this general phenomenon, Wright's (1962) comprehensive review of provenance differences strongly indicates that it is wiser to assume large subpopulation variations and to prove the assumption wrong than to ignore the possible existence of such variations. Variations among provenances and their possible uses have already been well described elsewhere. The discussion here is confined to the discerning of patterns of variation and the sources and uses of such patterns.

The traditional concept has been that selection for vegetative vigor should be limited to local sources, which are presumably best adapted to local environments. Strong support for this concept came from the classical studies on *Achillea* by Clausen and others (1948). They suggested that natural selection eliminates all migrants and genetic segregants not suited to local environments, and that vigorous growth was highly correlated with competitive ability and fitness. Further support for this view with forest trees was developed by Langlet (1936) and Wakeley (1954). In any test, growth would be expected to be best from the local source with some degree of decrease as a function of environmental distance.

Conflict with the traditional model was noted by Wright (1962), who observed that local provenances of Douglas-fir were not always the most vigorous. Careful analysis of loblolly pine performance in the "Southwide Pine Seed Source Study," by Wells and Wakeley (1966), demonstrated the existence of an optimal growth zone along the southeastern coastal border of its range. Genotypes from this zone outperformed all others in local tests up to 200 miles inland. Similarly, growth potentials for genotypes from more central populations and moderate climatic sources of black walnut were considerably superior for growth vigor far north of their present range (Bey 1970). In addition, optimal climatic and soil regions near the centers of the ranges of slash pine (Squillace 1966b, 1966c) and of ponderosa pine (Conkle 1973) produced genotypes which are superior far outside of their local regions.

It would be valuable to know why natural selection has not produced locally optimal vigor or, perhaps, fitness. It may be that vegetative vigor is not as well correlated with reproductive fitness as we foresters might suppose, especially since we measure vigor in plantations and not under natural conditions (Squillace and Kraus 1959). Certainly, however, vegetative vigor and reproductive fitness cannot be completely independent. More exact models and tests of the relationship between vigor and fitness are re-

quired, but the following discussion assumes that less than maximal fitness is being maintained in some populations. Accidental drift and restricted migration could produce random deviations around generally fit area means, but the regular patterns that have been observed cannot be so explained. It seems more likely that variations in environments have more critical effects on fitness values in ecologically marginal areas, and that the lack of response to selection for vigor is itself a defensive response to variability in environmental requirements. The concept of the existence of stable, optimal populations evolving under variable environments has been extensively developed by Levins (1968), who showed that maximum fitness over several generations can be achieved by a population that is not maximally fit in any one environment but reasonably good in all. If variations in climate, soil, or other environmental factors are large, it can be advantageous for populations to remain more conservatively adapted to the harsher environments. In forest tree species, especially those on ecologically marginal and variable sites, it can be advantageous not to respond to selection for what may be only transiently favorable site factors. Only in more stable, optimal areas would fine adjustments to environments add to the long-term fitness of the species.

One mechanism to dampen response to selection is a high migration rate among populations. Antonovics (1968a) has recently shown that among perennial organisms even limited amounts of pollen migration can strongly inhibit immediate responses in gene frequency to selection. Hence, currently unfavorable alleles can be maintained at intermediate frequencies if migration is effective. If the correlation between vegetative vigor and fitness is high, however, strong selection will clearly tend to produce vigorous local performance even with pollen migration (Endler 1973). But if the correlation with commercially important traits is low, the breeder should consider provenance selection in regions of optimal ecological development and minimal environmental stress as defined by the species itself.

An altogether different feature of provenances which may require special measures for selection is the genetic variance within seed sources which itself may vary among populations. If few populations are selected for breeding, then they should contain as much of the potentially useful genetic variants as is possible to obtain. In some cases, only large populations of species with considerable migration will be selected, and little extra care will be required. In other cases, however, the populations may be relic stands, plantings of limited parental origins, or even mixtures of a few clones as sometimes occur in Japanese Cryptomeria stands (K. Sakai, personal communication). In those cases, remedial efforts to regenerate genetic variations may be useful. One may select trees only from those populations with a large effective population size and genetic variance, or select among several stands of different origins to assure a low coancestry among the selects.

Much provenance research involves the discernment of relations between environmental and yield factors. For example, after a local provenance test the breeder often wishes to estimate the relations between environmental variables at the seed source and performance in his plantation. The interest for population genetics lies in determining the extent of genetic segregation in allelic frequencies and whether substantial genetic variance exists within or between stands. The extent to which variation in traits of interest is determined by environmental factors indicates the relative strength of directional selection and migration versus drift and other random forces in determining allelic frequencies. The analysis of multiple regression in several traits simultaneously is therefore of value in interpreting genetic population structure. The genetic covariance matrix among traits, estimated after interpopulation effects are removed, represents the multivariate analog of the simple genetic variance within populational subdivisions. One might wish to simplify interpretation by using canonical or principal component analysis, but the total regression and residual genetic covariance on all the traits should also be estimated.

The matrix of $\frac{p(p+1)}{2}$ genetic variances and covariances among p traits is therefore desirable to estimate, and general linear hypotheses (on additive or dominance effects in multivariate space, for example) can be tested by multivariate analogs of univariate analyses of variance. Thus, maximum-likelihood testing of the dispersion matrix among provenances or among half-sib families in several traits should be performed, and cluster analyses should be attempted to discern communities of similar provenances.

It is often difficult to define the location of optimum provenances where regular patterns of response exist, especially if plants have been moved much by natural or human endeavors. If high-vigor zones occur at random, then only complete or random sampling would locate them with any known probability. However, if trends exist, even with local error variations, the breeder may wish to weight his sample in favor of areas most likely to produce good genotypes.

Variables in source environment, such as seasonal rainfall, elevation, soil type, and length of growing season, that influence various expressions of yield can be identified. And for tests at one planting location, multiple regressions of all yield variations on all environmental-source variables can be determined.

For any one yield variate, the surface of response can be estimated if enough source environments are sampled and maxima and minima are estimated on those surfaces by standard linear or nonlinear regression. The sampling problems are no different from any other multiple regression problem, except that the combinations of environmental variables are not subject to simple manipulation but must be sampled as they exist in nature. If the

objective of provenance testing is to determine general variations and means, then general, range-wide or random sampling may be best. However, if the objective is to estimate the location of optimal regions, then heavier sampling around expected optimal regions is desirable. For example, if more moderate environments than available in local sources are expected to yield more vigorous trees, a pattern of sampling using the local source as an extreme and suspected optimal regions as centers of sampling may be feasible. Combinations of environmental variables may be sought which, in the dimensions of those variables, are constructed in concentric circles or rectangles with replicated center points. Such variables can be very efficient estimators of the surface for maxima near the suspected region. The surfaces may be as simple as the quadratic (Namkoong 1967) or some more complicated asymptotic functions (Sarvas 1970) which require heavy sampling in regions of maximum curvature, but all benefit from planned sampling of environmental variables. Provenances can also be analyzed for not only mean differences, but also for differences in reaction to sites. The results may indicate different levels of genotype-site interactions which can be studied by regression types of analyses (Butcher and others 1972). The problem for most programs, however, is that several traits are of interest simultaneously, and that simultaneous estimation and a unified form of evaluation are required. Estimation problems for the multivariate case are not especially difficult, but they require the estimation of a matrix of regression coefficients instead of a simple vector (Namkoong 1967). Except for problems with missing data, the only new concepts involved are associated with the distribution of multivariate moments, and they should cause little difficulty for the forester. The greater practical problem is that the optimum environment for one trait may not be optimum for others, and hence selection of an optimum set of environmental variables is not simple.

If the value functions for the combined traits of interest are independent among traits and can be well approximated by a linear function, environments can be evaluated in terms of that linear function. The evaluation can be made as if a single-value trait was being measured, since, under the assumption of linearity, relative values of traits do not change, regardless of the actual levels of the trait variables. Nonlinear value functions are discussed in greater detail in chapter 4 in the section on evaluation. For the present discussion, it is sufficient to state that a solution for an optimal environmental vector may indicate a combination of environmental variables which does not exist in nature. For example, for a given value function, its maximum within the space of environmental variables may lie at a combination of say low winter temperatures and high winter rainfall. This combination may not exist. To obtain trait combinations in provenance selection, then, mixtures or hybrids from complementary regions may provide material for future selection. For example, source A may

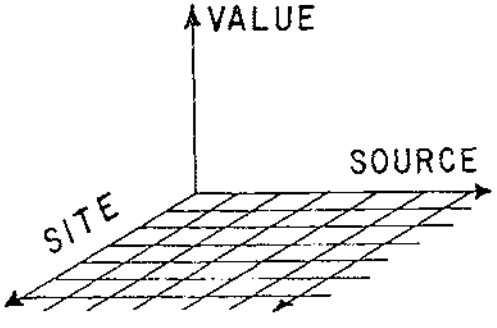
promise good growth but little resistance, while source *B* promises high resistance but poor growth. If both traits are necessary and not otherwise simultaneously available, a hybrid would contain both traits at intermediate gene frequencies and could promise greater breeding gains. While the estimation and selection problems are general for all breeding evaluations of a multivariate nature, breeders dealing with provenance selection will immediately be faced with choices of mixing the most useful sources for multiple traits in the base population.

In addition to problems of evaluating a multiple-regression surface for multivariate decisions, a tree breeder is seldom interested in one planting site. He must usually have to consider what single source or combination of sources may be suitable over a range of sites and how they will change for a set of planting environments. Complete sampling of all sources on all sites is desirable but often not feasible. Efficient sampling for testing suitable provenances on a sequence of sites would require that some changing subset of sources be tested on each site if some choice in source sampling is possible. A partial sampling design may be like:

		SOURCE				
		A	B	C	D	E
	a	X				
	b	X	X			
	c	X	X	X		
Site	d	X	X	X	X	
	e		X	X	X	X
	f			X	X	X
	g				X	X
	h					X

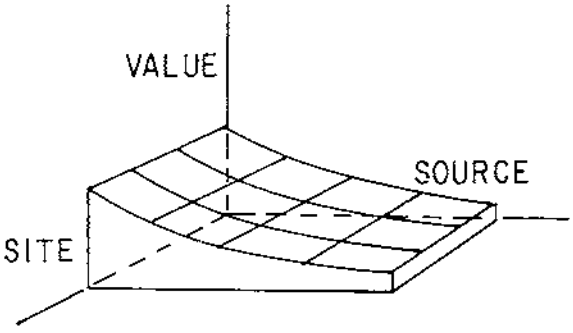
Overlapping sources among planting sites are required to determine general source effects and to distinguish between source × planting site interaction effects and general source average performances.

A complete factorial sampling of all sources on all sites would provide a complete picture of value at each combination, and we could then describe a 3-dimensional factorial response surface of value:

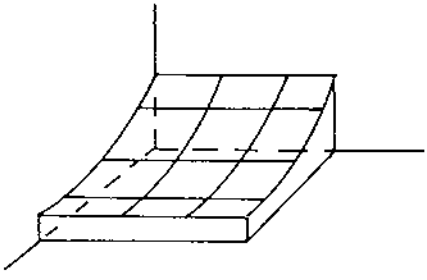


If sources and sites are identically ordered and the order has some relation to value, some simple surfaces may be described for some simple hypothetical results.

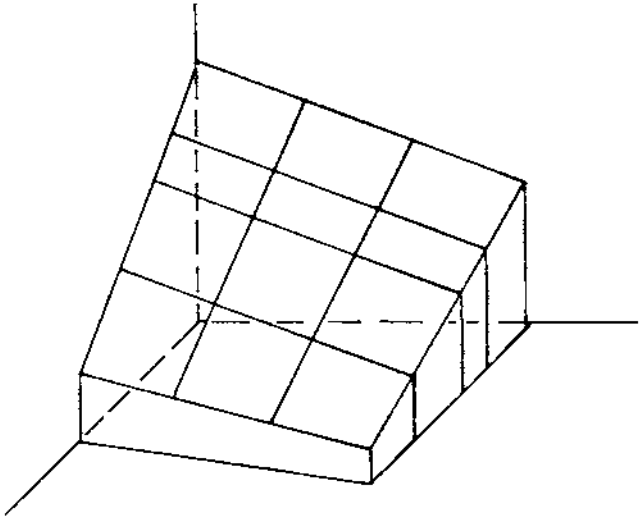
If sources differed but were identical in response to all sites, the value surface would be like:



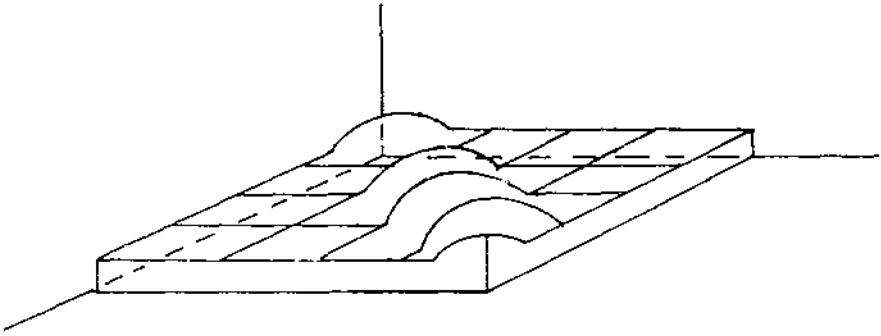
If sites differed but had identical effect on all sources, the value surface would resemble:



If both sites and sources differed but no interactions existed, the value surface would resemble:



If interaction existed, irregular surfaces would display various forms. A condition of local sources always being best would look like:



Less regular and mixed surfaces would have to be generally expected. To the extent that each site has unique optimal sources, the site-by-source interaction can be expected to be high. To the extent that sources perform consistently over different sites, the provenance source will have a large effect at the expense of the interaction component. It is such cases as these which were alluded to in the discussion of optimal ecological zones. The gain achievable is directly estimated by the mean differences observed.

In tests designed to evaluate provenance selection in which only a sample of all possible provenances is taken, gain estimation is a simple analysis of ordinary gain estimates by regression. If the interaction component is high and selection is to be generally among the best test sources, then the expected selective advantage for starting with the best sources is the selection differential $\times$ h^2 (provenance); where h^2 (provenance) is covariance (provenance test value, breeding value) $\div$ variance (test values), and where the selection differential is the difference between the population mean and the mean of the selected provenance. In this case, the numerator covariance of the provenance h^2 will be largely the interaction variance component plus any contributions due to persistent provenance performances.

If a mixture of sources is selected for starting the breeding population, materials with different gene frequencies will sometimes be mixed in the breeding. Any dominance and epistatic effects will then produce genetic recombinations and genetic variances in the F_2 generation unforeseen in the parental or initial crossing generations. It would, of course, be beneficial to start into recurrent selection either for a hybrid system or for general combining ability with a hybrid base population with some experimental information on the importance of nonadditive effects. However, if most provenance crossing displays additive and averaging effects and little dominance or heterosis, as appears true for most pine species, then provenance selection is simply a higher organizational form of family selection. Then, selection may be made in a tandem fashion—first provenance or source, then family and

individual. Simultaneous selection on an individual-tree basis is also possible if some index weight is given to family and provenance collateral relatives in judging individual worth. The effects of linkage disequilibrium, however, will be felt for several generations.

To the extent that provenance selection is deemed worthwhile, some question exists as to how much gain can be achieved by second or multiple stages of refined selection of stands within general provenance areas. If the initial sampling was small, then the response surface is poorly estimated and the model form of the surface may not even be detectable as being especially good or poor. In such cases, the breeder may: (1) choose the best among those he has sampled, (2) estimate and select from an optimum environment, or (3) resample the population for further testing. While further testing increases the chances for greater precision and gain, time and experimental costs also increase, and one might be better off to start a breeding operation if the benefit of better provenance selection can be overcome by a generation of within-population breeding. Thus, for example, the value of the selection of seed-production areas or stands may not be worth much extra time or testing if stand heritabilities are low and pollination is uncontrolled. The more intensive the initial sampling was, the less chance that either a more precise surface estimation would indicate other optima or that stands other than those actually sampled would be much better for starting a selection program. Provenance selection may occasionally be all that is desired if breeding in even minimal programs cannot be supported (Wright 1971). Thus, one generation of well-designed and intensively sampled provenance tests is often as much testing as is desired for starting either hybrid- or single-population breeding. Subsequent population developments would then proceed either to develop the separate parental population for hybrid production or single populations for some form of recurrent selection for general combining ability.

CHAPTER 4
TESTING AND ESTIMATING VALUE IN FOREST TREE BREEDING

It is clear that defining, measuring, and using gene effects in a breeding program, or simply understanding the variations and operations of natural events, involve highly complicated studies in which optimum solutions may be difficult to produce. Efficient estimation, breeding population development, and seed production in a breeding agency all require careful design. In addition, since the breeding program requires that all tasks be integrated, the various phases of the complete program often require simultaneous operation. Design problems can be highly complicated and arduous, since they involve genetics, statistics, and mathematics applied to the practical problems of testing, selecting, and breeding trees for a multitude of purposes. Nevertheless, forest tree breeders are required to also conduct tests which require unusually large amounts of time and space, and hence demand efficiency in achieving all of the experimental goals sought. Testing trees, families, or provenances for selection in forestry is complicated by changing environments and changing requirements for data on performance of different relatives on different planting sites. Because of these problems, the plant resources required for testing must be efficiently allocated.

In this chapter, testing and evaluation techniques are discussed as additional objectives of efficient breeding. The use of information on relatives in a linear function and the evaluation of multiple traits, also in a linear function, are discussed. The use and evaluation of correlated trait selection and response are examined and nonlinear value functions are discussed. Genotype-by-environment interaction and competition models are then described. In the following chapter, strategies for developing integrated research and development programs in forest genetics are discussed.

INDEX ON RELATIVES

One method of increasing selection efficiency is to use as much information as may be available on the performance of various kinds of relatives. The more relatives that exist, the more precisely the genetic value is measured, and the closer their relationship to the units being selected, the more reliance can be placed on their performance as supplements to the individual's own performance. Combining the information, which may indicate for example that

an individual is good but its half-sib family and parents are poor, requires that the performances be put in a form that permits fair weighting of the contrasting data to give precise selection and hence maximum gain when the units are selected.

A linear function is a simple, reasonable form for a composite index value: $I=b_1x_1+b_2x_2+\ldots$, where the x_i are the variables for each kind of relative, and the b_i are the index weights to be determined. Then the true value of a tree, V, can be estimated with some error by $V=I+e$, where e is the error in estimating true value by the I index and is to be minimized. Then, as in any selection scheme, expected genetic gain in value $E(\Delta G_V)$ when some imprecision in selection exists can be approximated by the regression function $E(\Delta G_V)=h_I^2(\bar{I}_s-\mu_I)^2$, where h_I^2 is the regression heritability of the index values, and $\bar{I}_s-\mu_I$ is the selection differential between the mean index value of those selected and the general population mean index value. The regression heritability $h_I^2=Cov\ (I,V) \div \text{Var}\ (I)$. For normally distributed traits (x variables can be expected to be approximately normally distributed especially if the x's are means), the expected selection differential $E(\bar{I}_s-\mu_I)=(z/p)\sigma_I$, where, as previously discussed, z is the ordinate of the standardized normal distribution at the truncation point, and p is the proportion selected. Then $E(\Delta G_V)=z/p\ \frac{\sigma_{VI}}{\sigma_I}=z/p\ \rho_{VI}\sigma_V$.

Since σ_V is fixed in the population, and z/p is chosen by the breeder to satisfy demands previously discussed, we maximize gain by maximizing the correlation ρ_{VI} or by maximizing the error variance of I around V. Using least squares procedures as in multiple regression, an underline to indicate a vector, and a prime to indicate a transposition, the relationships which we require are:

$$P\underline{\hat{b}}=\underline{\text{Cov}\ (x_i,V)}$$

or

$$\underline{\hat{b}}=P^{-1}\ \underline{\text{Cov}\ (x_i,V)}$$

where

$$P=\begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} & \cdots \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

the matrix of phenotypic variances and covariances,

$$\underline{b}=\begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{pmatrix}$$

the vector of weighting coefficients, and

$$\text{Cov } (x_i,V)=\begin{bmatrix}\text{Cov } (x_1,V)\\ \text{Cov } (x_2,V)\\ \cdot\ \cdot\ \cdot\\ \cdot\ \cdot\ \cdot\\ \cdot\ \cdot\ \cdot\\ \cdot\ \cdot\ \cdot\end{bmatrix}$$

the vector of genetic covariances between the relatives and the genotypes being tested.

It can be further seen that since:

$$\hat{\underline{b}}=P^{-1}\text{Cov } (x_i, V),$$

and $$I=\underline{b}'\underline{x},$$

$$\text{Cov } (I, V)=\underline{b}'\text{Cov } (x_i, V),$$

and $$\sigma_I^2=\underline{b}'P\underline{b}.$$

Therefore, $$\sigma_I^2=\underline{b}'PP^{-1}\text{Cov } (x_i, V)$$

$$=\text{Cov } (I, V),$$

and therefore, $h_I^2=I$.

Since $$E(\Delta G_V)=(z/p)\sigma_I h^2,$$

$$E(\Delta G_V)=(z/p)\sigma_I,$$

in the scale of the index measures taken on the relatives, and gain can be estimated as:

$$E(\Delta G_V)=(z/p)\sqrt{\underline{b}'P\underline{b}}$$

or $$E(\Delta G_V)=(z/p)\sqrt{\text{Cov}'\ (x_i, V)P^{-1}\ \text{Cov}\ (x_i, V)},$$

assuming that P, $\underline{\text{Cov } (x_i, V)}$, and b are all well estimated.

This kind of index has some potential use in forestry when sib, parental, and clonal data can all contribute to the estimate of value of a tree (Namkoong 1966b). We can get an intuitive feeling for how the index gives weights to the various relatives if we ignore the phenotypic covariances in P and instead look at only the phenotypic variances.

Then $$\hat{\underline{b}}=\begin{bmatrix}\sigma_{x_1}^{-2} & & \\ & \sigma_{x_2}^{-2} & \\ & & \sigma_{x_3}^{-2}\end{bmatrix}\begin{bmatrix}\text{Cov } (x_1, V)\\ \text{Cov } (x_2, V)\\ \cdot\ \cdot\ \cdot\end{bmatrix}$$

Therefore, $\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \end{bmatrix} = \begin{bmatrix} \text{Cov } (x_1, \text{V})/\sigma_{x_1}^2 \\ \text{Cov } (x_2, V)/\sigma_{x_2}^2 \\ \text{Cov } (x_3, \text{V})/\sigma_{x_3}^2 \\ . \\ . \end{bmatrix}$

We can see that the closer the relationship to the tested genotype is, the higher the covariance will be. The better we can estimate the breeding value of a genotype by reducing nongenetic variation, the lower the phenotypic variance ($\sigma_{x_i}^2$) will be, and hence the greater the b_i coefficient will be. For those species which can be tested clonally, we would have direct measures of the genotype and hence, the covariance of ramets with ortet would be the total additive genetic variance. Also, clones can be planted in several locations and replications and their value determined with minimum error. Hence, if clones can be used, their weighting will be very high (Libby 1964).

The variances in the diagonal of the P matrix are more easily estimated from the variance of various family means or clonal means. The covariances as between, say, full-sib and half-sib family performances in a balanced experiment may be zero but in unbalanced experiments may not be. Maximum likelihood estimates of the index can still be computed (Henderson 1963). Independently estimated means contribute no covariances to the P matrix. The other genetic covariances between an individual selection unit value and the various relatives' means may be simple genetic covariances as for parent-offspring relations, but may involve more complicated relationships between an individual and its family if the individual itself contributes to the family mean. In such cases, finite population correction factors can be used and the standard indices estimated (Henderson 1963).

The more complicated the relationships involved, and the more different kinds of relatives are estimated, the less well are the various covariance matrices estimated and, therefore, the poorer are the estimates of the optimal b coefficients. In a fairly extensive mixture of crossing and selfing data from a diallel estimation experiment, Cockerham and Matzinger (1966) found that simplified weighting procedures may often prove to be at least as good as the complete least squares analyses. In fact, as analyzed by Williams (1962) and Patel and others (1962, 1969), poor estimation of the P matrix can, over several trials, lead to poorer correlations of indices with true breeding value than simplified weighting procedures on the basis of genetic correlations among relatives or cost and precision of estimates.

Thus, for selection, very extensive tests of many kinds of relatives may not be worth construction of separate experiments, even

if they would be useful if they are otherwise available. More remote relatives have little to add to selection precision and may cause more problems in estimation of value than their limited assistance is worth. Furthermore, the use of relatives is most helpful in cases of low individual heritability, and even progeny testing may not be worth the cost in time or effort required (Namkoong 1970a) unless heritability is very low. Thus, while it is useful to have data on relatives, selection gain alone may not be worth the cost of extra matings and plantings to obtain the data. Each program, however, must make that cost analysis for its own benefit valuations. It must also be considered that multiple traits are often simultaneously selected for and that the desirability of using information on relatives usually varies among traits. For example, selection for growth may have a high heritability while selection for disease resistance may be low. In addition, if the correlation between them is negative, one may be forced to obtain the additional information for simultaneous selection from progeny tests (A. E. Squillace, personal communication). Thus, a complete evaluation of progeny testing for obtaining information on any kinds of relatives generally requires a multiple-trait evaluation.

INDEX ON TRAITS

In the above discussion, we have generally assumed the existence of a single measure of value on each unit of selection for which the various relatives are measured in some common way. In general, however, several traits are selected for and the simultaneous improvement of all traits is often desired. Alternatively, methods of improving one trait at a time in a tandem sequence or of simply using truncation selection for each trait independently to arrive at the same selection differential have been shown to be poorer than simultaneous index selection (Young 1961, 1964). One method of reducing the several-trait measures to a single scale is essentially the same as for selection with multiple relatives —a linear index function with weights estimated to maximize the gain in value. Similar to the previous discussion, a linear function is appropriate for independent evaluations of the traits, each of which increases in value in a linear form. While this is clearly a poor approximation, it may not be bad for small changes in each component trait.

The index we wish to build would thus weight the trait variables in a linear function: $I=b_1y_1+b_2y_2+\ldots$, and each trait would have some relationship to value V as before. The only added complication now is that value is some function, presumably linear, of each trait's true value, $V=a_1g_1+a_2g_2+\ldots$, where g_i are the inherited or true genetic values in the selection units, and a_i are the economic weights in the linear value function. As before, the maximization of value requires the least squares estimates

for the $\underline{b}$ coefficients from:

$$P\hat{\underline{b}} = \underline{\mathrm{Cov}(y_i, V)}$$

or $$\hat{\underline{b}} = P^{-1}\underline{\mathrm{Cov}(y_i, V)}.$$

The new problem is that the covariance of each variable y_i must be taken with the linear function of value, $V = a_1g_1 + a_2g_2 + \ldots$. Since we are in a breeding operation, the covariance of a phenotypic measure (y_i) with its genetic value (g_i) in a selection unit is the genetic variance or, more often, the additive genetic variance of the trait $\sigma_{A_i}^2$. Similarly, the covariance of a phenotypic measure on trait i (y_i) with the genetic value of another trait (g_j) is the genetic covariance or additive genetic covariance between traits $\sigma_{A_iA_j}$.

Then, since value $V = a_1g_1 + a_2g_2 + \ldots$,

the $$\underline{\mathrm{Cov}(y_i, V)} = \underline{\mathrm{Cov}\ (y_i, a_1g_1 + a_2g_2 + \ldots)}$$

$$= \begin{pmatrix} a_1\sigma^2_{A_1} & + a_2\sigma_{A_1A_2} & + a_3\sigma_{A_1A_3} & + \ldots \\ a_1\sigma_{A_1A_2} & + a_2\sigma^2_{A_2} & + a_3\sigma_{A_2A_3} & + \ldots \\ a_1\sigma_{A_1A_3} & + a_2\sigma_{A_2A_3} & + a_3\sigma^2_{A_3} & + \ldots \\ \ldots & + \ldots & + \ldots & + \ldots \end{pmatrix}$$

$$= (G)\underline{a},$$

where G is the genetic (additive) covariance matrix, and $\underline{a}$ is the vector of economic weights.

Then $$\hat{\underline{b}} = P^{-1}G\underline{a}.$$

As before, the expected gain in value, using optimum weights and assuming a linear economic model, is:

$$E(\Delta G) = (z/p)\,\sigma_I h_I^2$$

where $$h_I^2 = \mathrm{Cov}(I, V) \div \mathrm{Var}(I).$$

Since $$\hat{\underline{b}} = P^{-1}G\underline{a}$$

and $$I = \underline{b}'\underline{x}, \text{ and } V = \underline{a}'\underline{g},$$

then $$\mathrm{Cov}(I, V) = \underline{b}'\mathrm{Cov}(y_i, g_j)\underline{a}$$

and $$\mathrm{Var}(I) = \underline{b}'P\underline{b}$$

$$= \underline{b}'PP^{-1}G\underline{a}$$

$$= \underline{b}'G\underline{a}$$

$$= \underline{b}'\mathrm{Cov}(y_i, g_j)\underline{a} = \mathrm{Cov}(I, V).$$

Therefore, $h_I^2 = 1$

and $E(\Delta G) = (z/p)\sigma_I$

in the units of measure used to derive I. Even if a linear economic function is adequate, poor estimates of P or G lead to the same problems in estimating the optimum index weights as previously discussed. The investigations of Williams (1962) and Patel and others (1962, 1969) were directed to these kinds of indices and resulted in recommendations that the estimated $\underline{b}$ weights were better when the linear, additive genetic variances were high relative to nonadditive genetic variances and that estimates of the coefficients were restricted to within reasonable limits.

A more general condition for selection index construction includes cases in which it is wished to keep some traits in the population unchanged. While essentially similar in form to the indices constructed above, the value is to be maximized under constraints which require zero-valued functions to exist (Tallis 1962). Kempthorne and Nordskog (1959) state the restrictions in algebraic form as linear functions of genotypic values $\underline{c}'\underline{g} = 0$, and maximize the value function using Lagrangian multipliers. The optimum $\underline{b}$ estimate then is:

$$\hat{\underline{b}} = [I - P^{-1}GC(C'GP^{-1}GC)^{-1}C'G]P^{-1}G\underline{a},$$

where C is the matrix of coefficients of the restricting equations and the other matrices are as previously defined. If $C = I$ as for the case of no restrictions, the equation reduces to the familiar

$$\hat{\underline{b}} = P^{-1}G\underline{a}.$$

A different kind of index is required if dominance types of gene action are used, as in hybrid or mixed breeding systems. In such cases, the genetic value of the entries is also dependent on dominance effects and dominance and additive-by-dominance genetic variances. These genetic variances and covariances, however, are defined according to their hybrid population statistics as developed by Stuber and Cockerham (1966).

CORRELATED RESPONSE

The effects of selection for one set of traits on changes in other traits are clearly of great interest to foresters, since many forests are subject to multiple simultaneous demands and future forests are subject to selection for different sets of traits. Maintaining variation in the forests by maintaining huge populations or by selecting to maintain some intermediate mean values may both be useful, though large population size is easier to use and likely more conserving of variance. A more critical problem is that genetic correlations and hence correlated responses to selection are notoriously variable from generation to generation. If the

correlations among traits are due to nongenetic sources, they can change with environmental or cultural variations. If they are partly genetic, then they may change as linkages change or as they influence the relative effects of pleitropy or epistasis on trait correlations.

It may also be useful to select trees for improvement in one trait like yield by measuring a more easily observable associated trait like height growth. The associated trait may be measurable under less environmentally variable conditions or measurable several years earlier than the trait of direct economic value. Selection efficacy depends on the nature of the correlation between the traits.

Faced with problems associated with poor estimations and changing genetic correlations, greater assurance of achieving gains can be given if the numbers of traits are limited to a few with relatively assured values and breeding is followed with large population sizes.

One form of index selection on correlated traits which would be extremely valuable in forestry is selection on a set of juvenile traits for mature tree performance. Several juvenile traits that can be easily measured may, by themselves, have an economic value of zero but still be useful if correlated with one or several mature tree traits. In terms of single pairs of juvenile-mature tree traits, Nanson (1970) has clearly demonstrated that for many traits in a wide variety of forest tree species, these correlations are high enough that substantial savings in cost and rate of genetic gain can be achieved by selection early in the life cycle. The gain from selection on x on the correlated trait of value y, using linear approximations, is:

$$E(\Delta G_y) = i\sigma_x h_x^2 b_{x,y}$$

when $h_x^2 = \sigma_{Ax}^2/\sigma_{Px}^2$, the heritability of trait x, and $b_{x,y}$ is the regression between the genotypic value of y on the genotypic value of x. The numerator of the regression is usually restricted to the additive genetic covariance between the two trait performances σ_{AxAy}, while the denominator is σ_{Ax}^2, the additive genetic variance of x.

Then $$E(\Delta G_y) = \frac{\sigma_{Ax}^2}{\sigma_{Px}} \cdot \frac{\sigma_{AxAy}}{\sigma_{Ax}^2}$$

$$= i\frac{\sigma_{AxAy}}{\sigma_{Px}} = i\, r_{Px,Ay}\, \sigma_{Ay}$$

or as Falconer (1960) states:

$$E(\Delta G_y) = i\frac{\sigma_{Ax}^2}{\sigma_{Px}} b_{AxAy}$$

$$= i\frac{\sigma_{Ax}}{\sigma_{Px}} r_{Ax,Ay}\, \sigma_{Ay}$$

$$= i\frac{\sigma_{Ax}}{\sigma_{Px}} \frac{\sigma_{Ay}}{\sigma_{Py}} r_{Ax,Ay}\, \sigma_{Py}$$

NONLINEAR RELATIONS

As suggested in chapter 1, means and variances are descriptors of population behaviors useful primarily as initial approximations to biological phenomena. Similarly, the linear covariances and correlations described above are useful first approximations to the actual relationships among population measures. True linearity of relationships among traits should not be assumed; it is a rare exception in the real world. Few traits are linearly related and, even when measuring the same trait in different relatives, the conditions of the testing or ages of the relatives may differ. Hence, the covariance between say offspring and parent is often not one-half of a genetic variance, but is one-half of a genetic covariance. This relationship, however, may not be linear. In such cases, curvilinear regression adjustments are often made to linearize the parameters of the model and multiple regression coefficients used for each trait. These introduce no new theoretical problems, are useful second approximations to reality, and are about all that can now be done without using nonlinear mathematics.

If we can then assume that breeding can be efficiently performed for a given set of values, the central problem in forest tree breeding is defining and measuring value when the traits themselves are complicated by environmental interactions and their economic effects are nonlinear. Several forms of nonlinearities are fairly common in forestry. Some such problems cannot be linearized by logarithmic or polynomial transformations. Slightly more difficult to handle are cases in which discontinuities exist in the relationship between physical measures and value, such as between stem diameter and stem value when it jumps from pulp size to pole, saw-log, and veneer-log sizes. In addition, degree of past resistance may exhibit a relatively flat value function until some minimal levels are reached, after which a linear function may exist until high resistance levels are reached and increments in resistance add little to final crop value, especially if some natural thinning is expected. In many such cases, approximate value functions may be assigned and linearized, even when multiple discontinuities exist. Only slightly more complicated and difficult are those cases in which the value of one trait depends to some extent on the value of other traits—when trait values are interdependent. The joint value function then requires some iterative evaluation as, for example, when both volume growth and wood quality are interdependent and both depend on survival and pest resistance levels. For example, pest resistance may be of relatively little value at low-growth levels but can assume an exponential value function at high-yield levels. Similarly, increasing wood yield under high risk of mortality may be of low value until mortality rates can be dropped enough to warrant investment in growth improvement. However, even such joint value functions, with various points and lines of discontinuities and with various forms of curvature, present problems only in locating a direction for maximizing value

on an uneven, but known, surface. Thus, on some nonlinear surface of value of possible trait combinations, the breeder can seek to determine the direction cosines of the line which would give him maximum gain. This line may be either in the direction of maximum value gain for very small changes in trait values, for example, the gradient, or in the direction of the optimum trait combination which may require a long-term breeding effort and more than what maximum immediate gains could promise. In the case of a truly linear value function, these lines are identical and the index coefficients which are derived by traditional methods are the direction numbers of the planes of equal value which are perpendicular to the gradient. Determinations of the direction numbers of planes of equal value or the associated direction cosines of the line normal to those planes are essentially similar operations as one can be determined from the other. Therefore, other than forcing one to work harder to determine the value function and to change the direction of maximizing value gain according to the present mean value of the population, the above nonlinearities cause no theoretical problems in breeding in whatever direction is determined to be optimum.

A different class of problems are generated when, in addition to any nonlinearities, the actual value function is not precisely known. In this case, errors in estimating true values can cause some trait values to be relatively decreased in breeding value if they happen to be more sensitive to the uncertainties of future values. This is a particularly acute problem in forest tree breeding where the time interval for single generations from selection through breeding, planting, and harvest would often involve 20 to 60 years. While these times can be expected to decrease, the uncertainty factor will always be present. Since breeding operations for particular hybrids or with any recurrent selection objectives obviously require projections of value into an unknown future, only temporary gains can be achieved if breeding objectives change within generations. Long-term gains are necessarily limited to relatively few traits of persistent value such as survival, growth under wide site variations, and perhaps some pest resistances. Short-term objectives can include more traits but even then must include evaluations based on unknown technologies applied at various stages of forestry between planting and final conversion to economic return. Any of the stages, including silviculture, harvesting, mill technologies, and market variables, which can change within a generation between seed production and value conversion, can be more quickly altered than breeders can affect their product values (Namkoong and others 1966). In other breeding programs with shorter breeding cycles, such as with dairy cows bred for high butter fat content in milk or with tobacco bred for nicotine content, marked changes required strenuous changes in modifying otherwise adapted breeds and lines. Thus, the choice of traits for breeding when economic value uncertainty exists can drastically affect the value of the whole breeding operation, and

considerable care is required in choosing breeding for long-term gains on some traits and short-term (one generation or more) gains on others. Foresters will have to be far more conservative, since rapid breeding generations to alter breeds will not generally be possible (Stern 1972).

If some uncertainty is involved in determining a value function for a given generation of selection and breeding, the optimum strategy may not be to maximize an expected average gain, which may be known only with high error of estimate. Clearly, if the uncertainty were such that the error on predicted values was very narrowly distributed around a mean, then one might wish to treat the case as a deterministic one. In what might be the more common situation, however, a relatively high variance on the value function exists, and an estimated average function may have too high an error to attempt a definition of optimum breeding direction cosines. An alternative strategy is to determine simultaneous confidence limits within which, at given probabilities, the optimal direction exists. Breeding evaluations may then be made on the assumption that a maximum error and an associated minimal gain can be made for a given range in direction cosines. Alternatively, a breeder may wish to minimize the probability of certain errors occurring, such as some limit on misdirection, and may instead choose direction cosines with minimal gain objectives.

Under conditions of high uncertainty, such that the error distribution on the predicted value function is too large to reasonably derive an estimated mean with any reasonable probability of accuracy, the concept of maximizing expected value may have to be totally discarded. In such cases, it may be possible to describe several value functions for the combined traits, no one of which is any more likely than the rest to reflect the actual value function at harvest time. Then, maximizing a minimum expected gain may be a more reasonable, if conservative, strategy to follow since it would guarantee that certain minimal values can be achieved regardless of which value function actually exists at harvest. The optimum trait direction for the breeder to follow is thus defined as that which imparts the highest minimal gain which can be achieved. As an oversimplified example, consider only one trait, wood specific gravity, which is highly correlated with cell-wall thickness, and assume that one of four situations illustrated in figure 11 may occur:

(1) Thin-walled fibers become very valuable and wood of low specific gravity has high value.
(2) Thin-walled fibers are of some value but the loss in fiber yield of low-specific-gravity wood almost offsets the value of the wood-quality gain.
(3) Cell-wall thickness does not affect value and the increased fiber yield of high-specific-gravity wood increases its value.

(4) Thick-walled fibers are of some value and accentuate the increased value of high-specific-gravity wood.

If these four adequately define all value functions, the X marks in figure 11 represent the extreme points of the possible solution set. The value of specific gravity which maximizes the minimum gain is indicated by a circle and would be the optimum point toward which the population should move. It can be seen that discontinuities in the value functions would cause a problem in arriving at a solution.

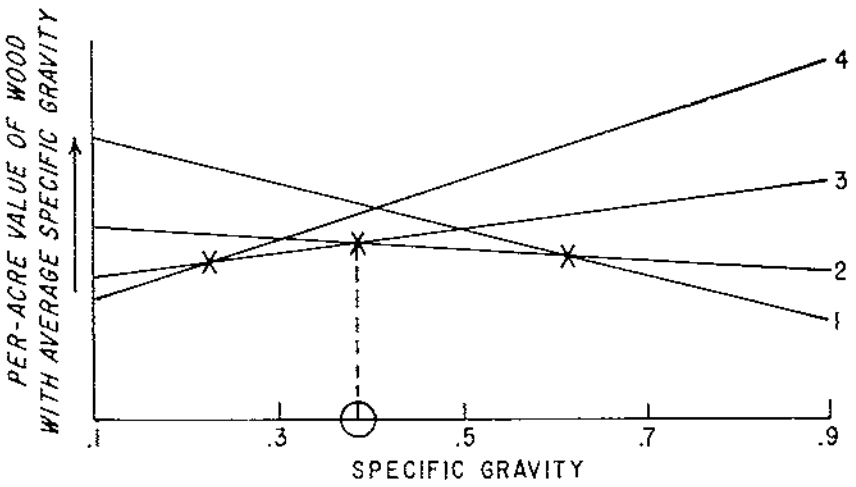


Figure 11.—Maximin problem for values of specific gravity in four possible situations: (1) strongly decreasing value, (2) moderately decreasing value, (3) moderately increasing value, and (4) strongly increasing value.

In any real situation, the economic functions would be more complicated and involve more variables, but they could be easily solved since the solution of maximin games can be found by standard linear programming (Hadley 1964). Difficulties with continuous, nonlinearizable functions remain (Owen 1968), but with modern computers good linear approximations for small intervals can be made, and the expanded set of restrictions can easily be handled. All problems of discontinuities and boundary values have been eliminated. Therefore, solutions for such linearizable functions in these essentially two-person, zero-sum games can always be found, and an optimum point or direction defined. Therefore, a direction for moving the breeding population can be defined even in these situations.

Therefore, even facing uncertainty of a high degree, the breeder can determine optimum trait combinations and can evaluate trees with respect to the directions so determined. While it is also clear that, due to prediction problems and problems of estimating genetic and phenotypic means and variances, only a restricted number of traits should be included in breeding programs, it is also true that

even mild selection pressures can be beneficial in keeping favorable alleles in population. Such alleles can be useful in later generations of breeding or in special breeding orchards and would be useful to maintain even if their immediate usefulness is limited and of uncertain values. In these cases, once a basic set of traits and their optimal direction cosines, or associated direction numbers, are determined, restricted indices may be additionally useful (Tallis 1962).

GENOTYPE X ENVIRONMENT INTERACTION

Economic and estimation uncertainties are not the only factors which make the direct application of breeding theory especially complicated in forestry. Changes in the environments of forests are also becoming more rapid and widespread through the direct efforts of silviculturists as well as through the accidental impact of human and other influences. The economics of land use with forestry alternatives clearly depend on forest values which are in turn affected by genetic potential values and the control exercised by breeders. At the same time, values clearly depend on how society controls forest operations within the general economic system. Under an intensive system of planning, genetic control of forest characteristics and values can be one management control variable for use in conjunction with silviculture and other technologies capable of affecting forest values. If the environment can be predicted to change, then coordinated changes in culture and genetic composition of forests can give extra benefits if genotype - environment interactions exist. For example, genotypes may compose an ordered set of entries in a field trial in which several environmental states may be sampled and relative performances estimated. Regardless of whether either genotypes or environments have any average overall effects, various combinations of particular genotypes on particular environments positively or negatively depart from their average performance due to the special reaction of one to the other in the combination. If the environments can be ordered, then differences among genotypes' response functions other than general mean or simple scale differences may be considered to be interaction effects for which some nonlinear functions may be fit and found to vary among genotypes. For many plant species, relationships among genotypes seem to be nearly linear with respect to stability measures or to measures of relative performance of genotypes over several environments (Freeman and Perkins 1971). However, this need not be true for the range of site variables which forest trees must face. In many forestry cases, the entire multivariate response surface to multiple environmental gradients should be estimated along with differences between genotypes so that specific sets of genotypes can be recommended for specific classes of environmental factor combinations as previously described. Regression analysis would indicate the extent to which regularity of response can be predicted. It could also indicate the distribution of the residual errors, ac-

cording to various subdivisions of regression sums of squares and interactions, and hence the linearity of genotype × environment interactions (Eberhart and Russell 1966; Freeman and Perkins 1971). In such cases, the entire package of genotypes with environmental specifications might be more easily developed, as has been the experience with other crops which are developed for given fertility and water regimes (Robinson and Moll 1959). For forest trees, such factors as planting-site preparation, spacing, fertility, and growing region for general climate or soils might be sufficiently controllable that testing for responses to some specified standardized conditions might give worthwhile benefits. However, very close specification of environments shrinks the area within which the breeding population is ideally suited, and total gain may suffer if the breeding program must sustain too many populations adapted to special sites instead of a good overall average adaptability. An alternative goal of breeding may then be to select for lack of response to a wide variety of conditions and for good average performance, as suggested by Finlay and Wilkinson (1963). In addition to any average value benefits, this system may have a greater uniformity of response to uncontrolled site variation, can have value as a more consistent or reliable planning factor, and can increase forest values by increasing their uniformity. For some plant species, however, highest average yield is associated with instability (Tai 1971), and different environments may induce different kinds of genetic variance to be displayed by the same organisms (Perkins and Jinks 1968). Then, uniformity of performance is most readily produced by different genotypic mixtures.

It would clearly be advantageous to use genotypes which uniformly perform best in all sites. There is some indication that such phenomena may exist, but to the extent that such uniform goodness does not exist, some compromise is required between maximum adaptedness with special breeds and limits to the number of special populations which can be specified for geographic areas or other site restrictions. Thus, preliminary surveys of the dimensions and extent of environmental variations are desirable to test the form and importance of these genotype × environment interactions as well as any changes in the genetic variances (King 1965; Ledig 1970; Squillace 1970). Descriptions and classifications of genotypes according to similarity of responses can be useful to determine the existence of subsets of environmental variables and subsets of genotypes (Hanson 1970). Regression types of analyses of genotypes on environments can greatly benefit both the analysis of forms of interactions and the practical use in breeding (Perkins and Jinks 1968; Freeman and Perkins 1971).

Differences in degrees of stability and response to favorable and unfavorable environments have been found for slash pine (Snyder and Allen 1971). The only major difference between these methods and standard regression analyses is that the environments are scaled according to the average performance of the genotypes and

not by known, measurable variables. Freeman and Perkins (1971) construct an ANOVA for t genotypes (with t–1 degrees of freedom), s environments (with s–1 df), and an interaction (with (t–1) (s–1) df). The s–1 df for environments are partitioned into 1 df for a sum of squares due to an average linear regression and s–2 df for the remainder. The interaction is partitioned into sums of squares due to genotypic differences in their linear regressions (with t–1 df) and a remainder (with (t–1) (s–2) df). Eberhart and Russell's (1966) partitioning is slightly different: the s–2 df of the remainder from environmental regression and the (t–1) (s–2) df of interaction remainder are used to construct a sum of squares of deviations of environments from linearity with s–2 df for each genotype. Both methods, however, are essentially similar in seeking linear regressions and ANOVA's for testing those models.

For more general analyses of nonlinear and multiparameter responses, the same breakdown of degrees of freedom for multiple regression can be followed. For such cases, unbalanced designs with several environmental measures may find greater use, especially for forest trees with complicated response patterns. These experiments would require a careful allocation of possible treatments or environmental degrees of freedom into specified treatment combinations.

While special treatment combinations can be most efficiently designed for purposes of regression analysis as above, or in more complicated response surface estimations (Box and Lucas 1959), it will seldom be possible to test all genotypes on all site and silvicultural variable levels. The location of replicated treatment combinations around maximum response and maximum curvature zones, as previously suggested, may not be feasible. More often, it will be necessary to specify standard environments which can test the range of responses of interest or to specify indicator genotypes which provide some idea of the existence and form of genetic interactions for given site variations. In addition to the effects on selection, the existence of interactions causes bias in estimating genetic variances of single-location experiments as previously discussed. However, even if the component of variance due to interactions is small, its effect on selection can be significant (King 1965). As a minimum program for testing, at least an average site would have to be sampled by all genotypes. More generally, to the extent that seedlings are available for testing, environmental factor combinations representing the breeder's best guesses on major site subdivisions should be sampled. If major genotype × site interactions exist, then selections can be made for specific sites for special breeds or for a single, generally useful breed.

Site sampling should follow the general principles of Box and Lucas (1959) to span as many site dimensions of present and future utility as possible. If balanced designs incorporating all genotypes across all selected sites cannot be installed, then unbal-

anced designs will find some utility. Recognizing that some interactions of special genotype-site combinations will not be observed, and if mean estimation is still of major importance, then partially balanced factorials of genotypes × sites can be profitably used. These may either involve a single, completely connected, partial factorial in which some overlap between genotypes and site combinations exists so that complete least squares of main effects can at least be determined, or some subblocking can be used instead. If genotypes are subblocked to gain experimental (mean estimation) efficiency, some loss occurs in that genotypes in different subblocks cannot be compared. The general design and analysis problems considered in chapter 7 are directly applicable to these problems. Subblocking can be arranged in several ways to avoid putting all genotypes on all sites. The families may be completely separated into mutually exclusive sets and each tested on one set of sites which are also subdivided into mutually exclusive subsets. In such a pattern, no information on sites or genotypes in separate sets can be recovered. By making partial overlaps of families among sites, comparisons of families in different sets can be made, and different average site effects can also be compared. This can be arranged by using certain families as common checks on all sites or alternatively, on some sites, using all families as a basis for adjusting other sites and genotypic combinations. In complete blocking designs, using a series of different, overlapping families to connect sequences of sites may also be used. For these testing purposes, various partially balanced incomplete blocking arrangements can be used. They will be the same for comparing genotypic means as for any other kinds of treatment means. In particular, the use of partial balance with respect to genetic blocking or replication within sites can help to preserve balance with respect to major site variables (Schutz and Cockerham 1962). The balanced designs are distinctly preferable for testing means, for determining regressions on ordered variables, and for security of the material and analysis in future years. If necessary, however, designing partial balance can offer experimental efficiency of scarce materials.

For purposes of estimating the general size and form of genotype × interactions, unbalance is not as great a problem in the design and analysis of experiments for estimating variance components or regression-type responses as it is for testing. At least initially, however, installing such experiments will be quite difficult. Experimental geneticists may be largely limited to providing some replicated genetic identities to use as split-plots in large plot silvicultural experiments. Locational or site differences have been thought to be more significant than year differences in annual crops because wide site variations can be easily sampled (Rojas and Sprague 1952), but persistent site × year × genotype differences also exist. The form of interaction effects is still highly variable and can affect experimental design as well as breeding procedures (Comstock and Moll 1963; Hanson 1964). Within the

set of factors and variations in each that might be considered, factors which change within the breeding cycle are generally considered to be part of the uncontrolled variations contributing to error variance. Hence, trees with good average response or lack of response to those factors might be given added value for uniformity of response to noise variations. Stability over these and unpredictable future variations is then an important positive value and can be parameterized and tested (Hanson 1970). Similar methods as developed by Finlay and Wilkinson (1963) and Eberhart and Russell (1966) have been found useful in studying provenance variations in Jack pine (Morgenstern and Teich 1969). Year-to-year climatic variations may often be linear, and therefore easily handled where breeds have to be planted over all of the years of seed production regardless of the climatic variations. Major yearly differences, however, may involve drought, fertility, or early planting effects of major significance which may be adjustable and hence, if properly sampled, can be a major site variant controllable by silvical measures.

The genotype × site interactions that occur among replicates within planting zones cause an additional problem in testing among a large number of entries within replication blocks. Since the use of check varieties or genotype entries is limited by their own interaction potentials, Schutz and Cockerham (1962) recommend the use of blocks of genotypes confounded in replication blocks, for selection as well as estimation experiments in preference to replications of complete blocks of entries or checks. They find it more efficient to subblock entries to reduce within-replication error at the expense of recovering interblock information and selection among entries in different blocks.

Sampling site variations by splitting major-site plots with genotypically distinct subplot differences may be a satisfactory compromise between independent experimentation and economic necessity. Standard split-plot analyses may then be performed within the treatment level combinations afforded by the silviculturist. In addition, any family structure such as male full-sibs within female half-sibs would allow for some further breakdown of the genotype × environment interactions into additive and dominance gene effects. For example, consider two environmental variables V and W. Let locations, replications in locations, families, and individuals in families be considered as random samples from infinite populations and the cultural treatments be considered fixed effects.

The linear model yield equation is:

$$Y=\mu+L_l+R_{r/l}+V_v+W_w+(VW)_{vw}+(VL)_{vl}+(WL)_{wl}$$
$$+(VWL)_{vwl}+e_a+F_f+(FL)_{fl}+(FV)_{fv}+(FW)_{fw}$$
$$+(FVW)_{fvw}+(FVL)_{fvl}+(FVWL)_{fvwl}+e_b+e_w$$

where

L_l = lth location effect,
$R_{r/l}$ = rth replication effect in the lth location,
V_v = vth level effect of treatment V,
W_w = wth level effect of treatment W,
$(VW)_{vw}$ = interaction deviation of the vth with wth levels of treatments V and W,
$(VL)_{vl}$ = interaction deviation of the vth level of treatment V with location l,
$(WL)_{wl}$ = interaction deviation of the wth level of treatment W with location l,
$(VWL)_{vwl}$ = interaction deviation of the vth level of treatment V with the wth level of treatment W with location l,
e_a = major plot error,
F_f = fth family effect,
$(FL)_{fl}$ = interaction of the fth family with the lth location,
$(FV)_{fv}$ = interaction of the fth family with the vth level of treatment V,
$(FW)_{fw}$ = interaction of the fth family with the wth level of treatment W,
$(FVW)_{fvw}$ = interaction of the fth family with the vth and the wth levels of treatments of V and W,
$(FVL)_{fvl}$ = interaction of the fth family with the vth level of treatment V and location l,
$(FWL)_{fwl}$ = interaction of the fth family with the wth level of treatment W and location l,
$(FVWL)_{fvwl}$ = interaction of the fth family with the vth and wth levels of treatments V and W and location l,
e_b = minor (family) plot error,
e_w = within minor plot error.

The location and interactions can then be further partitioned into linear and nonlinear regression factors, with their associated sums of squares.

If the family members are grouped into plots, the analysis may be made as on a split-split-plot, with the replicate-location plots designated as major plots, the V by W treatment factorial as subplots, and families as sub-subplots. If, in addition, reciprocals or closer relatives are grouped into compact family blocks, a further degree of hierarchy in genetic and error components would exist. In the following, we assume random location of all family plots. The fixed nature of the cultural treatments allows certain simplifications to be made in the tabulation of expected mean squares. From the analysis of variance presented in table 2, estimates of the variance components can be easily derived. A slightly different but essentially similar analysis would follow if the family members were randomly placed within the treatment plots.

Table 2.—Analysis of variance for families

Source of variation	df	Expected mean squares
Locations	l–1	
Replications in locations	$l(r$–1)	
V	v–1	
W	w–1	
$V\times W$	(v–1) (w–1)	
$V\times L$	(v–1) (l–1)	
$W\times L$	(w–1) (l–1)	
$V\times W\times L$	(v–1) (w–1) (l–1)	
Major plot error	(vs–1)$l(r$–1)	
Families	(f–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rv\sigma_{fwpl}^2+rvw\sigma_{fpl}^2+rvwp\sigma_{fl}^2+rvwpl\sigma_f^2$
Families×locations	(f–1) (l–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rv\sigma_{fwpl}^2$
Families×V	(f–1) (v–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rw\sigma_{fvpl}^2+rpw\sigma_{fvl}^2+rplw\sigma_{fv}^2$
Families×W	(f–1) (w–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rv\sigma_{fwpl}^2+rpv\sigma_{fwl}^2+rplv\sigma_{fw}^2$
Families×$V\times W$	(f–1) (v–1) (w–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rp\sigma_{fvwl}^2+rpl\sigma_{fvw}^2$
Families×$V\times L$	(f–1) (v–1) (l–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rpw\sigma_{fvl}^2$
Families×$W\times L$	(f–1) (w–1) (l–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rpv\sigma_{fwl}^2$
Families×$V\times W\times L$	(f–1) (v–1) (w–1) (l–1)	$\sigma_e^2+r\sigma_{fwpl}^2+rp\sigma_{fvwl}^2$
Minor plot error	(f–1) (vw–1) (r–1)	$\sigma_w^2+p\sigma_e^2=\sigma_s^2$
Within plot error	($fvwrl$) (p–1)	σ_w^2

If the sampling of families was structured to reflect a hypothesized structuring of the wild population (i.e., racial hierarchies), then the σ_f^2 may be broken down into these components. If the families were structured as to types of relatives, the σ_f^2 may be interpreted in terms of genetic variances. For instance, if the families were unrelated half-sibs, σ_f^2 is the covariance of half-sibs which, under assumptions of no inbreeding of progenies and insignificant epistatic variances, is one-quarter of the additive genetic variance (σ_A^2). The single-locus genetic model leading to the latter covariance interpretation is:

$$Y=\mu+\alpha_{y_m}+\alpha_{y_f}+\delta_{y_m y_f}+e_i+(\alpha e)_{y_m i}+(\alpha e)_{y_f i}+(\delta e)_{y_m y_f i},$$

where α_{y_i} = additive effect of the jth parental allele, $j=m, f$,
$\alpha_{y_m y_f}$ = dominance effect of the m with f parental alleles,
e_i = a general environment effect,
$(\alpha e)_{y_j i}$ = interaction of the ith environment with the additive effect of the jth allele,
$(\delta e)_{y_m y_f i}$ = dominance × environment interaction effect.

The covariances of different individuals under the above assumptions are:

Cov (maternal half-sibs over treatments) $=\sigma_f^2=\frac{1}{4}\sigma_A^2$,

Cov (individuals with the same environment over families) $=\sigma_E^2$, and

Cov (individuals with the same environment and maternal parent) $=\sigma_f^2+\sigma_E^2+\sigma_{fE}^2=\frac{1}{4}\sigma_A^2+\sigma_E^2+\frac{1}{4}\sigma_{AE}^2$, $\sigma_{fE}^2=\frac{1}{4}\sigma_{AE}^2$,

where σ_{AE}^2 is the interaction of the environmental factor with additive genetic effects. If the breeding design had been a hierarchy of females (f) and males (m) within females, the analysis of variance (table 2) would be exactly the same, but with a subdivision of the family and the family interactions sums of squares. Such an analysis is presented in table 3 with the major plot analysis omitted. If σ_{AE}^2 is high, it is included in the estimates of σ_A^2 made from analyses within location sites, and is properly included as part of the usable genetic variance for breeding for specific sites. For general site breeding, however, the σ_A^2 estimated over all sites is the only usable genetic variance for breeding purposes. However, it should again be noted that these are statistical averages. The indications of high σ_{AE}^2 will generally warrant more detailed investigation of the forms of environmental response displayed by the sampled genotypes and the existence of different amounts of genetic variance in different environments.

Again, the environmental and interaction components can be further partitioned into regression parameters of the interaction model.

If we again assume the linear model with a generalized environment effect, the covariances are:

Cov (maternal half-sibs over environments) $=\sigma_f^2=\frac{1}{4}\sigma_A^2$,

Cov (full-sibs over environments) $=\sigma_{m/f}^2+\sigma_f^2=\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2$, $\sigma_{m/f}^2=\frac{1}{4}(\sigma_A^2+\sigma_D^2)$,

Cov (individuals in same environment over families) $=\sigma_E^2$,

Cov (individuals with same environment and maternal parent over males) $=\sigma_{fE}^2+\sigma_E^2+\sigma_f^2=\frac{1}{4}\sigma_{AE}^2+\sigma_E^2+\frac{1}{4}\sigma_A^2$, $\sigma_{fE}^2=\frac{1}{4}\sigma_{AE}^2$,

Cov (individuals with same environment and maternal and paternal parents) $=\sigma_{Em/f}^2+\sigma_{fE}^2+\sigma_E^2+\sigma_{m/f}^2+\sigma_f^2$

$=\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2+\sigma_E^2+\frac{1}{2}\sigma_{AE}^2+\frac{1}{4}\sigma_{DE}^2$, $\sigma_{Em/f}^2=\frac{1}{4}(\sigma_{AE}^2+\sigma_{DE}^2)$.

Since these variances can be expected to be different, their existence implies error heterogeneities that will cause testing error rates to be imprecisely determined. If the dominance-by-environment interaction is high, then breeding would have to in-

Table 3.—Analysis of variance for nested sibs (major plots omitted)

Source of variance	df	Expected mean squares
Females	$d-1$	$\sigma^2_{e_f}+rpvw\sigma^2_{lm/f}+rpvws\sigma^2_{lf}+rpvwl\sigma^2_{m/f}+rpvwlm\sigma^2_{f}$
Males/females	$d(s-1)$	$\sigma^2_{e_m}+rpvw\sigma^2_{lm/f}+rpvwl\sigma^2_{m/f}$
Females×locations	$(d-1)(l-1)$	$\sigma^2_{e_d}+rpvw\sigma^2_{lm/f}+rpvws\sigma^2_{lf}$
Males/females×locations	$d(s-1)(l-1)$	$\sigma^2_{e_f}+rpvw\sigma^2_{lm/f}$
Females×V	$(d-1)(v-1)$	$\sigma^2_{e_f}+rpw\sigma^2_{vlm/f}+rpws\sigma^2_{vlf}+rplw\sigma^2_{vm/f}+rplwm\sigma^2_{vf}$
Males/females×V	$d(s-1)(v-1)$	$\sigma^2_{e_s}+rpw\sigma^2_{vlm/f}+rplw\sigma^2_{vm/f}$
Females×V×locations	$(d-1)(v-1)(l-1)$	$\sigma^2_{e_f}+rpw\sigma^2_{vlm/f}+rpws\sigma^2_{vlf}$
Males/females×V×locations	$d(s-1)(v-1)(l-1)$	$\sigma^2_{e_s}+rpw\sigma^2_{vlm/f}$
Females×W	$(d-1)(w-1)$	$\sigma^2_{e_f}+rpr\sigma^2_{wlm/f}+rpvs\sigma^2_{wlf}+rplv\sigma^2_{wm/f}+rplvm\sigma^2_{wf}$
Males/females×W	$d(s-1)(w-1)$	$\sigma^2_{e_s}+rpv\sigma^2_{wlm/f}+rplv\sigma^2_{wm/f}$
Females×W×locations	$(d-1)(w-1)(l-1)$	$\sigma^2_{e_f}+rpv\sigma^2_{wlm/f}+rpvs\sigma^2_{wlf}$
Males/females×W×locations	$d(s-1)(w-1)(l-1)$	$\sigma^2_{e_s}+rpv\sigma^2_{wlm/f}$
Females×V×W	$(d-1)(v-1)(w-1)$	$\sigma^2_{e_f}+rp\sigma^2_{vwlm/f}+rps\sigma^2_{vwlf}+rpl\sigma^2_{vwm/f}+rplm\sigma^2_{vwf}$
Males/females×V×W	$d(s-1)(v-1)(w-1)$	$\sigma^2_{e_s}+rp\sigma^2_{vwlm/f}+rpl\sigma^2_{vwm/f}$
Females×V×W×locations	$(d-1)(v-1)(w-1)(l-1)$	$\sigma^2_{e_f}+rp\sigma^2_{vwlm/f}+rps\sigma^2_{vwlf}$
Males/females×V×W×locations	$d(s-1)(v-1)(w-1)(l-1)$	$\sigma^2_{e_s}+rp\sigma^2_{vwlm/f}$
Female plot error	$(d-1)(vw-1)l(r-1)$	$\sigma^2_{e_s}+ps\sigma^2_{e_f}=\sigma^2_{ef}$
Male/female plot error	$d(s-1)(vw-1)l(r-1)$	$\sigma^2_{w}+p\sigma^2_{e_m}=\sigma^2_{e}$
Within male	$(dwvwlr)(p-1)$	σ^2_{w}

clude specific combinations of parents or hybrid breeding systems to utilize their responses to particular environments. In contrast, general combining ability breeding methods would be applied to use additive genetic variance-by-environment interactions.

Partially balanced designs would, of course, have different analyses. The above completely balanced analysis is exceedingly cumbersome and the variance components estimates would carry high errors. Unbalanced designs could estimate the same components with far greater efficiency. Designs for testing can be amalgamated with designs for variance component estimation through Gaylor and Anderson's (1960) L-shaped designs. All genotypes are represented on a subset of sites, while all sites are sampled with a subset of genotypes. In addition, where even the basic subset of sites cannot be sampled by all genotypes, partially balanced factorials or blocked factorials should be used.

To the extent that the environmental variables are ordered and have some regular form of effect on the yield variables, such as a polynomial function, the sum of squares due to the interactions can be further partitioned into such effects as linear or quadratic interactions.

Not only do genotypes vary in their average responses to environmental variations, but the degree and form of that variation are also genetically determined. Some genotypes are highly variable in their response to a set of different environments while others with the same average performance and grown on the same set of environments are more homeostatic (Butcher and others 1972). It has been frequently observed that the degree of inbreeding also can affect the general level of response to environmental differences; the higher the inbreeding, the greater the variation as measured by the interaction (Allard and Bradshaw 1964). However, there is no evidence that the type of gene action involved in average performance determines the degree of the interaction. All forms of gene effects have been associated with genotype × environment interactions (Robinson and Moll 1959).

The utilization of data on the response form in either breeding for high response to controllable site factors or uniformity of response to any kind of site factor depends on how uniformity is evaluated. If it is of positive value as implied by Hanson (1970), its value should be considered positive in evaluating trees for use in an uncertain future environment. However, if the future distribution of environments is known, then the mean value of the population will be the determining factor and uniformity of response will not affect value except as it can increase mean tree value directly. If the analysis of value is made under uncertainty and a value function is taken as some form of a minimax strategy, then uniformity has some incremental value in that it would give high minimal values to those trees with high uniformity.

In all of these experiments it is assumed that plot arrangements permit a complete evaluation of trees or families over as large a part of their life cycle as necessary in what might be understood

as reasonable competitive and spacing environments. However, it is clear that these factors are not simply defined, nor are there uniform conditions for suggesting that standard environments can be found. The effects of missing trees and missing plots are problems of experimentation which induce estimation problems and error heterogeneities and which affect test error rates. However, since these are common statistical problems not peculiar to forestry, the reader is referred to texts on experimental design and analysis.

Some designs to test spacing effects have been proposed (Namkoong 1966a), as have plot arrangements that maintain regular spacing for each family through successive thinnings (W. J. Libby, personal communication). But little work has thus far been attempted on intergenotypic competition effects among families of the same species or provenance.

COMPETITION

Competition effects among genotypes, as distinct from spacing effects, require that genotypic interactions be analyzed and that performance be defined in terms of other genotypes. At the interspecies level, genetic competition effects are clearly significant in forming plant and animal communities. At the finer levels of restrictions on growing space, competition control has been a major silvicultural tool. While interspecies competition may occur throughout the life cycle and environmental variables may affect all factors of a tree's growth and development, spacing effects themselves involve crown, stem, and root in complicated interactions (DeWit 1960). However, forest ecologists are still untangling the interrelationships among trees caused by proximity and competition for limited space, other special competitive interactions among specific genotypes caused by chemical or special time-dependent effects also have to be studied. In a series of studies on rice, Sakai (1955, 1965) defined intergenotypic competition as any departure in plant performance exhibited when a plant is competing against other genotypes rather than in a pure stand. Thus, ordered or unordered sets of competing genotypes at the same spacing and constant in other environmental factors have been noted to have suppressing or enhancing effects relative to pure-stand performance, and parameters and genetic variations with respect to the special competition environment have been described (Sakai 1961). Extending this work to forest trees, Sakai and Mukaide (1967) and Sakai and others (1968) clearly note that these special effects can substantially increase the total variance in mixed genotype stands over pure stands. Hence, such effects may be of considerable importance in controlling values and forest uniformity. Not only may uniformity be increased by selecting similar acting genotypes, but growth may be enhanced by selecting genotypes which somehow complement

and mutually benefit their selected neighbors. While too little is yet known of how the interactions operate and whether they are strongly dependent on other site factors of significance, further investigation is clearly warranted in forest trees (Adams and others 1973). Huhn (1969, 1970a, 1970b) defined the effect of genotype X on the growth of another genotype Y as the competitive influence of tree X (W_x) on the tree Y which has a competitive ability F_y. His projections suggest that genetic variances in both types of competitive effects can stabilize in forests. Similarly, Mather (1969) indicates that competition can have stabilizing effects on polymorphisms in natural environments. While his definitions are different from Sakai's, the results suggest a useful parameterization and a method for estimating the significance of this kind of competition in forests.

If competition can have as significant an effect on plant breeding, as suggested by Allard and Adams (1969), such as to force a complete reevaluation of plant breeding methods due to the peculiar stabilities that natural stands may have generated, then tree breeders can start developing their populations with the precautions of including wide growth variations in competitive tests using grouped tests, as suggested by Schutz and Brim (1971). In fact, the expansion of models to include competition and its spacing and density-dependent effects have generated renewed concern with the classical concepts of population genetics and their ability to account for the existing variations in natural population (Mather 1969; Ayala 1971). If it can clearly have profound effect on natural evolution, then for foresters starting with relatively natural populations and selecting for increasingly cultivated environments, the demand is clear for experimentation on these effects.

While most of the above studies have been made on mixtures of pure breeding lines, genetic segregation and intermating among competitors in breeding populations can also be studied as has Huhn (1970a). In terms of how selection affects the composition of competitive interactions if the genes which affect competitive ability are quantitatively inherited, Griffing (1967) has developed a model which is parallel to his classical model of selection effects. The allelic combinations which result from truncation selection and reconstituting a breeding population by crosses among that combination of selected individuals is traced. In definitions similar to Huhn's (1969), Griffing assumes that alleles have both direct effects on their own genotype's growth as well as associate effects on those with which it competes. Thus, instead of a simple $d_{ij}=\alpha_i+\alpha_j+\delta_{ij}$ genetic model, he defines ${}_{i_1 j_1}d_{i_2 j_2}$ as being the genetic value of individual 1 in the presence of individual 2, and hence for populations of size 2, the array of genotypes with allelic frequencies p_i and p_j is:

$$\sum_{i,j} p_i p_j (A_i A_j) \times \sum_{i,j} p_i p_j (A_i A_j) = \sum p_{i_1} p_{j_1} p_{i_2} p_{j_2} (A_{i_1} A_{j_1} A_{i_2} A_{j_2})$$

Genotype value $A_{i_1}A_{j_1}$ as expressed in $(A_{i_1}A_{j_1}, A_{i_2}A_{j_2})$ is ${}_{i_1j_1}d_{i_2j_2}$,

$$
\begin{aligned}
{}_{i_1j_1}d_{i_2j_2} = {}_{d}a_{i_1} + {}_{d}a_{j_1} + {}_{d}\delta_{i_1j_1} + {}_{a}a_{i_2} + {}_{a}a_{j_2} + {}_{a}\delta_{i_2j_2} + {}_{da}(aa)_{i_1i_2} \\
+ {}_{da}(aa)_{i_2j_2} + {}_{da}(aa)_{j_1i_2} + {}_{da}(aa)_{j_1j_2} + {}_{da}(a\delta)_{i_1i_2j_2} \\
+ {}_{da}(a\delta)_{j_1i_2j_2} + {}_{da}(\delta a)_{i_1j_1i_2} + {}_{da}(\delta a)_{i_1j_1j_2} + {}_{da}(\delta\delta)_{i_1j_1i_2j_2},
\end{aligned}
$$

where ${}_{d}a_{i_1}$ = direct additive effect of allele A_{i_1},

${}_{d}\delta_{i_1j_1}$ = direct dominance effect of $A_{i_1}A_{j_1}$,

${}_{a}a_{i_2}$ = associate additive effect of A_{i_2} as measured on $A_{i_1}A_{j_1}$,

${}_{a}\delta_{i_2j_2}$ = associate dominance effect of $A_{i_2}A_{j_2}$ as measured on $A_{i_1}A_{j_1}$,

${}_{da}(aa)_{i_1i_2}$ = additive × additive interaction effect between direct allele A_{i_1} and associate allele A_{i_2},

${}_{da}(a\delta)_{i_1i_2j_2}$ = additive × dominance interaction between direct allele A_{i_1} and associate genotype $A_{i_2}A_{j_2}$,

${}_{da}(\delta a)_{i_1j_1i_2}$ = dominance × additive interaction effect between direct genotype $A_{i_1}A_{j_1}$ and associate allele A_{i_2},

${}_{da}(\delta\delta)_{i_1j_1i_2j_2}$ = dominance × dominance interaction effect between direct genotype $A_{i_1}A_{j_1}$ and associate genotype $A_{i_2}A_{j_2}$.

These interaction effects are not epistatic effects but average intergenotypic effects due to allelic effects in the sense of affecting competitive phenotypes. The total genotypic variance for ${}_{i_1j_1}d_{i_2j_2}$ is:

$$
{}_{G}\sigma^2 = {}_{d}\sigma_A^2 + {}_{d}\sigma_D^2 + {}_{a}\sigma_A^2 + {}_{a}\sigma_D^2 + {}_{da}\sigma_{AA}^2 + {}_{da}\sigma_{AD}^2 + {}_{da}\sigma_{DA}^2 + {}_{da}\sigma_{DD}^2,
$$

and the covariance between direct and associate effects is:

$$
{}_{da}\sigma_A = 2\Sigma p_{i_1}({}_{d}a_{i_1})({}_{a}a_{i_1}).
$$

The consequences of ignoring the existence of interactions among competing genotypes in selecting individual trees can then be traced with the same simplifying assumptions that Griffing used to derive the noninteractive solution. However, by also tracing the associate effects of the trees selected, he derives a selective value of:

$$
W_{i_1j_1} = 1 + \left(\frac{s}{\sigma^2}\right)_{i_1j_1}{}^{d}
$$

and the gametic array of selected individuals is:

$$
(1/2)\,\Sigma\, p_{i_1}p_{j_1}w_{i_1j_1}(A_{i_1} + A_{j_1}).
$$

Random mating among these selected parents then generates a new mean of approximately:

$$\Sigma\, p_{i_1}p_{j_1}p_{i_2}p_{j_2}\left[1+\frac{s}{\sigma^2}\left[{}_d a_{i_1}+{}_d a_{j_1}+{}_d a_{i_2}+{}_d a_{j_2}\right]\right](_{i_1j_1}d_{i_2j_2})$$

$$=\frac{s}{\sigma^2}\left[{}_d\sigma_A{}^2+{}_{da}\sigma_A\right].$$

This reduces to the familiar sh^2 if the covariance of direct and associate effects is zero, which implies that there is independence of the two effects and that, on the average, we can select on individual performance with impunity. However, if a strong competitor is a vigorous tree which suppresses its neighbor, then a negative covariance can exist and gain can be substantially reduced. If the covariance is positive, a benefit is obtained over the case of ignoring competition effects.

If some form of group selection is used in which groups of size (gr) 2, 3 . . . n are chosen for mixed growth properties, the selective value of groups is taken as the contribution of both direct and associate effects. Then under group selection:

$$w_{i_1j_1,\ i_2j_2}=1+\frac{s}{\sigma^2}(gr)\,\tfrac{1}{2}\left({}_{i_2j_1}d_{i_1j_2},\ \ {}_{i_2j_2}d_{i_1j_1}\right)$$

and selection of groups within which random mating occurs yields a mean of approximately:

$$\tfrac{1}{2}\frac{s}{\sigma^2}(gr)\left[{}_d\sigma_A{}^2+2_{(da)}\sigma_A+{}_a\sigma_A{}^2\right].$$

While the latter factor can never be negative and never smaller than ${}_A\sigma^2$ in the individual case, the second factor can be small depending on testing ability.

Extending these results to groups of size n can benefit selection. However, if testing is limited, it might be easier to use direct individual selection with separate measurements such as crown diameter, root exudates, etc., to establish the form of competitive influence and to compose populations without direct group testing. This is essentially the recommendation of Toda (1956), who recommended selection for growth with narrow crowns in *Cryptomeria*.

If competition as well as other ecological variables represent factors affecting uncertain variations in future environments, breeding goals must be modified either to maximize some criteria of value in spite of future variations or to interactively develop breeds capable of responding to changing competitive environments. When future trends can be predicted, selection for special conditions to utilize any interaction effects can be developed and enhanced as suggested above. Similarly, for pest resistance development, the coevolution of predator and prey populations can be modeled as a special form of competition and dual evolution de-

veloped in maximally useful or minimally harmful directions. In such cases, simple monitoring or directed breeding on predator populations may be required, but coevolutionary systems between the populations should require only small extensions of existing theory.

CHAPTER 5
TREE BREEDING PROGRAMS

At best, the formulation of a tree breeding program is difficult. Even when gene effects, variances, correlations, etc., are estimable, sources of genetic variations are predictable, and breeding systems and testing procedures are also operable, the integration of all such functions into a breeding program is likely to be complex. The breeder still has the considerable task of integrating his functions into the silvicultural and forest management systems. Even assuming that a planting program is well established according to forest removal schedules and sociopolitical necessities, the breeder must project the desired profile of genotypes as it may change over many generations, since such profiles can be expected to change. Since environmental and economic changes are predicted with high error, the breeder must determine the relative merits of developing special breeds for special needs, or more generally adapted breeds. In general, one population intended for long-term development would not be maximally improved for traits of enduring value if selection for more ephemeral objectives is also imposed on that population. Selection for traits that degenerate in value within a breeding generation is clearly a costly waste, since selection effect on other traits is somewhat decreased. Thus, alternate means of controlling trait characters by silviculture, harvest or processing and conversion techniques have to be investigated within an overall forest management system. For those variations in the physical or economic environment which remain unpredictable, we require breeding strategies that at least maximize minimum gain or satisfy other less conservative criteria which may be developed.

If traits can be initially allocated among the various means of value control, such as genetic, cultural, and engineering, there will undoubtedly be further interactions among such factors as the form and value of say silvicultural versus genetic improvement. Refined management tools such as critical path analyses can be employed to find the best solutions. Obviously, a breeding program for improved response to intensive silviculture depends on the simultaneous application of intensive cultural techniques and on the existence of particularly responsive trees. In addition, breeding for survival and pest resistance alone may induce cultural investments not otherwise considered feasible. Hence, joint consideration would require breeding for cultural response as well.

Breeding to produce commercial seed that is more uniform in some traits can have an effect on silviculture that is not now foreseen. Reliable responses to treatments and reduction of tree-to-tree and stand-to-stand variation can reduce risk and increase the benefit derived from any management program. In general, reduction in such variation requires a special breeding or selection effort, since most breeding programs do not decrease genetic sources of variance and indeed are designed to avoid that. However, more uniform growth types might develop through drastic reduction of the lower end of the value scale by both silvicultural and genetic techniques. To some extent, the management control of forest values must depend on the extent to which the agencies' breeders and silviculturists control forest values. If the agency grows its own wood, then direct optimization of all control options can be utilized. However, if only a small portion of the total product is self-controlled, then reliance must be placed on uncontrolled or indirectly controlled sources of wood and other forest products. In that case, the forest manager may require special breeding opportunities to balance a general, uncontrolled, supply profile of wood types instead of direct manipulation of the entire forest product array. Public agencies or organizations devoted to a wide range of uses will clearly have broader objectives than those that produce only fiber or those that can more closely control the environments in which the trees will be planted.

An array of methods exists by which the breeder can genetically manage the character composition of individual trees and the mixture of types in single populations or in sets of separate stand types. He must coordinate his activities with changes in ecological-silvicultural management and economic management of various forest products. According to some general forest systems analysis, an optimum operation would require that some traits be controlled exclusively by product conversion methods, even if genetic control is potentially useful. Once an array of trait distributions is determined and site subdivisions are established, some sites may require breeding for heterosis in many traits while others may require breeding for additive gene actions on most. For each population, a mixed breeding operation may be required to deliver the trait combinations desired. Each site may also be planted to a mixed assortment of trees from different populations, but each breeding population can also be expected to have to mix trait types. A mixed breeding system using heterosis for some traits and additivity for others may then be found useful.

When breeding populations must be kept small because of operational limitations or by deliberate choice of maximizing selection differential at the expense of population size, developing small local populations split off from a larger regionally adapted population may be a useful way to organize a continuing breeding program. Thus, larger immediate gains for traits or sites of special interest can be attained by intensive breeding within small populations. As the smaller populations lose their ability to respond to

recurrent selection or to selection for new traits, or as inbreeding depression becomes a more serious problem, the breeder may reenter the larger population for at least a partial exchange of genotypes. The larger population would then have been bred for general adaptedness and for fewer traits which possess persistent value, such as total yield and pest resistance. Cooperative regional programs would thus permit individual agencies to maximize short-run gains without losing general ecological or economic adaptation. Furthermore, in the smaller populations, it may be possible to breed for uniformities which may be too risky for long-term development (Namkoong and others 1971).

Alternative exchange programs among agencies with small, different, genotypic compositions are also possible within ecological zones and represent a modification of the replicated breeding populations proposed by Baker and Curnow (1969). If the breeding objectives have been sufficiently similar, then the selection of particularly good combiners among replicates expands the parent population. However, the advantages of replicate selection are not obtained if the allelic frequencies failed to diverge among replicates (Madalena and Hill 1972). Therefore, cooperative planning to maintain individual and population replicate identities is required in any kind of segmented population development.

Many selection problems would be considerably reduced if breeding generation times were reduced and juvenile-mature tree correlations increased by either reduced harvest time or more precise estimation of harvest values from youthful seedlings. Not only would gains be more rapidly made, but opportunities for using a wider range of breeding methods, such as single-cross, backcross, and rapid breeding, would become available. In addition, the number of traits it is feasible to include in breeding control programs would increase as predictions of future sites and values become more certain. Thus, early flowering experiments, such as conducted by Stern (1963), and the possibilities of developing early harvest can provide breeders with many opportunities to investigate alternate breeding methods for their utility in forests.

BREEDING PROGRAMS

Breeding programs thus must encompass the whole range of activities in which breeders must engage. In addition to testing and selection, which involve economic and silvicultural projections, the breeder must see that the breed populations are optimally developed, that estimates of genetic means and variances are obtained with precision, and that commercial seed production is maintained at acceptable levels. In this book, the various operations have been separated in order to describe the separate goals and methods required for each phase, but most actual operations cannot afford to run separate programs for variance component estimation, controlled crossing to develop breed population, testing for selection, and seed production. Each tree breeder must

operate within his physical, biological, and financial constraints to accomplish some minimum objectives. If the primary value of breeding is genetically improved seed production and some inefficiencies can be tolerated, the breeder may be forced to forgo any mean, variance, or site response estimation. By choosing only traits which he believes have high heritability, he may also forgo testing and simply start a mass-selection program. He may even ignore the development of an ancestrally controlled breeding population and continually reselect a few parents for each subsequent generation without regard to ancestry. Such a minimal program is essentially the same ancient tradition of farmers who saved the best seed for the next crop. The modern tree breeder will undoubtedly be more aware of both the opportunities and limitations of various breeding alternatives. He may also be able to make small investigations on the distinctions between genetic and environmental sources of variation, and he will avoid some of the limiting effects of inbreeding in small populations. His methods, however, may not ultimately be much different than that which a genetically untrained but intelligent forester might develop. If by some cooperative programs, or the increase in his own capacity to develop more sophisticated programs, he can make controlled matings and experimental plantings, then the alternatives expand for generating large immediate and future gains and adapting his population to the changing needs of the forests.

The requirements of seed production will often be an independent consideration and may most often be handled in specially constructed orchards involving very few parents. The most genetically restricted type of production orchard would be established from cuttings or scions from a single selected genotype or self-fertilized seed. Such an orchard would obviously have little potential for generating better genotypes, but would, on the average, give one of the best gains possible from currently available genotypes. The next most limited production orchard would contain only two genotypes with only the single-cross seed being commercially produced. More complicated genotypic crossings for pure- or hybrid-population productions then follow. They may involve varying degrees of controlled pollination ranging from the production of specified full-sib families, through partial control by dusting with a selected pollen mix, to open pollination among all genotypes in an orchard. While it is not always possible, production orchards can often be separated from other phases of the program.

The generation of an optimum breeding population to create cumulatively better genotypes for production orchards is the principal objective of the breeder and is the principal problem we have considered in this book. The basic operation we have considered is the reduction of a base population to a few parents, then the regeneration of an improved large population from the selected parents and the sequentially repeated production of large populations by crossing among a smaller set of selected parents.

The number of parents may be only as large as the number of production orchard parents or may be as large as several hundred genotypes in a hierarchal or factorial breeding system. The crossing among these parental genotypes may be restricted and controlled among all possible parental combinations, or even uncontrolled, in which case the production and breeding orchards may be genetically identical. Since one purpose of breeding control is to maximize both the recombination of genes and the effective population size, the control of pollinations among as many crosses as possible is generally desirable. This control might be accomplished in stages rather than all at once, if time permits. The earliest crosses may also be done for other purposes, but supplementary crosses to increase the controlled ancestral breeding population can easily form one of the later options for additional breeding. As large a progeny population as possible would then be desired, since the reselection of parents for the next generation will attain maximum progress if the selection differential is large. Since certain minimal parental numbers are required, we must have large progeny populations to select from and to have representation among several full- and haf-sib families, or cousins, as well as within families. As previously discussed, some compromise will have to be reached between maximizing the selection differential and including a minimum relatedness and number of parents in the breeding population. The choice of crossing pattern is clearly affected by the number of parents, since the greater their number, the more costly it is to make all possible crosses among them. Thus, a solution for both the mating pattern and parental number may have to be sought simultaneously. In general, however, operating costs may not be affected much by these choices, and the solutions may often be found independently—the minimum parental population size can be determined first and all possible crosses can be made among whatever number is chosen, as in a large diallel, or possible in some modified form of the factorial or hierarchal designs. Unbalanced designs cannot be ruled out as a deliberate choice, and for many purposes may be most suitable. If the breeding system used is at all complicated, the choice of parental number (and hence, selection differential) and mating pattern can be very complicated and would require that the breeder trace all expected gains against cost projections.

The testing of genotypes (by using relatives) for inclusion in a breeding program requires further compromises, since testing requires both a set of crosses in a mating design and an array of test sites and time to accomplish the evaluation before the final breeding population is selected. The problem for the breeder is that the large population from which he will select his breed parents must be reduced to a new set of selected parents in several stages over a longer time period. The overall heritability of traits will help determine the utility of such additional testing and may sometimes be high enough that testing for additional data is

not worthwhile (Namkoong 1970a). However, if testing is cheap and can be done quickly for low-heritability traits, a good environmental and mating testing design may be instituted to select more accurately. A compromise is then required if the same crossings are to be used for breed production as for testing. Not only will many crosses have to be discarded because their parents prove unworthy, but a design which may be good for testing, such as the factorial North Carolina design II, may be a poor breeding design since it would include a few common parents which produces a breed population with high coancestry. The objectives of testing may therefore have to be compromised by using estimates of breeding value which have somewhat higher errors of estimation than what optimally designed experiments may produce. The various partial and blocked diallels yield moderate testing errors as well as reasonable flexibility in selection for breed population development. If blocking, replicated subblocking, or both are desired for test efficiency, then the various disconnected designs can be used. Within disconnected blocks, partial diallels can be constructed, and, if desired, partial overlapping of blocks can also be included without any additional analytical problems.

The separation of breeding and testing operations is clearly to be desired, but its cost is higher and the increased testing efficiency would have to be balanced against the direct costs of additional experimentation as well as costs incurred indirectly in reduced selection differentials caused by misdirected efforts.

On the other hand, estimation experiments may be reasonably compatible with good breeding design. We may often wish to draw a subset of the breeding crosses for estimation of variances to determine levels and changes in genetic variances and covariances, particularly if new traits or changes in selection goals are incorporated. While the demand for balanced designs seems obvious, the conflicts may not be serious. Thus, diallel mating designs may find some favor for breeding programs which require simultaneous estimation and breed population production. Again, however, unbalanced designs can be very efficient for estimating variances and may be useful in breeding populations if inbreeding can be controlled.

The conflicts which arise from the contrasting requirements of testing and variance estimation may also be difficult to resolve and will more often have to be compromised by the overriding requirements of breed population production. However, if these two objectives can be separated from the other functions, the factorial mating design can be made reasonably efficient for both testing and estimation.

Burdon and Shelbourne (1971) offer a comprehensive review of some of these alternatives for testing, estimation, and breeding and their conflicting problems as they affect a breeding program in New Zealand. While each breeding agency would have different constraints and capacities, each breeder should consider the vast alternatives available for choice to develop his own uniquely suit-

able program.

Timing is a very important aspect of the genetics program itself and of its integration into the general forest management system. A flow chart or some other device is often useful to assure that matings for the breed population are properly timed for maximum gain rates, or that test data are available when needed for seed or breed production.

In some situations, no genetic research or development program is justified or efforts are best limited to provenance selection (Wright 1971). Product value may be low, genetic-gain potential may be low, or the efforts may be better spent in another way. Provenance selection offers no chance for cumulative improvement by reselecting among provenances, but if seed procurement and planting are used, mass selection and simple recurrent selection can provide advances at virtually zero marginal costs, opening the potential for further gains at higher marginal costs. In cases where only a small improvement effort is feasible, it would be highly advantageous for a government agency or regional cooperative to estimate genetic potential and preserve gene resources. For future programs, the assortment of species into programs and initial breeding steps would be greatly advanced if such data and material were available.

Since such actions have not generally been taken, forest geneticists have usually been forced to make preliminary judgments on scant data. In most cases, an action program should be designed to provide for material and data for the next breeding generation as well as improved stock for current plantings. While one may usually be able to reduce an operation to simpler mass-selection schemes, it is often more difficult to expand the complexity of a program unless some degree of controlled crossing is exercised. A precaution to take in reducing a program to mass selection is that the population size be kept much larger to assure a reasonable effective population size.

Controlled crossing and site sampling make possible various options for selecting among specific crosses and among specific genotypes for certain sites. Since gain rates will be heavily influenced by time, early selection for seed production and for breed development will be advantageous. Hence, the crossing and planting systems should be forced as early as possible into the operations. If properly coordinated, tree breeders can have the data and materials for most of their species in one generation.

CHAPTER 6
MODELS OF POPULATION GROWTH

Populations of trees, shrubs, small animals, and even human beings are composed of individuals that differ from each other greatly or slightly, depending on the characteristic studied and the space and time scales of the observations. Such variations are not only of interest in accurate descriptions of populations, but are the sources of the capacity to change. Thus, while averages are valuable descriptions of populations and are usually our first perceptions of their nature, the variations that exist are more important to studies of population dynamics. Parameterizing means and variations in both the spatial and temporal senses is the subject of this chapter. The simplest population growth models are developed first. Population genetics has been founded on these models, which are very simple and include no age- or density-dependent effects. More complicated models will increasingly be needed to describe genetic concepts that are being developed. Several of these concepts and models are briefly described in the marked (*) sections.

In the time scale of human activities, forest tree populations appear to have stable size and composition. In the time scale of the trees, however, the populations behave as most sexually reproducing organisms. They are constantly changing. As seedlings replace old patriarchs, as dominant trees suppress their neighbors, or as whole forests die or regenerate after an environmental disaster, the age and genetic composition change. Such changes are continually occurring by chance, or in direct response to changing environmental pressures. In the scale of human economic activities, only large-scale catastrophes in forests impinge on the general public awareness. The exceptionally rapid disappearance of the American chestnut was fast enough to be widely understood, whereas even the great retreats of longleaf pine and eastern white pine in one generation were less easily perceived. Thus, major changes in distribution of important species have been but barely perceived. Large-scale genetic changes which will forever affect the future evolution of our forest resources have also passed with little notice. For example, the reoccupation of former longleaf pine sites which is currently occurring is generated in part from introgressants with loblolly pine (Namkoong 1966c). The historic record indicates changes of far greater magnitude in the loss of whole forests in Western China, in the reduction of

the Cedars of Lebanon to relic stands, and in the advance of Scots pine and Norway spruce through Scandinavia. It is thus clear that grand species movements and fluctuations in population size and composition do occur over space and time, in spite of our perceptual limitations, and have molded the evolution of tree species accordingly. Variations also occur on a smaller scale within species. Through differential reproduction within stands, and migration, isolation, and selection among stands, it has become clear that a wealth of genetic variation now exists within most of the species studied by foresters (see, for example, "Second World Consultation on Forest Tree Breeding," Food and Agriculture Organization of the United Nations 1970). It is evident that most studies of genetic variation were generated by commercial interests in tree breeding and that many agencies will be controlling the evolution of some segment of the forests through their breeding activities. The growth of forests in the future will therefore be controlled to some extent by human activity. For this reason, the forest biologist will have to understand the dynamics of population change to guide future forest compositions.

The preoccupation of population analysts is to describe and predict the changing patterns in relative abundance of genotypes, forms, and taxa over space and time. This chapter will analyze relative numbers in population subdivisions as a measure of population growth and development. By thus focusing exclusive attention on such a numerical measure, a presumption is made that numbers are identified with success. Other criteria of success or goodness can be advanced, such as durability on a site or probability of existence at some future age, and for trees, longevity may then be a concomitant measure of success. However, numbers are relatively easy to analyze and provide a measure of success if correlated with probability of avoiding death or extinction of lines of descent. Thus, when using numbers, we often make the implicit assumption that random or nondiscriminatory causes of death will remove individuals in proportion to their occurrence in the population, that those causes of death are important, and that we can therefore measure probability of survival by relative frequency.

Another restriction in the scope of interest will be the exclusion of the influences of interspecific evolution and interspecific competition on relative survivals of genotypes or age classes within species. By thus ignoring the substantial effects that interspecies relationships can have on variation patterns within species, we severely limit the exact applicability of our analyses to real forests. Recent studies by Dawson (1969, 1972), Levin (1971), Pimentel and Soans (1970), Kojima and Huang (1972), and Greenwood (1969) have demonstrated that interspecific effects can be important in the evolution of a species. They have shown that competing or predator species such as an insect or disease can alter the gene frequencies in populations from what they would be in the absence of the alternate species. However, enough

of the real world may be well approximated by our models without competitive effects of this sort, so that we can deduce the consequences of reasonably complete models and analyze the mechanics of a large segment of population growth and evolution.

Within these limitations, the problem is to mathematically express the biological concepts of population growth with parameters for response to environmental pressures, and to determine the stability of relationships among age and genotypic components. By analyzing the effect of observable factors on birth, reproduction, and death of individual trees, we expect to derive the probable behavior of the whole population and even of population differences. The origins of group differences are considered to arise only from differences among individuals which evolve into separate populations.

Such problems can be most directly analyzed if all trees behave exactly as expected and if all environmental factors are exactly predictable. Even if the interactions were quite complex, we could deduce exact relationships for any given set of conditions and could describe the populational variations solely in terms of variation in the external factors. If the external environment were very simple, then we could determine the exact behavior of the population. On the other hand, if behavior is not precisely determined but some elements of chance variation in response exist, then the average effects do not completely describe all deviations in any single population. The element of chance or the probabilistic nature of the response implies an unknown or unknowable set of causes such that individual events are not predictable, and that only average, collective tendencies can be described. Thus, in a deterministic model, it may be ordained that during the evolution of an oak forest, the seedlings will have 20 percent mortality the first year and 20 percent of the remainder in the second and each succeeding year, and that the survivors will reproduce one viable seedling in their tenth year, two in their fifteenth year, etc. Then, we can derive the exact age distribution of any stand, at any time, by simply following the predetermined course of any given population. On the other hand, it is often more reasonable to say that the birth and mortality schedules vary somewhat from tree to tree and from stand to stand. We may say that the process is not exactly determinate but that 20 percent is an average mortality that is rarely exactly achieved. Then, the model contains errors in estimating occurrences and chance or stochastic variations in the cause-effect process. While the average survival may be the same in the deterministic as in the stochastic processes, we have now generated a process in which variation can exist among stands or in which variations from an average predicted course can exist during the history of a single stand's development. In this case, it will be important to know the extent of variation around predicted events. The probability of stand extinction may exist and is of critical importance to estimate. Hence, analyses of such stochastic processes are required to

confirm or modify expectations based on deterministic processes. We will follow a common sequence of analyses and shall first examine deterministic models and their analyses, then examine some stochastic variations and their analyses of the same effects. Thus, any changes in the age or genotypic distributions which may have occurred over the past 100 years may be better understood. In addition, more optimal changes in the next 100 years may then be planned in terms of averages and expected variations. We shall therefore be concerned not only with the average behavior of populations but also with such measures of variation as the variance and correlations of numbers in age classes and genotypes as they may change over space and time. Also, some populations and some genes will wax or wane in relative numbers and some will go to extinction, requiring us to also consider probabilities of those events.

THE SIMPLEST MODEL

As a first approximation, life parameters may be simplified into general propensities for an individual to survive, reproduce, and die at any time within some generational timespan. Age-dependent processes are ignored in this oversimplified concept, and all life events are lumped into these simple categories without reference to time of action or to any interrelatedness of action. For example, if we were to observe an isolated forest stand at intervals of 20 or 30 years, many trees would be present in sequential observations, but some would have died. In or near their place, others would grow and might die. Thus, it is possible to imagine tree populations as starting from small colonies and increasing in numbers over several generations by occupying border areas as well as by increasing stand density, as Bannister (1965) has observed for populations of *Pinus radiata*. If it is further assumed that at some upper limit to population expansion, the members of the population are removed in proportion to their relative frequencies, then population limitations would be maintained and predictions of relative frequencies would still be accurate.

Populations growing in such a manner increase at each time period at a logarithmic rate, as determined by the relative propensities of each tree to regenerate or die. At a rate of increase of m, an initial population of n individuals would increase to $m \cdot n$ in one time period. Using superscripts to denote time periods, with $n^{[0]}$ original trees, $n^{[1]}$ would be equal to $mn^{[0]}$, and in the second period, $n^{[2]}$ is increased to $mn^{[1]} = m(mn^{[0]}) = m^2n^{[0]}$. In general, for t time periods, $n^{[t]} = m^t n^{[0]}$, or in terms of logarithms:

$$ln\,(n^{[t]}) = ln\,n^{[0]} + t\,ln\,m, \qquad (1)$$

which is a linear function of $ln(m)$ with time being the only variable.

We can similarly model the growth of populations which reproduce, as above but on a more continuous time scale, and define

a propensity to increase (w) in terms of rates to increase for vanishingly small time intervals. Thus, the number at time t is $n^{[t]}$ and increases after a small time period (δt) to $n^{[t+\delta t]}$, and the amount of increase is written as $n^{[t+\delta t]}-n^{[t]}$. Then defining our propensity to increase (w) as the ratio of the increase in numbers to the numbers at the start of the interval, multiplied by the length of the time interval, we have defined:

$$w=\frac{n^{[t+\delta t]}-n^{[t]}}{n^{[t]}\cdot\delta t}.$$

Now allowing δt to become very small, we define w as the limiting value. As $\delta t\rightarrow 0$ in the above expression, we derive the differential form of that equation as:

$$w=\frac{dn\ (t)}{n(t)\,dt}=\frac{[d\ ln(n(t))]}{dt}, \tag{2}$$

where n is now a continuous function of time, $n\ (t)$, and ln is the natural logarithm. This is analogous to our definition of m in defining a logarithmic rate of increase in numbers with a rate parameter w in place of m. The m is often called Malthusian parameter. We can integrate (2) to get the same form as for the discrete process:

$$ln(n(t))=ln(n(0))+w\cdot\Delta t \text{ or } n\ (t)=n_oe^{wt} \tag{3}$$

In either case, the population is expected to grow at an exponential rate until density-dependent or other processes force a change in the model. Thus, if two tree species or populations were to invade a new site, both would increase according to their propensities to reproduce (for example, their fitnesses while the population was expanding), and the one with the higher replacement rate (w) would dominate by occupying increasingly more territory or by being more heavily represented among the replacements whenever chance mortality reduced total population size.

While this model of population growth has served as the main basis on which genetic models are built, many more complicated and more realistic models of population growth have been developed. They have not yet been extensively used in genetics. This shortfall of application demands the attention of foresters and other geneticists who are familiar with the ecology of their species. Interspecific competition, for example, can have a major effect on relative fitness of genotype and therefore should be included as a frequency-dependent effect in any generally useful population model (Dawson 1969; Greenwood 1969; Levin 1971). In addition, it is obvious that environments change and alter fitness values, that age-dependent processes vary significantly, and that competition and density dependence often induce significant frequency-dependent effects on survival and germination processes. Hence, the extension of models to include the multiple and variable effects of genes is clearly desirable. Such models should be developed to better reflect the causes and effects of variations. Recent

models of genetic fitness that show competition among and within genotypes in a logistic equation form (Clarke 1972) open many new channels of investigation into density-dependent gene actions of simple types. Before discussing the genetic models derived from the simple population models, some of the complications of population models deserve discussion.

DENSITY-DEPENDENT AND COMPETITION MODELS*

It is clear to most foresters that species differ in density-dependent processes and hence differ in their relative abilities to exist and reproduce under dense conditions. While many aspects of tree growth and of the environment are involved, the different tendencies to increase under crowded conditions may be generally modeled by decreasing the reproductive rate or increasing mortality as some upper limit in population size is approached. One commonly used model parameterizes this effect by multiplying the rate by a factor which decreases as population size approaches an upper limit a. Thus, the change in $n(t)$ with respect to time of equation 2 can be written as: $\frac{dn}{dt} = w \cdot n$ and by multiplying by a factor of $(1-\frac{n}{a})$ the differential equation of this density-dependent model is:

$$\frac{dn}{dt} = wn\ (1-\frac{n}{a}).$$

The two parameters w and a then determine relative population growth rates and, if survival of genotypes or populations is a simple function of relative frequencies, relative success is also dependent on the density of the population. Each genotype, for example, may endow its possessors with tendencies to larger or smaller w or a factors. At low density, the type with the larger w will be increasing faster, but it may suffer relative to any alternative type with a higher a at some higher density. For populations which do commonly increase in density to the point where death or birth is differentially affected, the a values will be important.

In a very rough and oversimplified sense, tree species may be divided into those with heaviest selection for their w factors versus those with heaviest selection for a factors. The old-field and pioneer species may be most heavily selected for their ability to reproduce quickly and invade new territories. Often, such species of the *Salix*, *Populus*, or *Pinus* genera deteriorate the environment for their own reproduction and require some disastrous type of site clearance in order to reproduce. Among these species, the more successful types will be those with a high w in the face of interspecific and intraspecific competition. In contrast,

*Graduate-level statistical training required for thorough understanding.

climax species which are elements of a stable community require an ability to grow under intense intraspecific competition, and the more successful of such types will be those with high a parameters. These models refer only to the effects of numbers regardless of the genotype of the competitors. When the type of competitor affects success, intergenotypic or interspecific effects require specification of the density of each type of competitor. The classical model of competition between species in which both self-regulation and interspecific regulation occur is the Gause-Volterra model (Slobodkin 1961). In this model, the density-limiting effect that a second genotype has on the first genotype is made proportional to the n of genotype 2, n_2, and would reduce the replacement rate by say αn_2. Similarly, the depressive effect of genotype 1 on the replacement rates of genotype 2 would be say βn_1. Then, instead of a rate depression of $(1-\frac{n_1}{a_1})$, the rate would be modified by $(1-\frac{n_1}{a_1}-\alpha n_2)$ for n_1, and by $(1-\frac{n_2}{a_2}-\beta n_1)$ for n_2, where a_1 and a_2 are the limiting density effects of their own types. The two growth rate equations would then be:

$$\frac{dn_1}{dt}=w_1 n_1\left(1-\frac{n_1}{a_1}-\alpha n_2\right)$$

$$\frac{dn_2}{dt}=w_2 n_2\left(1-\frac{n_2}{a_2}-\beta n_1\right).$$

Similar equations for as many community participants as desired can be constructed and a set of first-order, ordinary differential equations constructed. Community stabilities can then be directly analyzed and several have been (Vandermeer 1972), but not thus far in forestry.

While this logistic model may now account for a form of self-regulation, it is still a crude approximation to reality and is not unique in producing the kind of inhibition sought. Any number of polynomial functions or other nonlinear forms could have been hypothesized, including the addition of elements for insects or diseases that cause mortality according to the frequency with which the host type exists. Nevertheless, studies on the existence or nonexistence of stable equilibria are often couched in the terminology of the Gause-Volterra equations (Ayala 1972; Gilpin 1972).

AGE-DEPENDENT MODELS*

A different degree of complexity can be introduced into our model by considering the changes that occur in the life of individual trees in their capacities to survive and reproduce. It has been pointed out by several authors (for example, Lewontin 1965;

*Graduate-level statistical training required for thorough understanding.

Demetrius 1969; Anderson 1971; Anderson and King 1970) that life history variations can significantly alter the simple parameters of population growth. It is also apparent that means and variances in traits which affect life processes drastically change as trees mature (Namkoong and others 1972). Therefore, life-cycle effects must be included in any study of relative population growth rates and individual survival probabilities. In addition, concepts of interspecific competition in the logistic form of density dependence can be affected by age-density changes in competitive effects (Gilpin and Justice 1972).

If we again revert to the simplest model of logarithmic population growth without density dependence, and assume that propensities for survival and reproduction exist without any form of competition or self-regulation, and only add that the processes are age dependent, we can derive some interesting, limiting forms of population growth. To simplify our examples, let us consider that a forest tree grows in only three stages (seedling, A; sapling, B; and mature, C) and that survival and reproductive probabilities can be predicted for each life class. The methods and theories developed for three classes are directly extendable to any number of age classes that would be desired for a more complete model of truly different growth phases of a forest. Ages, for example, could be grouped into classes of 5, 10, or 50 years or any time interval. Pursuing our model with only three age classes A, B, and C, assume that a good measure of class A seedling survival into class B is 30 percent, for survival of trees from periods B to C is 40 percent, and that all trees beyond that age died but were able to produce living seedlings of class A before death. Suppose further that trees in the seedlings of class A could produce no seed, but that saplings in class B could produce enough seed to yield an average of four surviving seedlings (of class A) and that trees in the last period, C, would produce an average of two surviving seedlings (of class A) before dying. If the population is observed at some starting time with age class ratios of 3 seedlings (A) : 2 saplings (B): 1 mature (C) tree, we could then trace the expected growth of the population. For example, out of a total population of 600 trees, the 300 A trees allow only 30 percent to survive and to advance into the next age class, leaving only 90 B trees and producing no new A trees. Of the 200 B trees, 40 percent would survive leaving only 80 C trees, but 800 new A trees among them. The 100 C trees would not live beyond this period but would leave 200 A trees. The next generation would then have an age distribution with ratios of 10 of class A: 0.9 of class B: 0.8 of class C. Repeating the process with the new age class numbers, the subsequent development of the population can be traced from its original:

$$\begin{pmatrix} n_A \\ n_B \\ n_C \end{pmatrix} = \begin{pmatrix} 300 \\ 200 \\ 100 \end{pmatrix} \text{ to } \begin{pmatrix} 1{,}000 \\ 90 \\ 80 \end{pmatrix} \text{ to } \begin{pmatrix} 520 \\ 300 \\ 36 \end{pmatrix} \text{ to } \begin{pmatrix} 1{,}272 \\ 156 \\ 120 \end{pmatrix}, \text{ etc.}$$

If the process is continued, a few interesting patterns emerge. The total population size would tend to increase over the course of several generations but initially it fluctuates from 600 to 1,170, to 856, to 1,548, etc. A certain constancy emerges if we take the numbers in each generation as a ratio of the numbers in the previous generation. Taking the ratios of 1,170:600, then 856:1,170, this sequence is: 1.95, 0.732, 1.808, . . . , which would fluctuate around a final value of 1.1196 with smaller and smaller variations. The increase in size of each age class also tends to fluctuate about the same limiting value of 1.1196, and, if followed long enough, the ratios of numbers in the age classes would be seen to settle down to constant ratios of 1:0.2679:0.0239. If we had started with other age-class distribution than the (3:2:1) which we arbitrarily chose, the exact same result would have ensued and the same asymptotic ratios would have been reached. We could have predicted these results if we had considered that the process is simply one of tracing one-generational-step transitions from each age class either to death, or to the next age class, and also to their contributions to reproduction. Computation can be done by multiplying the same life probabilities sequentially, but it is more simply done by matrix multiplication.

If the $\begin{pmatrix} n_A \\ n_B \\ n_C \end{pmatrix}^{[0]}$ of the initial population is the source of the seedlings in the next measurement period ($n_A^{[1]}$), then the equation relating $n_A^{[1]}$ to the initial population is:

$$n_A^{[1]} = 0 \cdot n_A^{[0]} + 4n_B^{[0]} + 2n_C^{[0]}$$

$$= (0\ 4\ 2)\begin{pmatrix} n_A \\ n_B \\ n_C \end{pmatrix}^{[0]} = (0\ 4\ 2)\ \underline{n}^{[0]},$$

where $\underline{n}^{[0]}$ is the column vector of n for period 0. Similarly, it can be seen that:

$$n_B^{[1]} = .3n_A^{[0]} + 0 \cdot n_B^{[0]} + 0 \cdot n_C^{[0]}$$

$$= (.3\ 0\ 0)\ \underline{n}^{[0]}$$

and, $n_C^{[1]} = (0\ .4\ 0)\ \underline{n}^{[0]}$

Therefore, $\begin{pmatrix} n_A \\ n_B \\ n_C \end{pmatrix}^{[1]} = \begin{pmatrix} 0\ 4\ 2 \\ .3\ 0\ 0 \\ 0.4\ 0 \end{pmatrix} \begin{pmatrix} n_A \\ n_B \\ n_C \end{pmatrix}^{[0]}$

or $\underline{n}^{[1]} = M\ \underline{n}^{[0]}$, where M is the matrix of life coefficients and $\underline{n}^{[t]}$ is the expected number vector in each class of time t. The matrix M contains reproductive rates for each class on the first row, and the survival and advancement probabilities for growth into B and C classes in the second and third rows, respectively. The coefficients are defined for the time period considered and clearly determine the growth rate of the population. Thus, multi-

plying by a matrix M is equivalent to multiplying by successive powers of m in the simpler case. We can also project that if a steady state in the ratios among the age classes exists, then the vector of numbers must eventually have a common ratio. This fact can be expressed as:

$$n^{[t]} = Mn^{[t-1]} = M \cdot M \cdot n^{[t-2]} = M^2 n^{[t-2]} = M^t \cdot n^{[0]}.$$

Eventually, if $n^{[t]} = \lambda n^{[t-1]}$, then substituting in the above equation, $\lambda n^{[t-1]} - M\, n^{t-1} = 0$ and $(M - \lambda I)\, n^{[t]} = 0$, and, therefore, $M - \lambda I = 0$.

To satisfy this equation for any given M matrix, a certain set of λ values must be found which are called the eigenvalues and which are comparable to the Malthusian parameter used in simpler population models. Associated with each eigenvalue, there would be a vector of age class numbers n, called the eigenvector, which is useful in determining relative growth of the age classes. On positive M matrices such as would exist for life tables, we can invoke Frobenius' theorem on their eigenvalues (Gantmacher 1964) and can assert that there will always be one eigenvalue which is positive and noncomplex, with modulus greater than all other roots and with an associated eigenvector with all positive elements. Therefore, there always will be a solution to the system of equations $(M - \lambda I)\, n = 0$ and a vector of age class ratios to which the population will grow asymptotically. The largest eigenvalue λ is the asymptotic growth rate of the population, as well as of the age classes at their equilibrium ratios, and takes the place of m in projecting future population growth.

The only added complication which we have introduced to the exponential growth model is the fluctuating patterns that the actual numbers are expected to take due to the time that it takes for excessive numbers in any one age class to influence all age classes. Such fluctuations can strongly influence the population to endure large changes in total numbers, and if there are any changes in genetic or other frequencies associated with the ages of the trees, those measures also will vary until the age class distribution achieves some stability. Anderson and King (1970) have shown that such variations in age classes affect population gene frequencies whenever frequencies change with age class and that any changes in gene frequencies can, in turn, affect perturbations in age distributions. Thus, whenever the population is in age-class or gene-frequency disequilibrium, the time required to achieve stability can be long and a significant factor in maintaining some fluctuations without other cause. If we then look at the other roots to the matrix equations, we can see that this type of fluctuation can also be predetermined.

Using a matrix of real positive numbers, we can usually expect that the equations for the sequential replacement of age classes in trees from each preceding age class will be independent in the sense that the matrix of transition probabilities will be of

full rank r. We can also generally expect that all r of the roots of the matrix M will be distinct. Under these conditions, the transition matrix can be changed to give the same results in a form in which the effects of fluctuations can easily be traced in the expected progress of the population. Where the eigenvalue λ_i of the matrix M has an associated eigenvector $\underline{n}_{[i]}$, we can write: $(M-\lambda_i I)\underline{n}_{[i]}=0$ For the whole set of eigenvectors then, we can write:

$$\left[M-\begin{bmatrix}\lambda_1 & & 0\\ & \lambda_2 & \\ & \ddots & \\ & & \ddots \\ 0 & & \lambda_r\end{bmatrix}\quad [I]\right]\ \underline{n}_{[1]}\underline{n}_{[2]}\cdots\underline{n}_{[r]}=[0],$$

which is equivalent to, $[M-\Lambda I]\ N=[0]$,

where
$$\Lambda=\begin{pmatrix}\lambda_1 & & 0\\ & \lambda_2 & \\ & \ddots & \\ & & \ddots \\ 0 & & \lambda_r\end{pmatrix},$$

and
$$N=(\underline{n}_{[1]}\underline{n}_{[2]}\cdots\underline{n}_{[r]}).$$

The complete set of eigenvectors spans a vector space defined by M, and any vector of initial age distributions can be written as a linear function of the eigenvectors. Thus:

$$\begin{aligned} M\,N &= N\Lambda \\ M &= N\Lambda N^{-1} \\ M^2 &= N\Lambda^2 N^{-1} \\ &\vdots \\ M^t &= N\Lambda^t N^{-1}. \end{aligned}$$

In particular, for any given initial age distribution, $n^{[0]}$, the progress in numbers for each age class can be written either as $n^{[t]}=M^t n^{[0]}$ or as $n^{[t]}=N\Lambda^t N^{-1}n^{[0]}$ which is much easier to determine once Λ and N are found. From the latter equation, it is now clear that each eigenvector contributes its stable age-class distribution in proportion to the size of its eigenvalue and, hence, that for any number of time intervals, we can trace the growth of the population by summing the contribution of each eigenvector as weighted by its associated eigenvalue raised to the power of the number of time intervals. It is also clear that as the number of time intervals gets large, the largest eigenvector dominates the rest since we continue to increase the power of its eigenvalue contribution. During the initial generations, however, the effect of

having a disequilibrium age distribution on the attained age distributions is determined by the size of the other, often complex roots.

The development of comparable models for the time-continuous case is fairly direct and might often be an easier method of deriving statistics for meaningful cases, especially for those cases in which the number of age classes is very large or in which birth and death are continuous in any reasonable time scale. The basic model simply reduces the time interval to zero. Hence, the birth and death processes become continuous functions of time.

The model in which birth and death are continuing events has some illuminating features. If the population growth rate tends to eventually even out at a rate λ, then each age class also necessarily increases at the rate λ as the stable age distribution is also approached. Hence, the same population growth model exists as for the logarithmically growing population first considered, except that the age substructure of populations can now affect the growth of populations.

Two types of functions have been used in traditional populational and demographic studies. The first is essentially the same as the matrix model and is a useful introduction to the second, which is a time-continuous model. The first form uses the convenience of two time indices or variables, t and x, to distinguish different points in time. The numbers at time t, $N_{(t)}$, and are related to the numbers which existed x time periods ago, $N(t\text{-}x)$, by the population growth rate during that time interval, e^{rx}, the exponential growth for the simple matrix models. Thus, $N(t) = N(t\text{-}x)e^{rx}$, or given the $N(t)$, the numbers which had to exist x time periods ago had to be $N(t\text{-}x) = N(t)\,e^{-rx}$. In particular, the newly germinated class $B(t\text{-}x) = B(t)\,e^{-rx}$. On the other hand, the numbers in an age class at time t are also a function of the numbers in younger age classes and their survival rate. In particular, from germination to age class x, if the survival rate is $l\ (x)$, then the numbers in age class x at time t, $N(x,t)$, were germinated x time periods ago and survived for that long and, therefore, $N(x,t) = B(t\text{-}x) \cdot l\ (x)$. If we defined $m(x)$ as a fecundity measure, the total number of births at time t, $B(t)$, would be the fecundity of each class multiplied by the numbers in each age class or $N(x,t)\ m(x,t)$ for all age classes. Since $m(x)$ is the same for all time, then $B(t) = \sum_x N(x,t)m(x)$ and from above $B(t) = \sum_x B(t\text{-}x)\ l\ (x) \cdot m(x)$. Then if the B age class also follows the simple rules of proportionate growth, $B(t) = B(t\text{-}x)e^{rx}$, or conversely, $B(t\text{-}x) = B(t)e^{-rx}$; then $B(t) = \sum_x B(t)e^{-rx}\ l\ (x) \cdot m(x)$. Division by $B(t)$ yields $1 = \sum_x e^{-rx}\ l\ (x) \cdot m(x)$. Then, population growth-rate parameter r is a function of $l\ (x)$ and $m(x)$ schedules, and mathematical analysis indicates that for a stable population with $B(t) = B(t\text{-}x)e^{rx}$, there is an r that gives $\sum_x e^{-rx}\ l\ (x) \cdot$

$m(x)=1$ for any $l(x)m(x)$ schedule. In fact, there are as many r roots as age classes, but the largest positive root again determines the asymptotic behavior. The λ's of the matrix solutions are the e^r roots of these solutions.

In the continuous cases, a population with a given $l(x)$ and $m(x)$ schedule would produce progeny within a Δx time period at a rate of $l(x)m(x)\Delta x$. If Δx is reduced to an infinitely small dx, and we summed the progenies from parents of all ages, then total births over a single lifetime would be expected to be $\int l(x)m(x)dx$. In a population of individual trees of different ages, and at an instant in time (t), we might guess that the $l(x)m(x)$ schedules would be operating on an age profile in the population described by some function of x, say, $f(x)$. Then, new germinations for the whole population with the mixed ages in frequencies defined by $f(x)$ would be estimated by $\int f(x)\ l(x)m(x)dx$. Since $f(x)$ is the numbers of trees now alive in the x age classes, it must have been produced by the $l(x)m(x)$ schedules of past times. If $l(x)m(x)$ schedules are constant, then our $f(x)$ might be expected to be proportional to the $f(x)$ of some time (t) ago. Individuals which are age x now were germinated x time intervals ago and can now be parents of new progeny. Therefore, at this time (t) the number of births $b(x)$[1] from x aged individuals $=b(t\text{-}x)\ l(x)m(x)$ and hence the total births at time (t), say, $b(t)$ for all ages, are:

$$\int b(t\text{-}x)\cdot l(x)\cdot m(x)dx.$$

We also know that growth of any age class is expressed in our model as a multiple of the instantaneous population growth rate e^r for each age class, and therefore that the number of germinants now (at time t) is equal to the number of germinants in the past (at time $t\text{-}x$) multiplied by e^{rx}. Therefore, $b(t)=b(t\text{-}x)e^{rx}$ and, therefore, $b(t\text{-}x)=b(t)e^{-rx}$. Substituting in the above integral gives $\int b(t)e^{-rx}l(x)m(x)dx=b(t)$ and hence dividing by $b(t)$ gives, $1=\int e^{-rx}l(x)m(x)dx$. We can thus write an integral equation for any $l(x)m(x)$ function and solve for the r roots, the instantaneous intrinsic rate of natural increase.

Since population growth is an exponential function of time, it can be seen that the major effect of variations in the $l(x)$, $m(x)$ life table parameters on r is exercised during the early life stages. Variations in the $l(x)m(x)$ schedule in later life stages are relatively ineffectual. This observation was offered by Fisher (1958) and Lewontin (1965) and was also shown to be true by Demetrius (1969), who demonstrated its validity for the Leslie matrix model of discrete age classes. It can be clearly inferred from Keyfitz's (1968) results for the integral equation form. Both the $l(x)$ and $m(x)$ schedules of younger ages are not only most effective in modifying population growth but are, consequently, also most intensely subjected to selection pressures, those individuals with maximum r, dominating those with small r. For example, growth or other behavior differences which affect successful reproduction

of loblolly pine would be much more important at age 15 than similar differences at age 50.

EFFECTS OF VARIATIONS*

For any given life table, variation must be expected in any real population development. Hence, the expected growth rates will not be realized with complete precision. For example, if a constant probability of mortality exists in any population, some variations from the exact expectation of population size would be common. Thus, in our previous example, if the probability of survival from class A to class B was 0.40, the sampling variance in the population could be reasonably expected to be $\left[\frac{p(1-p)}{N}\right]=\frac{0.24}{N}$. Even though survival and reproduction probabilities have such variances, it is still possible to predict expected population growth and an expected profile of population age classes. Pollard (1966) introduced a method for tracing the progress in expected means, variances, and covariances among the numbers in the various age classes. When sampling error occurs and each tree survives and reproduces independently with its stated propensities, then the events for the whole population are sums of binomially distributed events for all of the age classes. It turns out that the means are the same as predicted for the deterministic model. However, if many population trials were conducted, the expected variance could be large and any single population could develop growth trends quite different from the average.

In addition to random individual-tree variations around average survivals, there are variations in the average survival and reproduction expectations themselves. These variations can be observed in the changing conditions of life according to whether a stand develops on new ground or must develop through overstory competition. Environmental conditions vary on different areas at the same time and on the same area at different times, affecting survival and the probabilities of producing viable seedlings. Hence, variations about some average expected-growth rate and population age distribution are created by variations in sites over time and space. The means and variances for age classes in the future can be predicted with a deterministic model by varying the Leslie matrix (Sykes 1969). In general, variations due both to sampling deviations and variations in the actual probabilities of life events (Weissner 1971) may cause variance in age class and population numbers to increase faster than the squared population size. The variances may therefore be so strong that they continually induce age distribution disequilibria (Namkoong 1972). Thus, even

*Graduate-level statistical training required for thorough understanding.

though the average age distribution might be predicted for some future time if we know or can estimate the parameters and their variations (Billingsley 1961), large fluctuations can persist, causing extreme events such as stand or line extinctions to be of central concern though otherwise not expected. In addition, if stabilities may only be approximated, then any real demonstration of population stability must rest on the presence of other factors such as density-dependent effects and cannot be predicted from self-induced processes such as we have investigated.

POPULATIONS WITH GENE DIFFERENCES

Another source of variation within the classes considered above significantly affects population growth. Clearly, the species we deal with are commonly much more complicated than those we have been considering. Their members vary in survival and competitive and reproductive capacities. Any genetically variable trait which affects survival or reproduction possesses genetic variation with respect to the replacement rate in the population. Thus, genotypic or allelic variations affecting that trait can cause measurable variations in fitness values. For simpler models, the population can be expected to change its average fitness towards the more reproductively effective type. It is clear that intraspecies genetic changes in fitness require some variations in fitness, just as in interspecies relations. The relationship between variance and rate of change was derived by Fisher (1958) as the "Fundamental Theorem of Natural Selection," which includes the effects of intermating among all possible genotypes.

When considering the growth of a single population with the simplest model, the rate of increase is a logarithmic function of population size: $\frac{dN}{dt}=rN$. With more than one type in a population, r is dependent on the numbers and growth rates of the different member types, and growth is dependent on relative values:

$$\frac{dN}{dt}=N_1r_1+N_2r_2$$

in which $\bar{r}$ is the weighted mean of r_1 and r_2 and might be simply expressed as:

$$\bar{r}=\frac{N_1}{N}r_1+\frac{N_2}{N}r_2,$$

and where $N=N_1+N_2$ and the r_i are parameters for the intrinsic rate of increase specific to each type. Since the proportions $\frac{N_1}{N}$ and $\frac{N_2}{N}$ can be expected to change, $\bar{r}$ can change and hence we

can write:

$$\frac{d\bar{r}}{dt}=\frac{d[pr_1+(1-p)r_2]}{dt}$$

where $p=\frac{N_1}{N}$, $1-p=\frac{N_2}{N}$,

then, $\frac{d\bar{r}}{dt}=r_1\frac{dp}{dt}-r_2\frac{dp}{dt}$

$$=(r_1-r_2)\frac{dp}{dt}.$$

Since the proportions depend on the changes in numbers of both classes,

$$\frac{dp}{dt}=\frac{d(N_1/N)}{dt}$$

$$=\frac{1}{N}\frac{dN_1}{dt}-\frac{N_1}{N^2}\frac{dN}{dt}$$

$$=p(r_1-\bar{r})$$

and, therefore, $\frac{d\bar{r}}{dt}=p(r_1-r_2)(r_1-\bar{r})$

$$=p(1-p)(r_1-r_2)^2.$$

It can also be derived that if r_1 occurs with frequency p, and r_2 occurs with frequency $(1-p)$, the variance in r:

$$\sigma_r^2=p(1-p)(r_1-r_2)^2.$$

Therefore, the rate of change in average population fitness is exactly equal to the variance in the fitness values among the various types which exist in the population.

Applying the same principle to the genotypes which exist in intermating populations requires only the further derivation of the expected frequencies of the genotypes. Otherwise, we assume that the simple logarithmic growth potential of the population with sustained growth rates for each type is a reasonable model. If the population possesses two alleles $(A:a)$ at a locus with frequencies $p:(1-p)$, the three zygotic genotypes, $AA:Aa:aa$, may occur in almost any set of relative frequencies according to how the gametes are combined. If no forces, such as assortative or disassortative mating, or meiotic drive, affect the pairing of gametes, then the frequencies of the zygotic types in the population would be expected to be in proportion to the probabilities of random association. In large populations, these frequencies would be p^2 for AA, $2p(1-p)$ for Aa, and $(1-p)^2$ for aa, regardless of the zygotic arrangements in the parental class. The equilibrium frequencies induced by random mating have been derived elsewhere very well (in particular, see chapters 1 and 2 in Li (1955)) and

under the label of Hardy-Weinberg equilibrium. We can then define an average fitness ($\bar{w}$) as:

$$\bar{w} = W_{AA}p^2 + W_{Aa}2p(1-p) + W_{aa}(1-p)^2.$$

It is also necessary to define an average effect of an allele, since the fitness contribution of an allele can change if the allele is in a tree with genotype Aa or AA. A reasonable definition for the average effect of A is the fitness of the zygotic types weighted by the frequency with which it is associated with A (in AA trees) and a (in Aa trees) alleles. Thus,

$$W_A = pW_{AA} + (1-p)W_{Aa}$$
$$W_a = p(W_{Aa}) + (1-p)W_{aa}$$

and
$$\bar{w} = pW_A + (1-p)W_a$$
$$= p^2W_{AA} + 2p(1-p)W_{Aa} + (1-p)^2W_{aa}.$$

Therefore, $\frac{d\bar{w}}{dt}$ can be derived in terms of changes in p and hence in terms of $\frac{d\bar{w}}{dp}$ and $\frac{dp}{dt}$.

From the above, we can obtain:

$$\frac{d\bar{w}}{dp} = 2(W_A - W_a)$$

and as in the simpler nongenetic cases:

$$\frac{dp}{dt} = \frac{1}{N}\frac{dN_A}{dt} - \frac{p}{N}\frac{dN}{dt}.$$

Since we assume logarithmic growth rates,

$$\frac{dN_A}{dt} = W_A \cdot N_A$$

and
$$\frac{dN}{dt} = N\bar{w},$$
$$\frac{dp}{dt} = pW_A - p\bar{w}$$
$$= p(W_A - \bar{w}).$$

Therefore,
$$\frac{d\bar{w}}{dt} = \frac{d\bar{w}}{dp}\frac{dp}{dt}$$
$$= 2p(W_A - \bar{w})(W_A - W_a).$$

In Fisher's (1958) notation this is:

$$\frac{d\bar{w}}{dt} = 2pa\alpha,$$

where $a = W_A - \bar{w}$, the average excess of a gene substitution and $\alpha = W_A - W_a$, the average effect of a gene substitution. We can also rewrite the above formula as:

$$\frac{d\bar{w}}{dt} = 2p(1-p)(W_A - W_a)^2,$$

which is the variance in average fitness of the alleles. Thus, Fisher (1958) derives his "Fundamental Theorem of Natural Selection": the rate of increase in fitness of any population at any time is equal to its genetic variance in fitness at that time. This same result is derived in chapter 2 for selection models. It should be noted that the genetic variance used above is the variance among average gene effects and does not include all of the genetic variances among the three genotypes which may be due to such gene effects as dominance.

The genetic variation is therefore of central importance in predicting and directing future genetic changes in populations. Our concern in population genetics must therefore be focused on the variances as well as the means of populations.

The models used to derive these variances are clearly very simple ones and refer only to our parameterization of population growth. For models of population growth of wider generality, in which the assumptions of simple logarithmic growth are clearly not acceptable in important ways, more elaborate models of gene effects on population behavior are required. Nevertheless, additive models do provide reasonable first approximations which are of great direct value and which provide rational first steps in analyzing population behavior. We use them throughout this book as first approximations.

CHAPTER 7
REGRESSION AND REGRESSION EFFECTS OF GENOTYPIC DIFFERENCES

If mean trends and variances, especially those with genetic interpretations, are important analytical measures, then their definitions, derivations from biological models, and estimation must be well known. In this chapter, the concepts of simple linear regression are briefly reviewed and extended to gene effects. In particular, the sum of squares due to regression on genotypes is related to the concept of measuring the effect of genetic sources of variation. Several topics of special concern to forest geneticists, such as weighted and nonlinear regression and multivariate regression, are included but are not necessary for chapter continuity. Experimental design and variance components are treated in chapter 8.

LINEAR STATISTICAL MODELS

Genetic and environmental factors determine phenotypic expression through multiple and intricate physiological pathways. For simple biological models, mean effects, variances, and covariances of those causal factors on the dependent response of any measurable trait such as size, vigor, or reproductiveness can begin to describe the ways variations exist in natural populations. The sampling and description of a multitude of traits which are affected by age-dependent changes in tree populations are well developed in forestry. The description and analysis of phenomena such as populational cause-effect relationships, or gene actions which are not directly observable, are less well developed. To describe the relationships between such independent variables as fertilizer levels or genotypes and such dependent variables as volume growth or disease resistance, a model of their actions can be built and the parameters of the model both estimated and tested. However, the reduction of such cause-effect relationships to an explicit mathematical form is usually a crude approximation. Only the simplest linear functional relationships have been thoroughly studied to obtain good estimators and testing procedures. Nevertheless, extensions to nonlinear effects and interactions among multiple variables are being developed and are also becoming more generally useful. Linear models remain the basis for these extensions and have in their own right been found to be generally useful.

The linear model simply assumes that for any level of a causal, independently manipulatable factor (X), an average response of a dependent variable (Y) is expected. Specifically, for any changes in causal factor X, a proportionate change in dependent factor Y is expected. The proportionality factor depends on the scale of measurement of X and Y, but it is assumed either that the factor is constant over the scope of any experiment or that the variables can be measured on scales which will linearize their relationship. Generally, the effect of variations in X on Y cannot be measured directly, but differences among repeated trials can be measured, and while each trial has some error in exactly reflecting the proportionality factor, the average response is assumed to be a more precise estimate of the proportionality than any subsample. If another set of trials were made, however, it would result in a slightly different cluster of points with which to estimate the linear regression response. It would therefore be reasonable to not only want regression-line estimates that are as unbiased and as precise as possible, but also to want some idea of how well the estimation was done. Then, it would be possible to determine if an estimate of the regression based on a different sample is close enough to the first estimate that the two can reasonably be said to represent the same relationship. If they are too different, some factors could reasonably be inferred to have influenced the estimated relationships so as to make the proportionality factors distinctive. Clearly, the more variation around the regression and the larger the error in estimating the regression, the larger is the probable error of the estimates and the more difficult it is to distinguish truly different regressions from poorly measured ones.

In more explicit language, a single-variable linear regression model can be written:

$$Y_{ij}=\alpha+\beta X_i+\varepsilon_{ij}$$

where Y_{ij} = dependent variable measured at the jth trial at level i of the independent variable,

α = base level of response in the absence of X and ε effects,

β = proportionality factor (regression coefficient),

X_i = ith level of the independent variable,

ε_{ij} = deviation of the jth trial at X_i from the exact linear response.

More generally, there can exist several factors which simultaneously affect y, such as different genes or site factors. Then, a simple linear model for multiple X variables, say p of them, can be written:

$$Y_{ij}=\beta_0X_0+\beta_1X_{1ij}+\beta_2X_{2ij}+\beta_3X_{3ij}+\ldots\beta_pX_{pij}+\varepsilon_{ij},$$

where the substitution of β_0X_0 for α is a convenience.
For several samples, say n of them, and for the general case where

each sample varies the level of each of the X's sampled, we can fix the j subscript for each of the n trial samples and drop the i subscript since levels and samples change together in this simplified case. Then the equations for each sample are:

$$Y_1=\beta_0X_{01}+\beta_1X_{11}+\beta_2X_{21}+\beta_3X_{31}+\ \ldots\ +\beta_pX_{p1}+\varepsilon_1$$

$$Y_2=\beta_0X_{02}+\beta_1X_{12}+\beta_2X_{22}+\beta_3X_{32}+\ \ldots\ +\beta_pX_{p2}+\varepsilon_2$$

$$\ldots=\ldots+\ \ldots+\ \ldots+\ \ldots+\ \ldots+\ \ldots+\ \ldots$$

$$Y_n=\beta_0X_{0n}+\beta_1X_{1n}+\beta_2X_{2n}+\beta_3X_{3n}+\ \ldots\ +\beta_pX_{pn}+\varepsilon_n$$

In matrix notation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{01}X_{11}X_{21}X_{31}\ldots X_{p1} \\ X_{02}X_{12}X_{22}X_{32}\ldots X_{p2} \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ X_{0n}X_{1n}X_{2n}X_{3n}\ldots X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

$$\underline{Y} \quad = \quad (X) \qquad \underline{\beta} \quad + \quad \underline{\varepsilon}$$

Where Y is the $nx1$ column vector of the dependent variable or measures of yield, (X) is the nxp matrix of X's, $\underline{\beta}$ is the $px1$ column vector of regression coefficients, and $\underline{\varepsilon}$ is the $nx1$ column of error deviations.

Since the X's and Y's are known and we wish to estimate the p regression coefficients, we should have at least as many equations as unknowns and, therefore, it is required that sample size $n\geq p+1$. Since each equation also has an unknown error term, however, the actual Y values are not exactly what they would be predicted to be by the X and β values. Therefore, even if there are more equations than regression coefficient unknowns, they cannot be solved very simply and there actually exists a wide choice of ways to determine the coefficients. For one X and one Y variable, only two points are needed to determine a straight line relationship, but with many more points, each with some error, our choice is not obvious. A common procedure is to derive the least squares estimators in which the equations of the model are used to determine a simple function of the error deviations in terms of the X's, Y's, and β's, and from which an explicit expression for the β's are derived in terms of the X's and Y's.

Different criteria of a "good" estimator, such as unbiasedness, efficiency, consistency, and sufficiency, may be formed. They are thoroughly discussed in many theoretical statistics texts. Several methods that accomplish these objectives in different ways which may be more direct or more general than the suggested least squares have been devised. The generally satisfactory maximum likelihood estimators are examples. The interested reader is referred to Kendall and Stuart (1963) for a thorough introduction to these concepts and several estimation methods. In the important

case where the errors are normally distributed, the maximum likelihood and least squares estimators happen to be the same and are unbiased, efficient, and sufficient. In the following, the notation of Searle (1971) is followed closely. A reasonable function to use is the sum of squares of the error deviations ($SSE = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \ldots \varepsilon_n^2$) since its minimization would give us an intuitively good result. The least squares refers to the minimizing of these squares, and by doing so, a function of the X's, Y's, and β's can be developed which gives good estimates of the β's. To do this, it is convenient to express the ε's as follows:

$$\varepsilon_i = Y_i - \sum_j \beta_j x_{ij} \text{ for } i = 1, 2, \ldots n,$$

or $$\varepsilon = \underline{Y} - (X)\underline{\beta}$$

in matrix notation. In this case, the

$$SSE = (\varepsilon_1\ \varepsilon_2\ \varepsilon_3 \ldots \varepsilon_n) \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

$$= \varepsilon'\varepsilon$$

$$= (\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

To minimize SSE with respect to appropriate choices of the β_i,

$$\frac{\delta(SSE)}{\delta\underline{\beta}} = 0$$

$$\frac{\delta(SSE)}{\delta\underline{\beta}} = \frac{\delta(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})}{\delta\underline{\beta}}$$

$$= \frac{\delta(Y'Y - 2\underline{\beta}'X'Y + \underline{\beta}'X'X\underline{\beta})}{\delta\underline{\beta}}$$

$$= \frac{\delta(Y'Y)}{\delta\underline{\beta}} - \frac{2\delta(\underline{\beta}'X'Y)}{\delta\underline{\beta}} + \frac{\delta(\underline{\beta}'X'X\underline{\beta})}{\delta\underline{\beta}}.$$

Since $Y'Y = Y_1^2 + Y_2^2 + \ldots + Y_n^2$, it contains no explicit β values, and $\delta(\underline{Y'Y})/\delta\beta = 0$.

Since $\underline{\beta}'X'\underline{Y} = (\beta_1\ \beta_2 \ldots \beta_p)(X'\underline{Y})$, the β_1 factor is a multiplier for each of the elements of the first row of $(X'\underline{Y})$. Therefore, $\frac{\delta(\underline{\beta}'X'\underline{Y})}{\delta\beta_1}$ is the first row of $X'Y$. Similarly, $\frac{\delta(\underline{\beta}'X'\underline{Y})}{\delta\beta_2}$ is the second row of $X'Y$, and, therefore, for all of the β elements listed in

column order, $\frac{\delta(\underline{\beta}'X'\underline{Y})}{\delta\underline{\beta}}=X'\underline{Y}$.

Since $\beta'X'X\beta=(\beta_1\,\beta_2\ldots\beta_p)\,(X'X)\begin{pmatrix}\beta_1\\ \beta_2\\ \cdot\\ \cdot\\ \cdot\\ \beta_p\end{pmatrix}$,

and letting $X'X=\begin{pmatrix}a_{11}\,a_{12}\ldots a_{1p}\\ a_{21}\,a_{22}\ldots a_{2p}\\ \ldots\ \ldots\ \ldots\ \ldots\\ a_{p1}\,a_{p2}\ldots a_{pp}\end{pmatrix}$,

the β_1 factor on the left can be seen to be a multiplier of the first element of the product:

$$(X'X)\begin{pmatrix}\beta_1\\ \beta_2\\ \cdot\\ \cdot\\ \cdot\\ \beta_p\end{pmatrix},$$

which is $a_{11}\beta_1+a_{12}\beta_2+\ldots a_{1p}\beta_2$. The β_1 factor on the right can be seen to be a multiplier of $(\beta_1\beta_2\ldots\beta_p)\,(X'X)$, which is

$$a_{11}\beta_1+a_{21}\beta_2+\ldots+a_{p1}\beta_p=a_{11}\beta_1+a_{12}\beta_2+\ldots+a_{1p}\beta_p.$$

Therefore, $\frac{\delta(\underline{\beta}'X'X\underline{\beta})}{\delta\beta_1}$ is the first element of $2(X'X)\,(\underline{\beta})$.

Similarly, for the $\frac{\delta(\underline{\beta}'X'X\underline{\beta})}{\delta\beta_2}$, it can be determined that this is equal to the second element of $2(X'X)\,(\underline{\beta})$, etc., and, hence, that for all $\underline{\beta}$ listed in column order:

$$\frac{\delta(\underline{\beta}'X'X\underline{\beta})}{\delta\underline{\beta}}=2(X'X)\,(\underline{\beta}).$$

The equation for minimizing SSE can then be written:

$$\frac{\delta(SSE)}{\delta\underline{\beta}}=\frac{\delta(\underline{Y}'\underline{Y})}{\delta\underline{\beta}}-2\frac{\delta(\underline{\beta}'X'\underline{Y})}{\delta\underline{\beta}}+\frac{\delta(\underline{\beta}'X'X\underline{\beta})}{\delta\underline{\beta}}$$

$$0=0-2X'\underline{Y}+2X'X\underline{\beta}$$

Therefore, $2X'\underline{Y}=2X'X\underline{\beta}$ is the equation which must be satisfied if SSE is to be minimized. Then, this least squares estimate of $\underline{\beta}$ would be: $\underline{\hat{\beta}}=(X'X)^{-1}(X'\underline{Y})$, assuming that $X'X$ can be inverted. If these estimates are used and are substituted in the SSE equation for β:

$$SSE=\underline{Y}'\underline{Y}-2\beta'X'\underline{Y}+B'X'\underline{\beta};$$

then using $\underline{\hat{\beta}}$, $SSE=\underline{Y}'\underline{Y}-X'\underline{Y}(X'X)^{-1}X'\underline{Y}$

$$=\underline{Y}'\underline{Y}-\underline{\hat{\beta}}X'\underline{Y}.$$

The $\underline{\hat{\beta}}X'\underline{Y}$ is called the sum of squares due to regression (SSR) or the sum of squares of reduction, or any of several terms to denote the extent to which the error or unexplained portion of the original total sum of squares ($\underline{Y}'\underline{Y}$) has been reduced by having adjusted the variation around the regression.

In summary, the least squares equation,

$$\underline{\hat{\beta}}=(X'X)^{-1}X'\underline{Y},$$

is the explicit equation we sought in terms of known X's and Y's. It leads to a residual sum of squares or sum or squares for error of:

$$SSE=\underline{Y}'\underline{Y}-\underline{\beta}'(X'\underline{Y})$$

$$=\underline{Y}'\underline{Y}-SSR.$$

Hence, we maximize SSR and minimize SSE by the choice of $\underline{\hat{\beta}}$ which gives the least SSE.

Since the population responds imprecisely, any resampling of the population, even at the same X levels, would produce Y's which would exhibit some variations. Therefore, variations in estimating $\underline{\beta}$ depend on the behavioral variance in $\underline{Y}$. Thus, for fixed X's,

$$E(\underline{\hat{\beta}})=(X'X)^{-1}X'E(\underline{Y})$$

since $\underline{Y}=(X)\underline{\beta}+\underline{\varepsilon}$

$$E(\underline{\hat{\beta}})=(X'X)^{-1}X'E[(X)\underline{\beta}+\underline{\varepsilon}]$$

$$=(X'X)^{-1}X'XE(\underline{\beta})$$

$$=\underline{\beta},$$

proving unbiasedness of the estimator. Also the variance in $\underline{\hat{\beta}}$ is:

$$E[\underline{\hat{\beta}}-E(\underline{\hat{\beta}})]^2=[(X'X)^{-1}X'E((X)\underline{\beta}+\underline{\varepsilon})-\underline{\beta}]^2$$

$$=E[(X'X)^{-1}X'\underline{\varepsilon\varepsilon}'X(X'X)^{-1}]$$

$$=(X'X)^{-1}X'E(\underline{\varepsilon\varepsilon}')X(X'X)^{-1}.$$

The question of estimation error can now be stated in terms of the middle element of this equation, $E(\underline{\varepsilon\varepsilon}')$, the expected nature

of the squares and cross products of the error terms. If the errors are random samples of a homogeneous process and are independent among themselves, then they share a common expectation of ε^2, namely σ^2, and a common expectation of the covariance among any two errors $\varepsilon_i \cdot \varepsilon_j$, namely 0, when $i \neq j$. Then,

$$E(\underline{\varepsilon}\underline{\varepsilon}') = I\sigma^2$$

and $$V(\underline{\hat{\beta}}) = (X'X)^{-1}\sigma^2,$$

which can be simply written in the more familiar notation for a single X variable as:

$$\frac{\sigma^2}{\Sigma X_i^2}.$$

With more than one $\hat{\beta}$, $V(\hat{\beta})$ is the variance-covariance matrix of β's.

HETEROGENEOUS AND CORRELATED ERRORS*

It may often occur that the errors are not independent samples of a common set of ε, that $E(\varepsilon_i^2)$ varies among samples, and also possibly that they are correlated and hence $E(\varepsilon_i\varepsilon_j) \neq 0$. In such cases, if we let $E(\varepsilon\varepsilon')S\sigma^2$, then the $V(\hat{\beta})$ using the previously derived estimators would be:

$$V(\hat{\beta}) = (X'X)^{-1}X'S^{-1}X(X'X)^{-1}\sigma^2.$$

Instead of deriving the estimator to minimize $\Sigma_{ij}(\underline{\varepsilon}\underline{\varepsilon}')$ when $E(\varepsilon\varepsilon') = S\sigma^2$, it may well be better to choose $\underline{\hat{\beta}}$ to minimize $E(\underline{\varepsilon}\underline{\varepsilon}'S^{-1}) = I\sigma^2$. Then the sum of the variances is $tr(\varepsilon\varepsilon'S^{-1})$, which equals $tr(\varepsilon'S^{-1}\varepsilon)$. This latter function is a quadratic form and in the regression model is:

$$\underline{\varepsilon}'S^{-1}\underline{\varepsilon} = \underline{Y}'S^{-1}\underline{Y} - \underline{Y}'S^{-1}X\underline{\beta} - \underline{\beta}'X'S^{-1} + \underline{\beta}'X'S^{-1}X\underline{\beta}.$$

This error can now be minimized by forming the equation,

$$\frac{\delta(\underline{\varepsilon}S^{-1}\underline{\varepsilon})}{\delta\underline{\beta}} = 0$$

which yields: $$\underline{\hat{\beta}} = (X'S^{-1}X)^{-1}X'S^{-1}\underline{Y}.$$

It can then be shown that $E(\underline{\hat{\beta}}) = \underline{\beta}$

and that $V(\hat{\beta}) = (X'S^{-1}X)^{-1}\sigma^2$, where $V(\hat{\beta})$ is the variance-covariance matrix of $\hat{\beta}$'s.

If we now consider the design problem of a scientist wishing to estimate his β's particularly well, and having some choice in

*Graduate-level statistical training required for thorough understanding.

how he selects his independent X variables, we have an inverse of the former problem in that we try to minimize $\hat{V}(\beta)$ by choice of X's for given ranges of β. In simple cases such as that of the simple linear regression, it is only necessary to maximize ΣX^2 for any β. For more complicated cases, the solutions often are not independent of β and require that regions of β be specified for any optimum solution. The design matrix (X) may then be chosen to minimize some function of the variance-covariance matrix within that region. A further problem is then entailed since the relationship of the (X) matrix to the V matrix is most often nonlinear and exists in p^2 dimensional space. Search procedures on irregular surfaces in n-space are most easily carried out on computers by procedures such as those developed by Marquardt (1963) for single yield variables or single functions of the V matrix.

NONLINEAR REGRESSION*

Another form of the regression problem is the general nonlinear regression equation in which the parameters cannot be transformed into a linear function. While we can often separate a linear error element, the other variables are often nonlinearized. Thus,

$$Y=\beta_0+\beta_1 X^{\beta_2}+\varepsilon$$

is nonlinear in the parameters and cannot be transformed into a linear form. Therefore, the estimation equations cannot be neatly separated between the known X and Y versus the β to be estimated, and it is not a simple matter to find $\hat{\beta}$.

Most procedures derive a good approximation which is easy to compute and then derive successively closer approximations. One such procedure is to transform the function into an approximately equivalent but more easily soluble linear form by taking the function's Taylor Series expansion and dropping as many terms as feasible; as more terms are dropped, the approximation becomes worse but also easier to compute. The compromise is most often made heavily in favor of computational ease. Thus, for any function of β, we can write its Taylor Series expansion as:

$$f(\beta)=f(\beta^0)+(\beta-\beta^0)f'(\beta^0)+\frac{(\beta-\beta^0)^2}{2!}f''(\beta^0)+\frac{(\beta-\beta^0)^3}{3!}f'''(\beta^0)+\ldots$$

where β^0 is a first guess at the true β and $f'(\beta^0)$ is the first derivative of $f(\beta^0)$, $f''(\beta^0)$ is the second derivative, etc. The superscript refers to the first guess and would advance by one each time the guessing process is repeated. In elementary calculus, Newton's approximation to the roots of equations required that $Y=f(X)=0$. Then choosing a guessed root, X^0, as the first ap-

*Graduate-level statistical training required for thorough understanding.

proximation to the roots gives:

$$f(X)=0=f(X^0)+(X-X^0)f'(X^0)+\frac{(X-X^0)^2}{2!}f''(X)+\ldots.$$

Then if the first guess is reasonably close, $\frac{(X-X^0)^2}{2!}$ will be small, as will all higher coefficients. Then, what we hope is a reasonable first approximation is:

$$0=f(X^0)+(X-X^0)f'(X^0)$$

or $X-X^0=\frac{-f(X^0)}{f'(X^0)}$.

In the regression case, we wish to find a minimum for the *SSE* and this process gives us a function of β:

$$\frac{\delta(SSE)}{\delta\beta}=0=f(\beta).$$

Then: $\hat{\beta}-\beta^0=-f(\beta^0)[f'(\beta^0)]^{-1}$

is our first approximation

where $f'(\beta^0)=\frac{\delta^2(SSE)}{\delta\beta^2}\Big|\beta=\beta^0$

For the multiple regression case of several β's, the $f(\beta^0)$ becomes the vector of functions $\frac{\delta(SSE)}{\delta\underline{\beta}}$ for each β and the $f'(\beta^0)$ becomes the matrix:

$$\begin{bmatrix} \frac{\delta^2(SSE)}{\delta\beta_0^2} & \frac{\delta^2(SSE)}{\delta\beta_0\beta_1} & \frac{\delta^2(SSE)}{\delta\beta_0\beta_2} & \cdots \\ \frac{\delta^2(SSE)}{\delta\beta_1\beta_0} & \frac{\delta^2(SSE)}{\delta\beta_1^2} & \frac{\delta^2(SSE)}{\delta\beta_1\beta_2} & \cdots \\ \cdots & \cdots & \cdots & \end{bmatrix}$$

The subscripts in the matrix refer to the variable of the regression equation for this estimation. Using the guessed values in these equations provides an approximation to our best estimator $(\beta-\beta^0)$. Increasing $\hat{\beta}^0$ by this amount gives us a new approximate $\hat{\beta}^1$. Using superscripts to denote the number of iterated estimates, and recognizing that $\hat{\beta}^1$ is only approximate, another estimate, $\hat{\beta}^2$, may be needed. Then, the $\hat{\beta}^1$ is used in the same formulas to provide a $\hat{\beta}^2$, etc., where the matrix above of $f'(\beta)$ changes at each iteration until the $\beta^i-\beta^{i+1}$ corrections are negligible.

For the case of $Y=\beta_0+\beta_1X^{\beta 2}+\epsilon$, the procedure simplifies still further (Namkoong and Miller 1968) in that the correction vector

$\beta^i - \beta^{i+1}$ is a function of only one of the $\hat{\beta}$ coefficients, and, hence, the sequence of estimates is not difficult to handle on a computer.

For the simple case of linear regression it can readily be derived that:

$$\delta(SSE) / \delta\beta = 2X'(\underline{Y} - X\underline{\beta})$$

$$\delta^2(SSE) / \delta\underline{\beta}^2 = -X'X$$

$$\delta^3(SSE) / \delta\underline{\beta}^3 = 0$$

and the exact solution is:

$$\beta - \beta^0 = -[f(\beta_0)]\ [f'(\beta_0)]^{-1}$$

and for $\beta^0 = 0$,

$$\beta = (X'X)^{-1}X'Y.$$

For more difficult nonlinear cases, the general problem is to find the $\underline{\beta}$ which minimizes the value of SSE over the entire surface of SSE values created by all possible choices of $\underline{\beta}$. In general, the surface will be irregular. If considered in this way, it is clear that the linear regression SSE is a simple quadratic surface $(Y - X\underline{\beta})'(Y - X\underline{\beta})$ where SSE is high at values of β which are far from our estimator $(X'X)^{-1}\ (X'Y)$ and low near that vector value. For example, with two β's and a single SSE, we can visualize a parabaloid, concave upward in the direction of increasing SSE and coming to a minimum at the point we would choose as the β_1 and β_2 coordinates (fig. 12). For less regular surfaces, search procedures have been developed for computers which use a variety of techniques to efficiently locate the $\underline{\beta}$ combination corresponding to the minimum SSE. The problems of finding the lowest minimum are more than computational where more than one local minimum exists and there are flat areas around the minima. The existence of such surfaces also indicates that widely different $\underline{\beta}$ may be almost equally good and hence that the model or the data cannot discriminate very well among widely different $\underline{\beta}$ estimates. In such cases, it behooves the analyst to consider the adequacy of his model or experimental design matrix. Nevertheless, programs such as Marquardt's (1963) can be used to estimate $\underline{\beta}$ and approximate variances of the estimates. Hartley's (1964, 1969) procedures can often be used to obtain exact confidence regions for $\underline{\beta}$.

MULTIVARIATE REGRESSION*

An expansion of simple linear regression of a different order is to multivariate models in which more than one yield variate Y_i is affected by the X variables. It is a particularly useful extension

*Graduate-level statistical training required for thorough understanding.

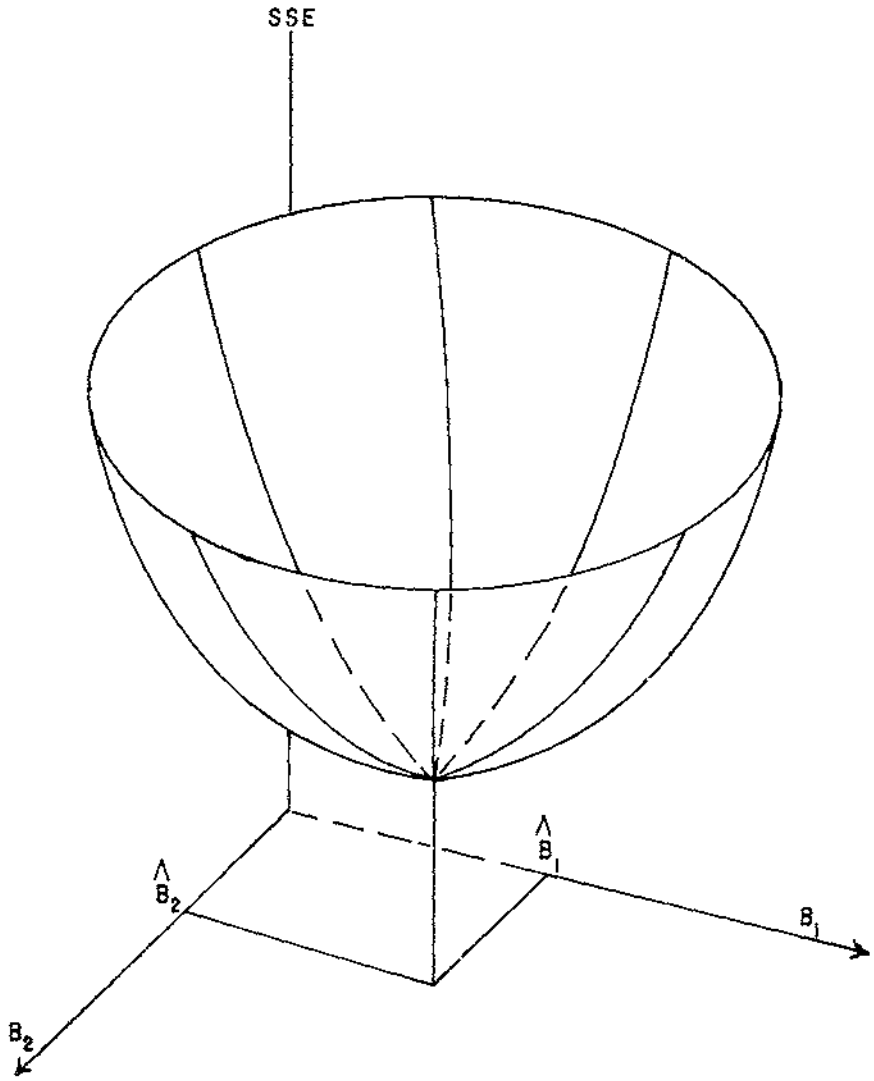


Figure 12.—A simple quadratic error surface created by relationship of the error sum of squares to the choice of β_1 and β_2 for a set of data.

in forestry where the cost and duration of experiments make it desirable to measure the response of several dependent yield variates in any one experiment. For tree breeders, an important problem lies in determining if between-population selection is more or less effective than within-population selection, whether the relative effectiveness varies among yield traits, and how information on the source environment can help define provenances and the location of good sources. When several intercorrelated yield variates determine the value of a provenance and each variate is affected by the same environmental variables, but in different ways,

the breeder must parameterize these relations in order to select among provenances. This parameterization is essentially multiple regression extended to study several traits simultaneously. For example, height growth and diameter growth are simultaneously affected by some independent variables, both are often measured, and both are often affected by genetic or soil factors in what may generally be similar ways. Of course, the dependent yield variates are not exactly correlated in their responses and our interest therefore centers on the pattern and strength of the joint responses. The greater the correlation among the yield variates, the simpler the problem becomes since the results in several variates can be predicted with increasing accuracy on the basis of the behavior of any one variate. In such cases, a single (or very few) functional relationships among variates would reduce our problem to univariate analysis, which we can choose either as a linear function of all the variates or a single convenient variate for regression analysis and predict the behavior of all other variates by that function. In that case, the only problem for the breeder is that of determining a value function among the yield variates. This determination can be an additional and a critical problem if the values of the variates do not assume a linear form. Consider, for example, a curvilinear relationship in the set of points representing the trees in terms of their yield variates such that at low stem growth rates, fruit yield is low, but that fruit yield increases with increased growth vigor up to a point beyond which increased growth rates are made at the expense of a decline in fruiting. If managed for the dual yield of stem and fruit, a problem can be seen to exist in deciding which combinations are best. The answer can vary widely, depending on the relative value placed on the two traits.

The nature of irregular value functions and their uncertainty is an acute economic problem which we will not consider at this point. For simpler linear models, however, extensive theory and methodologies have been developed. The analytical problem of describing the joint distributions of points in multivariate fields and analyzing regression functions on them is immense.

The serious investigator would profit by study of the distribution theory (Anderson 1958) and analytical methods as detailed in several texts such as Blackith and Reyment (1971).

To briefly familiarize the reader with some of the concepts, a few of the multivariate analogs of elementary univariate statistics might be useful to describe. Whereas, in the univariate case, a normal distribution has a probability density function (*pdf*) of

$$pdf = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\tfrac{1}{2}(X-\mu)^2/\sigma^2\right]$$

In the multivariate normal distribution, there are means and variances for each variate and also, covariances between them. The covariance matrix (A) takes the place of the variance and the

vector of $(\underline{X-\mu})$ takes the place of the $(X-\mu)$ in the univariate case. Then for the multivariate case:

$$pdf = \frac{1}{|A|^{1/2}\,(2\pi)^{p/2}} \exp\left[-\tfrac{1}{2}\,(\underline{X-\mu})'A^{-1}(\underline{X-\mu}\right]$$

In the bivariate case for variates x and y, where $x = X-\mu_x$ and $y = Y-\mu_y$,

$$A = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

and the $$pdf = \frac{|A|^{-1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\,\frac{(x^2-2xy\rho+y^2)}{\sigma_x^2\,\sigma_y^2\,(1-\rho^2)}\right]$$

where $\rho = \sigma_{xy}/\sqrt{\sigma_x^2\sigma_y^2}$, and exp [ ] is the exponentiation operation.

In regression theory, with a single y variate and single x variable, we had $f(y \mid x)$ as a mean of

$$\overline{Y} + \frac{\sigma_{yx}}{\sigma_x^2}(X-\mu_x)$$

and residual variance:

$$\sigma_y^2 - \frac{(\sigma_{xy})^2}{\sigma_x^2} = \sigma_y^2(1-\rho^2).$$

For multiple x variables, we have a regression:

$$(\sigma_{yx_i})\ (\Sigma_{xx})^{-1} f(y \mid x_1, x_2, \cdots)$$

with mean

$$\overline{y} + \begin{bmatrix} \sigma_{yx_1} \\ \sigma_{yx_2} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \Sigma_{xx}^{-1} \begin{bmatrix} X_1-\mu_{x_1} \\ X_2-\mu_{x_2} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

and residual variance:

$$\sigma_y^2 - (\sigma_{yx_i})'\,\Sigma_{xx}^{-1}\,(\sigma_{yx_i})$$

where Σ_{xx} is the covariance matrix among the x's. σ_{xy} is the covariance vector between y and the x's.

For multivariate regression with several x variables and y variates:

$$f(y_1, y_2 \ldots \mid x_1, x_2 \ldots)$$

has regression matrix estimates: $\Sigma_{xy}\,\Sigma_{xx}^{-1}$

has mean vector: $\underline{\mu_y} + \Sigma_{xy}\,\Sigma_{xx}^{-1}\,(\underline{x-\mu})$

and has residual covariance matrix: $\Sigma_{yy} - \Sigma_{xy}\,\Sigma_{xx}^{-1}\,\Sigma_{xy}$

where Σ_{yy} is the covariance matrix of y's

and Σ_{xy} is the covariance matrix of x's and y's.

Simple tests are also closely analogous to the univariate case.

In univariate t-tests, $t=\sqrt{N}\dfrac{(\bar{x}-\mu)}{\sigma}$

or $$t^2=N\frac{(x-\mu)^2}{\sigma^2}$$

and t^2 is distributed as an $F_{(1,\ N-1)\ df}$.

In multivariate tests, $$T^2=N(\underline{x-\mu})A^{-1}(\underline{x-\mu})$$

and $$\frac{T^2}{N-1}\frac{(N-p)}{p}\sim F_{(p,\ N-p)df}$$

where p=number of variates. For our purposes, it is only necessary at this time to consider the desirability of reducing the number of yield variates and how ordinary regression theory can be applied to problems involving more than one dependent variate. In simple multiple regression it is assumed that the relationships are linear between the independent regression variables and the dependent yield variates. Similiarly, for the simpler analyses, it is also generally assumed that the relationships among the dependent variates are linear. The problem of nonlinearity is exceedingly difficult to handle, since it requires that nonlinear multivariate distributions be specified in such a way that moments can be derived and the effects of independent regression variables also are derivable. It seems best at this time, when few data analyses are available, to linearize the joint measures as much as possible and to use standard linear theories until the effects of known nonlinearities can be predicted.

The first problem in most practical breeding studies is identification of the environmental variables of importance to one or more of the height, volume, or other dependent yield variates. In this part of the problem, one use of multivariate techniques is to reduce the number of yield variates that must be measured. The same techniques may be applied to the multiple environmental variables for elimination of variate redundancies (Kendall and Stuart 1966). The reduction can be accomplished by component analysis rather than by multiple regression. If, to start with, independent p variables may account for part of the genetic variance, the problem is to find whether collinearities exist among any subset of the variables. A collinearity exists between two variables when the occurrence of one variable at a specific level fully determines the other. A single linear relationship that completely describes the joint variation can then substitute for the two original variables, or, conversely, one of the variables is redundant. Similar reductions in redundancies may often be a significant aid in data interpretation. If there are three variables and

one is fully dependent on the other two, a collinearity in three-dimensional space exists and all of the variation is in a two-dimensional plane. Then, a single collinearity is said to exist and the rank of the space is two. If, in addition, all of the variation reduces to a single line, a second collinearity exists and only one dimension is required to include all of the variability in the three original traits. In component analysis, a series of lines (in the original p dimensions) is successively and orthogonally fit to reduce the residual variance about the lines. These lines are the principal component vectors. If a single line describes all of the variation in the variables, the first component would be that line. (The line is given in terms of its direction cosines in the space defined by the original variables.) If only a single collinearity exists, the remaining variability about the first line is all in one plane and hence is reducible to a single line which we choose to be orthogonal to the first.

These relations are illustrated in figure 13, in which the perpendicular line between the sample point and the first principal component is seen to lie in three, two, or one dimension as dimensionality decreases from three to one. If fewer than p one-dimensional transforms are required to account for almost all of the variance, then there are, perforce, linear dependencies among some of the p original variables. Dependencies imply that some of the variates which have been measured can be fully explained or replaced by a linear function of other variates. Therefore, the removal of at least one is desirable. A procedure utilizing the principal components already derived can be useful. The vector corresponding to the smallest root of the standardized covariance matrix (i.e., the correlation matrix) presumably represents almost a random vector in the orthogonal residual space. The variate which, in this vector, has the largest coefficient is that which can presumably be best explained by the others and hence is a likely candidate for discarding. The estimation of further components may then be repeated for discarding variates for as long as zero or near zero roots continue to exist.

This component analysis procedure for reducing dependent variates can also be applied to reduce the number of independent variables (Kendall 1961). Like other regression techniques which reduce the variate space, this method is subject to the usual restrictions on interpretation of cause and effect or anything other than simple association. It does possess some advantages over the more common stepwise procedures, but it is scale dependent. The standard procedures for variable reduction in multiple regression may therefore be more useful parts of an analytical system. When interest exists in finding the linear function of independent variables (X) which can best fit a linear function of dependent variates (Y), the regression coefficients for the X variables and the component coefficients for the Y variates can be simultaneously chosen to minimize the error variance of the principal component. The technique is known as canonical correlation analysis, and

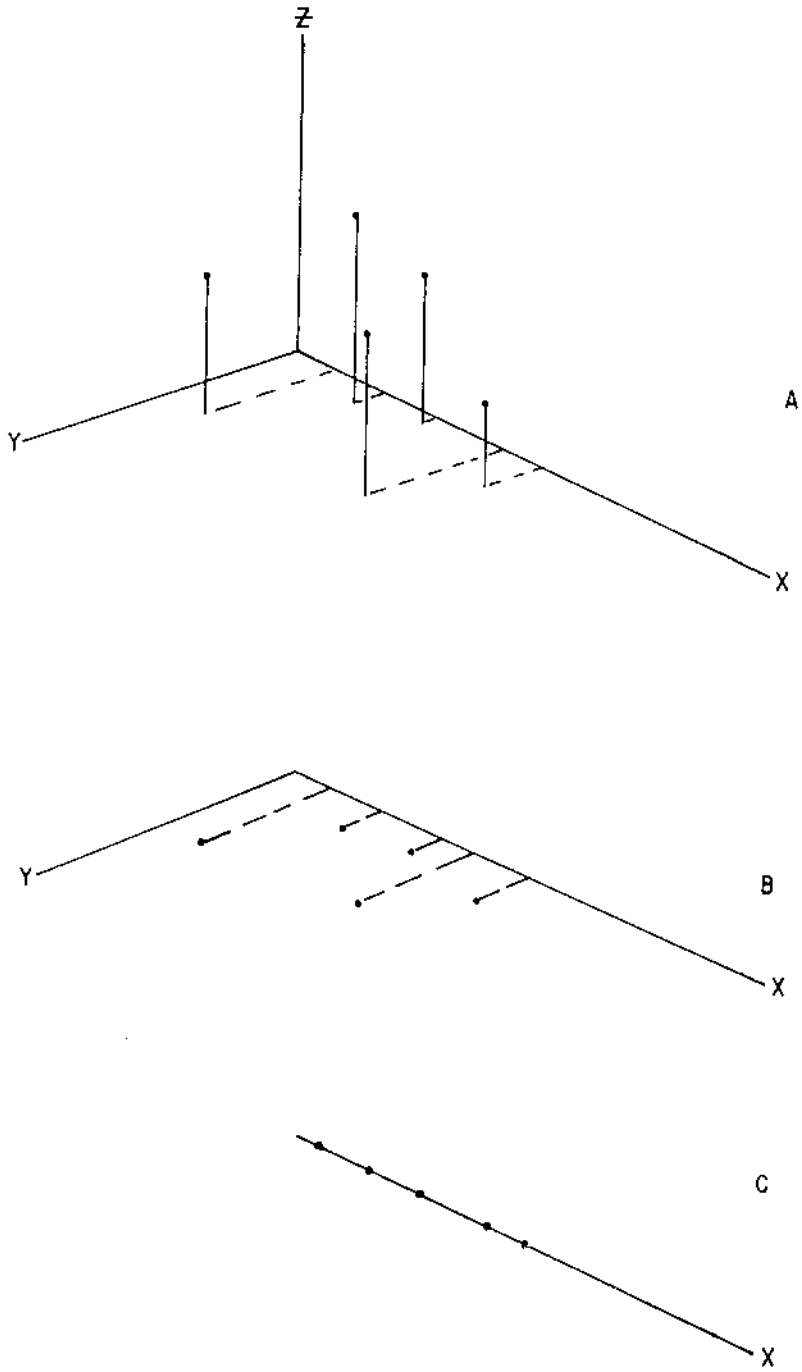


Figure 13.—Location of five sample points where variation occurs in three dimensions (A), two dimensions (B), and one dimension (C).

though not yet used very extensively in forestry, it is potentially useful in provenance analyses where several traits vary simultaneously in response to several site variables and the greatest degree of explainable variation is desired.

If consideration is restricted to a reasonable number of dependent yield variates and independent environmental variables, the problem of parameterizing the joint relations is often simply one of describing the regression effects. In the univariate case, a response surface is estimable and the combination of source-environment variables which gives the greatest expected progress in the dependent yield variate is identifiable. For example, if a simple quadratic surface is hypothesized and estimated, the estimated optimum levels of the X environmental variables that would give a maximum yield are derivable. For example, the dependent Y variable may be some measure of growth and the independent X variable may be fertilizer level or the latitude of the seed source, both of which may have some intermediate optimum level for maximum Y. The error in estimating the optimum point for maximum Y is also estimable. If a simple quadratic-response line of yield to a single environmental variable, for example, latitude, were to be estimated by $\hat{Y}=\hat{b}_0+\hat{b}_1 X+\hat{b}_2 X^2$, then the maximum likelihood estimate of the X corresponding to a maximum Y is: $-\frac{1}{2}\frac{\hat{b}_1}{\hat{b}_2}$. The standard error of the maximum Y can be estimated by such procedures as given by Kendall and Stuart (1963, ch. 10, p. 232) and the confidence belt estimated as that which would be appropriate for regressions derived from a normal distribution of errors of Y. Regions in which the environmental variables are of acceptable levels may then be set up by simple graphical or more sophisticated techniques. We may extend the number of different environmental variables to two or more, but remain in the univariate case and describe a quadratic response surface, for example, by:

$$\hat{Y}=\hat{b}_{00}+\hat{b}_{10}X_1+\hat{b}_{20}X_1^2+\hat{b}_{01}X_2+\hat{b}_{02}X_2^2+\hat{b}_{11}X_1X_2.$$

The levels of X_1 and X_2 giving maximum Y are:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = -\begin{bmatrix} 2\hat{b}_{20} & \hat{b}_{11} \\ \hat{b}_{11} & 2\hat{b}_{02} \end{bmatrix}^{-1} \begin{pmatrix} b_{10} \\ b_{01} \end{pmatrix}$$

Other nonlinear models increase the difficulty of estimation but are often more precise and useful (Anonymous 1961). The estimation of the regression coefficients of the linear model is well known and can be written in terms of estimating the vector $\hat{\underline{\beta}}$ by $(X'X)^{-1}X'\underline{Y}$, where X is the matrix of the levels of the sampled variables and $\underline{Y}$ is the vector of the yield variate at each level of the independent variables. If the model is extended to the

multivariate case, each dependent yield variate would be characterized in its response to the environmental variables by its specific vector of regression coefficients.

The estimation is a simple extension of the usual univariate procedures for estimating a $\hat{\beta}$ matrix instead of a vector. In linear models, the matrix $\hat{\beta}$ is estimated by $(X'X)^{-1}X'Y$ where the only difference from the univariate case is that Y is a matrix. Each row of the Y matrix is the vector of yield responses with respect to a given set of levels of the X variables, and the columns of the Y matrix represent the response of each yield variate to the series of sampled X variables. For n samples and one dependent variate,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{01} & X_{11} & X_{21} & \dots & X_{p1} \\ X_{02} & X_{22} & X_{22} & \dots & X_{p2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ X_{0n} & X_{1n} & X_{2n} & \dots & X_{pn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

For the same n samples and two dependent variates $\underline{Y}$ and $\underline{Z}$,

$$\begin{pmatrix} Y_1 & Z_1 \\ Y_2 & Z_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ Y_n & Z_n \end{pmatrix} = \begin{pmatrix} X_{01} & X_{11} & X_{21} & \dots & X_{p1} \\ X_{02} & X_{12} & X_{22} & \dots & X_{p2} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ X_{0n} & X_{1n} & X_{2n} & \dots & X_{pn} \end{pmatrix} \begin{pmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & \beta_{21} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \beta_{1n} & \beta_{2n} \end{pmatrix} + \begin{pmatrix} \varepsilon_{10} & \varepsilon_{21} \\ \varepsilon_{11} & \varepsilon_{22} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \varepsilon_{1n} & \varepsilon_{2n} \end{pmatrix}$$

Since the X matrix remains the same, (X) can be used as it was for simple regression, and changing the $\underline{Y}$ vector to $\underline{Y}_1$, the $\underline{Z}$ vector to $\underline{Y}_2$, and the subscripts on the β and ε elements by adding 1, 2, . . . etc., for the $\underline{Y}$ variate to which they refer,

$$(Y) = (X)\ (\beta) + (\varepsilon).$$

Using $\quad \hat{\underline{\beta}}_1 = (X'X)^{-1}X'\underline{Y}_1,$

and $\quad \hat{\underline{\beta}}_2 = (X'X)^{-1}X'\underline{Y}_2,$ etc.,

the matrix $\quad \hat{\beta} = (X'X)^{-1}X'\ \underline{Y}_1\ \underline{Y}_2 \dots$

$$= (X'X)^{-1}\ (X')\ (Y),$$

or in the Σ matrix notation $\quad = \Sigma_{XX}^{-1}\ \Sigma_{XY}.$

The covariance between the regression coefficients is estimated by $\Sigma_{XX}^{-1}S_{ij}$, where S_{ij} is the residual covariance between traits Y_i and Y_j.

The data of Wells and Wakeley (1966) on the performance of loblolly pine of various seed sources in a plantation in Dooly County, Georgia, will serve as an example. The independent X variables are the January minimum temperature (X_1) and summer rainfall (X_2) of the source locations, and the Y variates are survival (Y_1) and height (Y_2). These data are listed in table 4, and regressions are drawn in figure 14. The regression equations estimated on the nine source locations are:

$$\hat{Y}_i = b_{i00} + b_{i10}X_1 + b_{i20}X_1^2 + b_{i01}X_2 + b_{i02}X_2^2 + b_{i11}X_1X_2$$

$$\hat{Y}_1 = 62.9 + 2.91X_1 - 0.06X_1^2 - 1.98X_2 - 0.09X_2^2 + 0.09X_1X_2$$

$$\hat{Y}_2 = 135.5 - 6.79X_1 + 0.11X_1^2 + 0.59X_2 + 0.09X_2^2 - 0.08X_1X_2.$$

The total variance in Y_1 is 27.8, and in Y_2 is 3.10, and the covariance between them is -7.89. In matrix form, this total covariance matrix is:

$$\Sigma_{YY} = \begin{pmatrix} 27.8 & -7.89 \\ -7.89 & 3.10 \end{pmatrix}$$

After adjusting for regression, the residual covariance matrix is:

$$\Sigma_{YY} - \Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{XY} = \begin{pmatrix} 11.6 & -0.452 \\ -0.452 & 0.269 \end{pmatrix} = \begin{pmatrix} S_{ij} \end{pmatrix}$$

The sampling variance for all of the regression coefficients involving Y_1 is a product of 11.6 and the appropriate element of Σ_{XX}^{-1}. The sampling variance for those involving Y_2 is a product of 0.269 and the appropriate element of Σ_{XX}^{-1}, and the covariance of Y_1 and Y_2 is a product of -0.452 and the appropriate element of Σ_{XX}^{-1}.

The upper triangle of the Σ_{XX}^{-1} matrix is:

$$\begin{bmatrix} 9.819 & -0.1528 & -2.260 & -0.0641 & 0.1079 \\ & 0.0025 & 0.0668 & 0.0008 & -0.0024 \\ & & 8.895 & -0.0677 & -0.1876 \\ & & & 0.0020 & 0.0003 \\ & & & & 0.0048 \end{bmatrix}$$

Hence, the sampling variance of b_{110} for Y_1 is (11.6) (9.819) = 113.9 and of b_{210} for Y_2 is (0.269) (9.819) = 2.64. The sampling variance of b_{101} for Y_1 (11.6) (8.895) = 103.24 and b_{201} for Y_2 is (0.269) (8.895) = 2.394. The sampling covariances of b_{110} and b_{101} are similarly derived as (11.6) (-2.260), and the sampling covariance of b_{101} for Y_1 and of b_{201} for Y_2 is (-0.452) (9.819) = -4.439.

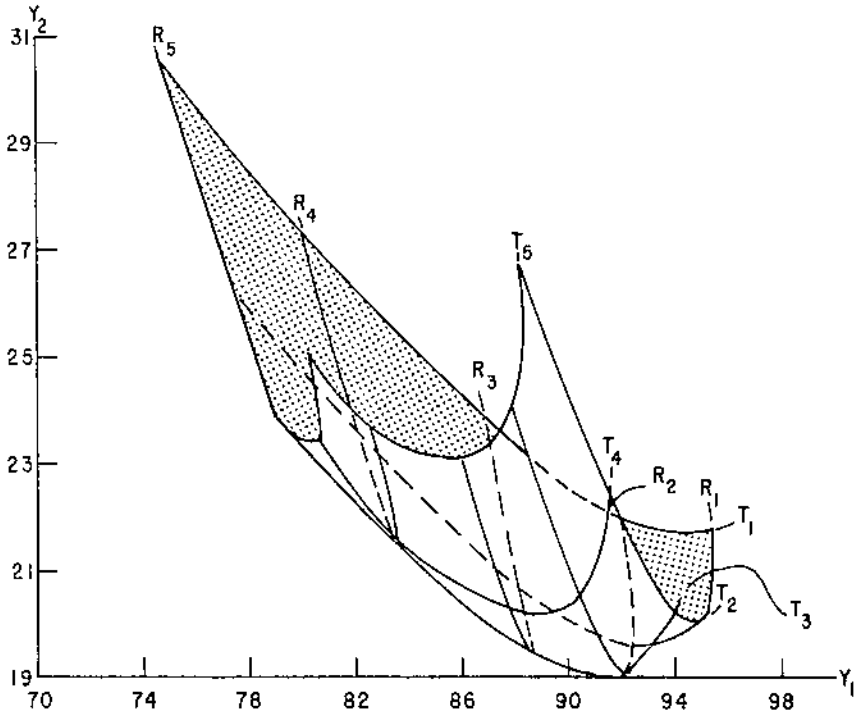


Figure 14.—Quadratic response surface of the simultaneous effects of rainfall and temperature on survival and height growth.

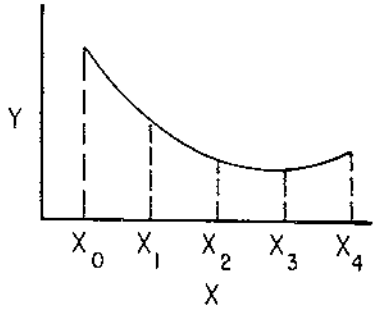
Table 4.—Loblolly pine seed and source performance in Dooly County, Georgia
(Wells and Wakeley 1966, tables 2 and 13)

Seed source	Source		Performance	
	January minimum temperature (X_1)	June–August rainfall (X_2)	Survival (Y_1)	Height (Y_2)
	°F	*Inches*	*Percent*	*Feet*
Eastern Maryland	31.4	12.9	87.3	22.3
Southeast North Carolina	36.9	19.0	73.3	24.2
Eastern North Carolina	35.9	18.4	82.2	23.2
Southwest Georgia	40.3	15.4	80.4	23.2
Northern Alabama (1)	32.7	12.7	85.8	20.4
Northern Alabama (2)	34.9	13.9	85.6	20.4
Southeast Louisiana	42.5	16.9	77.3	25.3
East Texas	38.8	9.2	85.3	23.3
East Arkansas	33.2	10.4	91.2	20.6

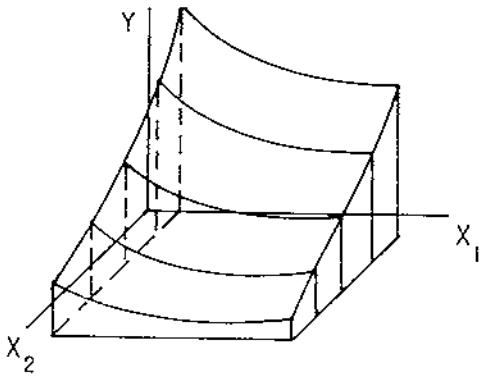
A problem for the breeder in provenance selection is how best to find a region for seed source sampling or its unique set of en-

vironmental conditions that simultaneously, optimally affect the important yield variates. If we establish a space defined by these dependent variates, it would be possible to locate a point representing the vector of the several dependent variates corresponding to each set of environmental conditions. A combination of independent variables denoted by a vector valued $\underline{X}$ is thus associated with a vectorial representation of the dependent variates. Changes in $\underline{X}$ over the independent variables, which may be environmental variations, define a surface of changing values in the $\underline{Y}$ variates. The problem then is to define some joint evaluation of all $\underline{Y}$ variates and then to find the maximum value of that joint value function. An optimum $\underline{X}$ may not be unique, however, depending on the shape of this $\underline{Y}$ surface and the value function used.

Geometrically, the simple regression problem is to define some functional relationship between an independent variable X and an average dependent response variate Y.

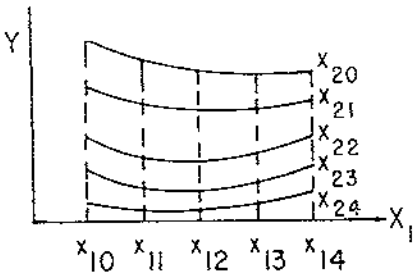


Expanding the case to more than one X in multiple regression requires description of the response surface of Y to the X's.

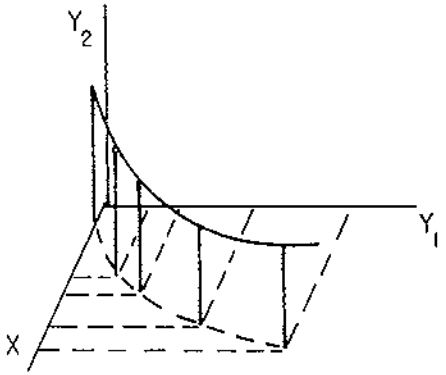


This may be projected onto the Y, X_1 plane when X_{20}, X_{21}, X_{22}, and X_{23} are the projections of Y at the various levels of X_2. Considering now that two dependent Y variables may exist for a single independent X variable, there generally is only one mean

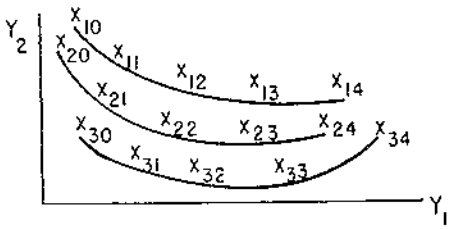
joint response point (Y_1, Y_2) for each point in X. In that case, the response of Y_1 and Y_2 to X_1 is a line in three dimensions.



This may be projected onto the Y_2, Y_1 plane.



Then similar variations in a second X variable can be projected onto this plane.



This kind of projection then graphically displays the joint response of two dependent Y variates to two independent X variables. It should be noted, however, that the proviso was stated that there is only one mean joint response point to each point in X. In fact, however, there are residual variances, in each Y variate, and some residual covariances and correlated response of Y_1 and Y_2 at each X point. In terms of provenance analyses, these are the intrapopulational variances and covariances which may be of considerable interest.

More formally, the linear relations are describable by the equation $\underline{Y}=\underline{X}\ (\beta)$ where $\underline{Y}$ is the vector of yield values, $\underline{X}$ is the vector of the environmental variables, and β is the matrix of regression coefficients previously described. The problem is to pick the vector $\underline{X}$ which will maximize some appropriate value function of $\underline{Y}$. Since these relations are scale/dependent, reduction to a standardized or other established basis is recommended. Simple linear combinations of the variates may always be constructed, but often the gain achieved by selecting for conditions maximizing one trait will not be optimum for other traits. It may often be best to pick the environmental vector which will assure that none of the variates suffer much loss and which thus will maximize the minimum gain. Of course, it is possible to construct solutions to maximize the sum or product of all variates. For any criteria, the surface representation will allow one to examine how the alternative solutions will differ and will be helpful in determining intermediate solutions.

The conclusion of Wells and Wakeley (1966)—that the fastest growing trees in the Dooly County, Georgia, plantation were from regions with warm winters and wet summers—is generally supported by the present anaysis. While the quadratic surface model may not be the best fit for the data, it is evident that warm winter at the seed source favors growth but not survival and that low summer rainfall at the seed source favors survival but not growth (fig. 14.). The sampling of source environments in this plantation is not sufficient for more precise source evaluations. It is generally not wise to extrapolate from a model which at best may mimic a set of data, and to try to discern cause and effect, particularly when such high residual errors exist. However, the generally negative correlation between height and survival for both the total variables and the residuals after regression indicates that selection for both would result in opposing selection forces. The deviations from regression allow for some combined selection, and the data suggest that deviations exist in the direction of intermediate to high winter temperature and intermediate to low summer rainfall. If selected, these sources might best maintain both height and survival without great loss in either. Provenance hybridization to combine genes for both may be warranted. If survival is of less importance in this range, the desirable vector of selection would more heavily favor the warm winter-wet summer sources.

While the regressions are not well established, it is illuminating to consider how different vectors might influence the choice of source environments. An infinite number of functions could be written for the relationships between the yield variates. If equivalent and linear economic weight is given to survival in percent and height in feet, the expected value would be:

$$E\ (Y_1+Y_2)=198.4-3.88X_1+0.05X_1^2-1.39X_2+0.01X_1X_2.$$

Here X_1 has about twice the weighting of X_2. If height is so im-

portant that the scale of value is Y_1+10Y_2, the regression would be:

$$E\ (Y_1+10Y_2)=1{,}418-65.0X_1+1.04X_1^2+4.0X_2$$
$$+0.81X_2^2-0.71X_1X_2.$$

In this linear weighting of Y_1 and Y_2, the value (or objective) function is a straight line. In a multiplicative model of value, the value function would be a hyperbolic line. In the linear form, and with a 1:10 weighting, a negative weighting is given to winter temperature (X_1) and a positive weighting to summer rainfall (X_2). Both linear cases indicate that several X_1, X_2 combinations could provide good value yields but also that there is a unique minimum. Since only minima exist, the analysis is useful for indicating unique sources to avoid rather than unique sources to choose. On the other hand, hybrids of the alternately good sources may be quite valuable.

These regression solutions for provenance selection are simply multivariate extensions of univariate theory. Estimation problems in multivariate analysis are only slightly more involved, though distribution theory is often much more complicated.

In addition to the statistical questions, the chief difficulties are in deciding whether to put selection emphasis on between-provenance or within-provenance differences and in interpreting the pattern of genetic variance and covariance in terms of population structure. These methods merely facilitate consideration of the joint processes of several variates under the influence of several environmental variables. Since many problems in forestry involve the simultaneous evaluation of several traits on single trees or populations due to the multiple values that exist in forests, multivariate analyses are likely to be more commonly important in forestry than in other agricultural sciences. Not only are multiple uses of forest lands commonly required, but over the duration of a single forest population, changing uses will be imposed by human activities. Thus, for the forester, it would be most appropriate to consider univariate analyses as a special case of multivariate analyses rather than the multivariate case as an extension of the univariate case. Regardless of the need, we are still largely limited to univariate model approaches to genetic analyses.

LINEAR GENETIC MODELS

In this section of the chapter, the genetic variances are defined more exactly than in chapter 2. In the next section, the relationship between genetic variances and the variation between and within families is established. In the next chapter, some uses of the genetic variance between and within families are examined.

The commonly used models of gene action are simple extensions of the usual linear regression theory with the further complication that polygenic effects cannot be directly observed. While it is

always preferable to work with easily measurable, clear gene differences, this is not often possible in forestry and we shall limit our attention to genes with small effects relative to error variation in their expression. However, while it may be difficult to measure single gene effects, families can be created with easily measurable means and hence also easily measured variances among family means. It shall generally be assumed that certain kinds of families can be created and that their means and variances can be measured. Then family means and variations can be related to means and variances of gene effects.

The basic statistical method used in defining genetic variances is to define a linear gene-effect model and to partition the total variance due to variations in genotypic mean effects according to the amount which can be accounted for by simple linear effects. More complicated and inclusive models can then be constructed by simply extending the linear models. In a similar way to genetic variables, a variable like soil fertility can be measured and its effects on growth described not only by means and regression coefficients but also by the amount of the variance in tree growth yield which is caused by known or measurable variations in soil fertility. In these more traditional forestry experiments, the importance of controlling fertility is then measurable by the intraclass correlation which gives the variance caused by the measured (for example, soil) variable as a ratio of the total uncontrolled variance. Much quantitative genetics work has a similar objective. When genetic effects cause some average yield differences and the variations due to such differences are identified as the genetic variances, then genetic control can affect forest yields.

As in all experiments, some more or less normal range of conditions is assumed or defined and extraneous causes of variation are controlled as much as possible. More complicated models which can account for additional variances are often introduced in factorial arrangements. These main effects due to linear and polynomial responses are measured on the basis of average or marginal effects over all levels of the other factors. Interactions among main effects can also be defined and measured. In genetic analyses, the effects of different loci are similarly structured where the main effects are the linear (additive) effects of each locus as measured over all other sources of variation, including both environmental and genetic effects of other loci. Dominance effects at each locus are analogous to the quadratic deviations of any factorial analysis. Epistatic effects are analogous to interactions among the main effects, and higher order epistatic effects among several loci are analogous to higher order interactions.

Consider, for example, under some conditions of age, spacing, and general location, that the average volume yield capacity of trees with genotype $A'A$ at the "A" locus is 1,000 units. Certainly, not all trees with this allelic combination will yield the same volume of wood since environmental variables and the genetic condition at other loci can also strongly affect volume

growth. As trees may respond to a preponderance of factors giving more or less than average growth, they will deviate positively or negatively and will therefore cause us to observe actual volumes in some distribution of values around the mean which was composed of different sources of variation. If the same conditions were to exist for trees with genotype *AA*, its growth might average 1,005 units. But if other sources of variation caused variances of 1,000 units in each genotype, the difference between genotypes would be difficult to distinguish even if we could identify the genotypic composition of the trees. If, in addition, genotypes with mean of 995 units also experienced the same variable conditions, the total populational variation over all genotypes would be only slightly greater than 1,000 and gene effects at this locus would have minor importance. If genotypes were randomly distributed over all other sources of variation, the variances due to genetic and other sources would simply be summed in the total population variance and the proportion of the total variance due to genetic variations at the *A* locus would be very small.

On the other hand, if variation from other factors were small, even differences of 5 units would be quite distinctive. In such a case, genetic sources of variation at locus *A* would make up a large portion of the total variance. It is also possible that regardless of the size of the environmentally induced variations, many genetic loci may have variations which also affect average volume production. Thus, a *B* locus with similar effects to the *A* locus in causing ±5 unit deviations could give the following average yields if the effects of *A* and *B* loci were independent and if the trees could be identified:

	AA	*AA′*	*A′A′*
BB	1,010	1,005	1,000
BB′	1,005	1,000	995
B′B′	1,000	995	990

Obviously, more extreme averages exist, around which the same environmental variations may cause dispersal, but the genetic sources of differences have increased the total variation. If the loci are independent in frequency and action, then the variance is simply added to the previous total variance and genetic sources of variation are thereby increased in importance. Several such loci could easily make the genetic variance a large part of the total variance. Hence, even if single-locus effects are small and unimportant, the total array of genotypes can have a major effect on volume yield in the population.

For quantitative genetics and for most breeding work, a useful working hypothesis has been that, unless traits are obviously controlled by very few loci, they are likely to have relatively small individual-locus effects with respect to both the environmental sources of variance and the total genetic variance due to all loci. Thus, even if environmental effects cause large variances with respect to individual-locus genetic effects, the genetic variances

can still contribute heavily to the total variance.

In addition to the relative effect of single genes versus other sources of variance, and the number of such genetic loci, the gene frequency is the third factor affecting the contribution of genetic sources of variance to the total. If one allele is very common, then most trees will be homozygotes with that allele and the rare heterozygotes and the even more rare, alternate homozygotes will have small effect on the total variance regardless of the size of their contribution to average growth differences. This effect of allelic frequency can be more explicitly seen in the exact formulations of genetic variances.

For the simple genetic factors alone, we can define simple models for single gene effects with average effects defined over some understood range of more or less normal environmental and genetic conditions. The three possible genotypes produced by two alleles, A and A' at a locus (AA, AA', and $A'A'$) can have any range of action for any kind of dominance relation between the alleles. For example, average genotypic performances of 1,005:1,000:995 for the three genotypes $AA:AA':A'A'$ display a condition of no dominance, while 1,005:1,005:995 is a classic condition of one allele exhibiting complete dominance. Overdominance is classically defined as the condition in which the AA' has a higher or lower mean than either AA or $A'A'$. (Some authors also use underdominance to indicate the condition of AA' having a lower mean than either homozygote.) It will always be possible to fit a linear regression of yield to a scale of genotypic effects and hence reduce the total variance due to gene actions at this locus by the amount due to this regression.

The difference between the homozygotes may be defined in an arbitrary way to establish a yield scale and define the heterozygote effect in terms of that scale. Though any system would be satisfactory, we adopt the notation of Comstock and Robinson (1948) and use $u:-u$ to define the deviations of $AA:A'A'$ around the mean, and au to define the mean deviation of the heterozygote AA'. Then, to describe the total variance in yield due to genetic differences at this locus, it is necessary to measure the mean differences and weight them according to their frequencies. To determine the portion that can be accounted for by a linear model, let the independent X variable take the values 2, 1, and 0, and let $Y-\mu=u$, au, and $-u$ for genotypes AA, AA', and $A'A'$, respectively. Then, by using the frequencies, f_i, of $q^2:2q(1-q):(1-q)^2$ for the three genotypes, the conventional formula for deriving the total variance (σ_y^2) can be applied:

$$E(Y^2)-[E(Y)]^2=\Sigma f_i Y_i^2-\overline{Y}^2,$$

where $Y_i=Y_{AA}, Y_{AA'}, Y_{A'A'}$,

respectively, and $\overline{Y}=q^2Y_{AA}+2q(1-q)Y_{AA'}+(1-q)^2Y_{A'A'}$,

then the variance due to all gene effects, σ_g^2, is:

$$\sigma_g^2 = 2q(1-q)\,[1+2(1-2q)\,a+(1-2q+2q^2)\,a^2]\,u^2.$$

The variance due to linear regression is the genetic variance of linear effects only (σ_A^2).

$$\sigma_A^2 = \frac{\mathrm{Cov}^2\,(xy)}{\sigma_x^2} = \frac{\Sigma f_i X_i Y_i - \overline{X}\,\overline{Y}}{\Sigma f_i X_i^2 - \overline{X}^2}$$

$$= 2q(1-q)\,[1+(1-2q)\,a]^2\,u^2.$$

If $q=0.5$, then $\sigma_A^2 = 2q(1-q)u^2$.

Since the variance due to nonlinear effects can be defined as the dominance genetic variance σ_D^2, and $\sigma_g^2 = \sigma_A^2 + \sigma_D^2$, then $\sigma_D^2 = 4q^2(1-q)^2\,a^2u^2$.

In a slightly more general way, we can tabulate the mean yields, or average value of the dependent Y variables, and their frequencies (P) and define average effects as:

	A	A'	Average effect	Frequency
A	Y_{AA}	$Y_{AA'}$	Y_A	$P_A = q$
Frequency	P_{AA}	$\frac{1}{2}P_{AA'}$		
A'	$Y_{A'A}$	$Y_{A'A'}$	$Y_{A'}$	$P_{A'} = 1-q$
Frequency	$\frac{1}{2}P_{A'A}$	$P_{A'A'}$		

$$Y_A = \frac{P_{AA}Y_{AA} + (1/2)P_{AA'}Y_{AA'}}{P_A}$$

$$Y_{A'} = \frac{P_{A'A}Y_{A'A} + (1/2)P_{AA'}Y_{AA'}}{P_{A'}}$$

and define $\alpha_A = Y_A - \overline{Y}$ as the average effect of A and $\alpha_{A'} = Y_{A'} - \overline{Y}$ as the average effect of A'.

We can see that $\Sigma P_i \alpha_i = 0$. We can also define a dominance effect as $(Y_{AA} + Y_{A'A'} - 2Y_{AA'})$.

Then, in a linear model of these effects for diploid trees we can write:

$$Y_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij},$$

where δ is the deviation of Y_{ij} from the expected Y_{ij} due only to μ and α effects. The total genetically caused variance is:

$$\sigma_g^2 = 2\sigma_\alpha^2 + \sigma_\delta^2 + 2\mathrm{Cov}\,(\alpha_i\alpha_j) + 2\mathrm{Cov}\,(\alpha_i\delta_{ij}) + 2\mathrm{Cov}\,(\alpha_j\delta_{ij}).$$

Covariances may exist due to nonrandom frequency of the joint occurrence of α or δ effects. These may be nonrandom due to inbreeding or nonrandom mating. In that event, the least squares fit of the model gives us biased estimates for α_i, α_j, and δ_{ij} as previously defined. If inbreeding exists, α_i and α_j are not independently drawn, since by inbreeding we mean there exists a higher fre-

quency of like genotypes mating than expected. Then Cov (α_i,α_j) exists, and using Wright's (1922) definition of an inbreeding coefficient F,

$$F=\text{Corr}(\alpha_i,\alpha_j)=\frac{\text{Cov}(\alpha_i,\alpha_j)}{\sigma_{\alpha_i}\sigma_{\alpha_j}} = \frac{\text{Cov}(\alpha_i,\alpha_j)}{\sigma_\alpha^2}$$

Without dominance effects,

$$\sigma_A^2=2\sigma_\alpha^2+2\text{Cov}(\alpha_i,\alpha_j)$$
$$=2\sigma_\alpha^2(1+F)=\sigma_A^2(1+F).$$

To include the effects of inbreeding and dominance, we can derive a more general solution by considering the complete model and deriving least squares estimators for all effects. Such a procedure would yield definitions:

$$\sigma_A^2=\frac{2q(1-q)}{1-F}[(q+F(1-q))(Y_{AA}-Y_{AA'})$$
$$+(1-q+Fq)(Y_{AA'}-Y_{A'A'})]^2$$

and $$\sigma_D^2=\frac{q(1-q)}{1-F}[(q+F(1-q))((1-q)+Fq)(1-F)]$$
$$(Y_{AA}+Y_{A'A'}-2Y_{AA'})^2.$$

It can be seen in these formulas that inbreeding changes frequencies of genotypes from their random mating frequencies and therefore affects the definition of σ_A^2 and σ_D^2 such that they may not be translated easily if we wish to estimate or define σ_A^2 or σ_D^2 for populations at a different level of inbreeding.

The extension to multiple allelic cases is direct, involving only the estimation of more interaction on epistatic parameters for interlocus effects and accumulating more main effects for the linear-additive components.

For an expanded model including two loci, it is necessary to consider all the additive and dominance effects at each locus and, in addition, to consider any interactions between the loci in terms of their respective additive and dominance effects. If we assume the simplified conditions of no inbreeding and random mating for each locus, and further assume that the various zygotic states are independent gametic associations, then the joint frequencies are simply products of the frequencies at each locus, and the variances can be expressed as a sum of the variances at each locus plus the variances due to the epistatic interactions. Again, least squares estimators and variances can be directly derived for any combination of frequencies and effects, as outlined by Cockerham (1954) and Kempthorne (1957).

A linear model for two loci, say the A and B as in the previous example, can be constructed for any kind of dominance effects at each locus and for any epistatic changes in the average effects at one locus due to changes in the other locus. Without epistasis

and with gene effects at the A locus causing average yields of 1,005:1,000:995 for $Y_{AA}:Y_{AA'}:Y_{A'A'}$, the linear model would have $u=Y_{AA}-Y_{A'A}=5$, $au=Y_{AA'}-\frac{Y_{AA}+Y_{A'A'}}{2}=0$, and $a=0$. If $q=\frac{1}{2}$, the frequencies of $AA:AA':A'A'$ would be $\frac{1}{4}:\frac{1}{2}:\frac{1}{4}$, and the additive genetic variance would be 12.5. If we were to raise or lower the average yields by any constant amount, say 5 units, corresponding to changing the state of the B locus from BB' to BB, or $B'B'$, then we would have a similar state of additivity at the B locus as at the A locus and the average Y values for complete additivity would be:

	AA	AA'	$A'A'$
BB	1,010	1,005	1,000
BB'	1,005	1,000	995
$B'B'$	1,000	995	990

If $q_B=0.5$, then σ_A^2 for the B locus is 12.5, and if the loci freely recombine, the total genetic variance is 25, the sum of the additive variances at each locus. If dominance existed at one locus, we might have an average A locus yield of 1,003.5:1,001.5:993.5, so that its $u=5$, $a=0.6$, and at $q_A=0.5$, giving $\sigma_A^2=12.5$, and $\sigma_D^2=2.25$. If the B locus were to remain as it was, with partial dominance at the A locus, additivity of the B locus, no epistasis, and $q_A=q_B=0.5$, the table of Y values would be:

	AA	AA'	$A'A'$	Mean
BB	1,008.5	1,006.5	998.5	1,005
BB'	1,003.5	1,001.5	993.5	1,000
$B'B'$	998.5	996.5	988.5	995
Mean	1,003.5	1,001.5	993.5	

There is still independence between loci in gene action, and the total genetic variance is still the sum of the variances at each locus, $\sigma_A^2=25$, $\sigma_D^2=2.25$.

If interactions exist such that genotypic differences at one locus are not constant, then epistasis exists. For example, when the additivity at the B locus varies from its constant $u=5$, $a=0$ to some other values, say $u=7.5$ when the A locus is AA, but then $u=10$ when the A locus is AA', and $u=2.5$ with $A'A'$, the average gene effects remain the same but an additional variance is caused by the interaction between loci. This condition is called interlocus epistasis or gene interaction. The average yields for this epistatic model, with $q_A=q_B=\frac{1}{2}$, would be:

	AA	AA'	$A'A'$	Mean
BB	1,011.0	1,006.5	996.0	1,005
BB'	1,003.5	1,001.5	993.5	1,000
$B'B'$	996.0	996.5	991	995
Mean	1,003.5	1,001.5	993.5	1,000

The total genetic variance is now larger than before while the genetic variance at each locus, based on average effects of each locus over all other genetic and nongenetic sources of variance, remains the same. Thus, even if the loci freely recombine and the locus means, frequencies, and variances are the same as before, the total genetic variance of 28.625 includes an additional 1.375 due to epistatic effects. Clearly, an infinite variety of epistatic models can be written with dominance levels changing at both loci and varying additivity levels. If the coupling and repulsive heterozygotes are similar, the nine zygotic states can be listed as:

	Locus A			Mean of B	Frequency
	AA	Aa	aa		
Locus B					
BB	Y_{AABB}	Y_{AaBB}	Y_{aaBB}	$Y_{\cdot BB}$	p^2
Bb	Y_{AABb}	Y_{AaBb}	Y_{aaBb}	$Y_{\cdot Bb}$	$2p(1-p)$
bb	YY_{AAbb}	Y_{Aabb}	Y_{aabb}	$Y_{\cdot bb}$	$(1-p)^2$
Mean of A	$Y_{AA\cdot}$	$Y_{Aa\cdot}$	$Y_{aa\cdot}$		
Frequency	q^2	$2q(1-q)$	$(1-q)^2$.		

The differences among these genotypic states can be described in terms of a factorial arrangement of genetic effects; the effects due to locus A (α) should be averaged over all levels of locus B, the effects of B (β) should be averaged over A, and the interactions between A and B. The effects of each locus are described as before in terms of linear additive effects and dominance deviations. The interactions could be described as interactions among linear effects at A by linear at B, linear at A by dominance at B, dominance at A by linear at B, and dominance at A by dominance at B. Thus, we can write the model of gene effects as:

$$\begin{aligned} Y_{ijkl} = \mu &+ \alpha_i + \alpha_j + \delta_{ij} + \beta_k + \beta_l + \gamma_{kl} \\ &+ (\alpha\beta)_{ik} + (\alpha\beta)_{il} + (\alpha\beta)_{jk} + (\alpha\beta)_{jl} \\ &+ (\alpha\gamma)_{ikl} + (\alpha\gamma)_{jkl} + (\beta\delta)_{ijk} + (\beta\delta)_{ijl} \\ &+ (\delta\gamma)_{ijkl}. \end{aligned}$$

The variances due to all of these gene effects can then be summarized as:

$$\begin{aligned} \sigma_{\underline{Y}}^2 = \sigma_A^2(_A \text{ locus}) + \sigma_A^2(_B \text{ locus}) + \sigma_D^2(_A \text{ locus}) + \sigma_D^2(_B \text{ locus}) \\ + \sigma^2_{A\cdot A} + \sigma^2_{A\ D} + \sigma^2_{D\cdot D}. \end{aligned}$$

The additive and dominance variances at the two loci are simply added together and their sum is the additive and dominance variances for the trait.

The only new variances are the three epistatic interaction components. These can be derived in exactly the same fashion as for the general case. Using the notation:

$$e_{22} = (Y_{AABB} - Y_{AABb}) - (Y_{AaBB} - Y_{AaBb})$$
$$e_{21} = (Y_{AABb} - Y_{AAbb}) - (Y_{AaBb} - Y_{Aabb})$$
$$e_{12} = (Y_{AaBb} - Y_{AaBb}) - (Y_{aaBB} - Y_{aaBb})$$
$$e_{11} = (Y_{AaBb} - Y_{Aabb}) - (Y_{aaBb} - Y_{aabb}),$$

then:

$$\sigma_{A^1}{}^2 = \frac{4q(1-q)p(1-p)}{(1+F)^2}\Big[[p+F(1-p)]\ [q+F(1-q)]e_{22}$$
$$+[p+F(1-p)]\ [1-q+Fq]e_{21}$$
$$+[1-p+Fp]\ [q+F(1-q)]e_{12}$$
$$+[1-p+Fp]\ [1-q+Fq]e_{11}\Big]^2.$$

$$\sigma_{AD}{}^2 = \frac{2p(1-p)q(1-q)(1-F)}{(1+F)^2}[q+F(1-q)]\ [1-q+Fq]$$
$$\Big[[p+F(1-p)]\ [e_{22}-e_{21}]+[1-p+Fp]\ [e_{21}-e_{11}]\Big]^2$$
$$+\frac{2q(1-q)p(1-p)\ (1-F)}{(1+F)^2}[p+F(1-p)]\ [1-p+Fp]$$
$$\Big[[q+F(1-q)]\ [e_{22}-e_{21}]+[1-q+Fq]\ [e_{21}-e_{11}]\Big]^2$$

$$\sigma_{DD}{}^2 = \frac{p(1-p)q(1-q)\ (1-F)^2}{(1+F)^2}[p+F(1-p)]\ [1-p+Fp]$$
$$[q+F(1-q)]\ [1-q+Fq]\Big[e_{22}-e_{21}-e_{12}-e_{11}\Big]^2.$$

EXTENSION*

A general algebraic expression for the two-locus model can be written in terms of more traditional regression effects as the linear (additive) effects at each locus, the quadratic (dominance) effects at each locus, the linear-by-linear interaction (additive-by-additive epistasis), the linear-by-quadratic interaction (additive-by-dominance epistasis), and the quadratic-by-quadratic interaction (dominance-by-dominance epistasis). The total variance would now include not only the $\sigma_\alpha{}^2$, $\sigma_\delta{}^2$, $\mathrm{Cov}(\alpha_i,\alpha_j)$ effects at the A locus and the $\sigma_\beta{}^2$, $\sigma_\delta{}^2$, $\mathrm{Cov}(\beta_i,\beta_j)$, at the β locus, but the $\sigma_{\alpha\beta}{}^2$, $\sigma_{\alpha\delta_\beta}{}^2$, $\sigma_{\beta\delta_A}{}^2$, and $\sigma_{\delta_A\delta_B}{}^2$, which are, respectively, the additive-by-additive, the additive-by-dominance at both loci, and the dominance-by-dominance variances. Also, the covariances between any of the elements due to linkage disequilibrium between the A and B loci must be included in a general model.

*Graduate-level statistical training required for thorough understanding.

Expansion to three loci increases the model elements, not only by the simple average effects at the new locus, but also by the new two- and three-way interactions or epistatic effects. The genetic variances include an additive variance, σ_A^2, at each locus; a dominance variance, σ_D^2, at each locus; an additive-by-additive variance, σ_{AA}^2, at each pair (A,B; A,C; B,C) of loci; and additive-by-dominance variance, σ_{AD}^2, at each pair of loci; a dominance-by-dominance variance, σ_{DD}^2, at each pair of loci; an additive-by-additive-by-additive variance, σ_{AAA}^2; three additive-by-additive-by-dominance variances, σ_{AAD}^2, σ_{ADA}^2, and σ_{DAA}^2; three additive-by-dominance-by-dominance variances, σ_{ADD}^2, σ_{DAD}^2, and σ_{DDA}^2; and finally, a dominance-by-dominance-by-dominance variance, σ_{DDD}^2. The total genetic variances are then increased by these new elements which contribute average performance variations and are also changed by three-way linkage disequilibrium effects. Thus, complete, multilocus systems can be built up. Their complexity expands rapidly, but they completely account for sources of variance suggested by basic linear statistical concepts.

Alternatives to the linear statistical models and analysis of variance types of estimators exist. One alternative system of defining genetic variances was suggested by Kenneth Mather and extended by Dickinson and Jinks (1956) and further by Hayman (1958, 1960b). Mainly applicable to homozygous lines and crosses among them, a basic genetic model of d, h, and $-d$ for AA, AA', and $A'A'$, with gene frequency u for A and v for A', is used to derive 6 or 8 variances and covariances. These can be used to test hypotheses about the sizes of the additive and dominance effects, the gene frequencies of favorable alleles, numbers of loci, and the presence of certain kinds of epistasis. The method is therefore very comprehensive for estimating gene actions in a sampled population, but is not likely to be of much use in forestry.

At its simplest, the method consists of estimating variances among parental means (D), variance among families made with a common parental line, say r, (V_r), the average of those variances ($\overline{V}_r$), variance among the various families' means ($V_{\bar{r}}$), the covariances of parents with offspring families within parental line r ($W_{\bar{r}}$), and the average of these covariances ($\overline{W}_r$). Each of these statistics has an expected value in terms of D, H_i, and F_i, and h functions which in turn are functions of the gene frequency and gene-action parameters. While Kearsey (1965) has shown that translations can be made between these statistics and the genetic variances as described by Cockerham (1959) and Kempthorne (1957), it is possible to make direct estimates of additive and dominance effects without the confounding of dominance effects in the additive genetic variances as involved in those previously described statistics. In addition, graphical interpretation of the analyses is particularly illuminating. The critical problem is the extension of such analyses to the general case of heterozygosity of parents, as partially developed by Dickinson

and Jinks (1956) and extended by Oakes (1967), and to the sampling errors associated with estimates and tests of similar hypotheses. For some special cases, Kearsey (1965) has examined the utility of these and the analysis of variance methods and concluded that for the same number of families raised, the Jinks and Hayman types of diallel analyses gave most information but severely restricted the population of parents which could be sampled.

Thus, the genetic variances can be partitioned into various measures of how genes affect phenotypes in some sampled populations as functions of their gene frequencies and genotypic frequencies. The assumptions required for analysis and estimation of those statistics, however, are often very restrictive. The problems of estimation will be investigated in the next chapter. It should be clearly noted that the parameters we speak of and the methods used to estimate them are not readily separable. Estimation methods are often dictated by the model parameterizations. It should also be emphasized that if inbreeding is an important factor in populations, a correlation will exist in the frequency with which the alleles at a locus will associate and that correlation among alleles at different loci will also occur. The average effects of alleles will therefore change if inbreeding levels change, causing the genetic variances to change. In such cases, a trajectory of genetic variances may be a more interesting statistic to estimate. In addition, the presence of linkage disequilibria and any disequilibria caused by sampling or by crossing previously isolated chromosomes generates correlations among loci, also making the genetic variance nonstationary. It is often impossible, therefore, to describe and estimate simple parameters relating to general genetic phenomena, but these first approximations have served well.

GENETIC VARIANCES IN TREE SPECIES

If we accept a certain vagueness about the exact meaning of our average statistics like σ_A^2, we still cannot escape the strength of the general conclusion that in almost any trait studied in almost any tree species studied, considerable variation is due to genetic sources. While there are some notable exceptions to this experience and while the record is somewhat biased because geneticists generally test traits they suspect of having some genetic variance, the results are too broad to dismiss. The current work on the extent of genetic variation in allelic polymorphisms indicates large amounts of residual genetic variance in presumably unselected or weakly directionally selected traits in many populations of plants and animals. The same may be true in tree populations with respect to the traits studied. Over the range of tree genera and species studied, for a variety of traits exhibited at different times of the life cycle, genetic sources of variance have generally been found whenever the variation present has been investigated

by appropriate analyses.

Genetic variance estimates have been derived for a broad array of tree species, but there is a heavy preponderance of commercial species for which intensive silviculture has led to breeding interest. The pines have been most intensively studied. Estimation experiments have been done with loblolly (Stonecypher 1966), slash (Barber 1964), eastern white (Kriebel and others 1972; Wright 1970), western white (Hanover and Barnes 1969), ponderosa (Callaham and Hasel 1961), Monterey (Nicholls and others 1964), jack (King and Nienstaedt 1965), Scots (Wright 1963; Ehrenberg 1966), red (Fowler and Lester 1970), and *patula* (Armitage and Burrows 1966) pines. The studies noted, parenthetically, are by no means all that have been done on even these pine species, and represent only a fraction of all studies which have indicated the existence of genetic variance in forest tree populations.

Other conifers which have been studied to estimate genetic variances include Douglas-fir (Campbell 1964), Norway spruce (Saeterstal 1963; Lacaze and Arbez 1971), and *Cryptomeria* (Toda 1961). Among the hardwoods, the various species and hybrids of *Populus* have received widest attention (Hattemer 1976; Wilcox and Farmer 1967). In addition, some estimates of genetic variances have been published for *Acer saccharum* (Kriebel and Gabriel 1969), *Juglans nigra* (Funk 1970), *Liriodendron tulipifera* (Kellison 1970), *Quercus rubra* (Kriebel 1965), *Platanus occidentalis* (Webb 1970), *Liquidambar styraciflua* (Wilcox 1970), *Betula verrucosa* (Tigerstedt 1966; Stern 1962), *Eucalyptus regnans* (Eldridge 1966), wattle (Moffett and Nixon 1963), and *Gleditsia triacanthos* (Grisjuk 1959).

Among these species, many traits have been studied—again largely those associated with commercially important features and mostly restricted to traits as expressed in young trees. In addition to the commonly measured growth and survival traits, wood quality has been widely and intensively studied (Smith 1967) by Zobel (1961) in conifers and by Bhagwat (1963) in poplars. Estimates have also been derived for root growth (Wilcox and Farmer 1968), stem form (Ehrenberg 1961), crown form (Barber 1961), branching characteristics (Strickland and Goddard 1966), leaf form (Kellison 1970), thorn morphology (Grisjuk 1959), seed morphology (Kraus 1967), and fruitfulness (Varnell and others 1967). In addition, genetic variances have been estimated for competitive ability (Sakai and others 1968); resistance to cold (Rudolph and Nienstaedt 1962; Dietrichson 1961), drought (Texas Forest Service 1957), insects (Wright and others 1967), diseases (Bingham and others 1969; King and Nienstaedt 1965), and transplant shock (Beineke 1967); rooting ability (Muzik and Cruzado 1958); grafting ability (Hanover 1962); and the yield of gum exudates (Squillace 1966a) and rubber (Burkill 1959). Some physiological traits have also been studied and genetic vari-

ances have been estimated for nutrient absorption (Walker and Hatcher 1965) and photosynthetic and respiratory rates (Ledig and Perry 1967).

It is also clear that measurements of a trait in different stages of the life cycle may represent somewhat different phenomena. Crown and branch characters change very rapidly in the early years (Snyder 1961), as do wood fiber characteristics (Zobel and others 1961). While it is reasonable to expect that traits which develop in sequence are closely correlated, the developmental mechanisms cannot generally be expected to remain under constant control, and hence something less than perfect correlation is to be expected. In particular, if a trait has different selective pressures with respect to survival at different ages, then we might expect the various kinds and levels of genetic variances to change somewhat over the life cycle. Thus, in height growth of Douglas-fir, the genetic variances among families within populations were found to decline over a 40-year period (Namkoong and others 1972). It was also found that the error variance tended to decrease when the trees were 15 to 20 years old, suggesting that height-growth control mechanisms do change as trees mature. Similar patterns were found for ponderosa pine up to 29 years of age. Thus, while some studies indicate little change in genetic variances through the juvenile period, more advanced ages may indicate quite different apportionments of the genetic and error sources of variance. It can, therefore, also be expected that the correlations of traits at the same and at different ages will be different. They may be expected to be large if the causal mechanisms are similar and small if the causal mechanisms are largely independent.

COVARIANCES OF RELATIVES

If the genetic variance in a population is defined as the variance among individuals caused by gene effects, then there would necessarily be no genetic variance among individuals which are genetically identical. Conversely, the genetic variance would make its full contribution to total variance if individuals were randomly chosen. Between these extremes, the amount of genetic variance exhibited depends upon the relatedness of individuals—the closeness of their parentage. In similar environments, close relatives are generally less variable among themselves than are nonrelatives because their genes were derived from a restricted population. Therefore, a correlation in their gene effects must exist. In this section, the correlation among relatives is defined in terms of genetic variances. In the next chapter, the correlation among relatives is defined in terms of estimated family variances. Therefore, relationship of genetic variances to estimated family variance components is derived.

For any two individuals, a genetic covariance would exist and can be written in terms of their genetic effects if there is some

probability ($\neq 0$) that their genetic effects are more likely to be identical than what would occur solely by chance in random mating. If pairs of individuals are randomly chosen from the whole population, then their alleles are expected to occur in the frequencies expected of that general population. If the pairs have a close relationship, then the nonrandomness can be measured by the frequency or probability that the alleles in the two individuals are identical in descent and exactly alike. Thus, for a linear model of average and dominance effects, as previously defined, we can derive the covariance between two individuals, X and Y, according to the probabilities that their alleles are the same:

Let $X=\mu+\alpha_{X_\sigma}+\alpha_{X_\varphi}+\delta_{X_\sigma X_\varphi}$

and $Y=\mu+\alpha_{Y_\sigma}+\alpha_{Y_\varphi}+\delta_{Y_\sigma Y_\varphi}$

where α_{X_σ} = average effect of allele from male parent of X,

α_{X_φ} = average effect of allele from female parent of X,

$\delta_{X_\sigma X_\varphi}$ = dominance deviation of allelic combination in X,

α_{Y_σ} = average effect of allele from male parent of Y,

α_{Y_φ} = average effect of allele from female parent of Y,

$\delta_{Y_\sigma Y_\varphi}$ = dominance deviation of allelic combination in Y.

With respect to the various genetic effects, the covariance of X and Y equals $[E(XY)-E(X)E(Y)]$ which contains:

$$E(\alpha_{X_\sigma}\cdot\alpha_{Y_\sigma})+E(\alpha_{X_\sigma}\cdot\alpha_{Y_\varphi})+E(\alpha_{X_\varphi}\cdot\alpha_{Y_\sigma})$$
$$+E(\alpha_{X_\varphi}\cdot\alpha_{Y_\varphi})+E(\delta_{X_\sigma X_\varphi}\cdot\alpha_{Y_\sigma})+E(\delta_{X_\sigma X_\varphi}\cdot\alpha_{Y_\varphi})$$
$$+E(\delta_{Y_\sigma Y_\varphi}\cdot\alpha_{X_\sigma})+E(\delta_{Y_\sigma Y_\varphi}\cdot\alpha_{X_\varphi})+E(\delta_{X_\sigma X_\varphi}\cdot\delta_{Y_\sigma Y_\varphi}).$$

The first four elements are additive variances and covariances, the second four are covariances of additive and dominance effects (not epistatic interactions), and the last element is the dominance variance or covariance. If the male parentage of X and Y is not random, then a certain probability exists that $\alpha_{X_\sigma}=\alpha_{Y_\sigma}$, and then $E(\alpha_{X_\sigma}\alpha_{Y_\sigma})=Pr(X_\sigma=Y_\sigma)\cdot E(\alpha_i^2)$. It was previously derived that $E(\alpha_i^2)=\sigma\alpha^2$, and in particular, was $(\frac{1}{2})\sigma_A^2$. Therefore,

$$E(\alpha_{X_\sigma}\alpha_{Y_\sigma})=Pr(X_\sigma=Y_\sigma)\sigma\alpha^2=Pr(X_\sigma=Y_\sigma)(\tfrac{1}{2})\sigma_A^2.$$

Similarly,
$$E(\alpha_{X_\sigma}\cdot\alpha_{Y_\varphi})=Pr(X_\sigma=Y_\varphi)\cdot\tfrac{1}{2}\sigma_A^2,$$
$$E(\alpha_{X_\varphi}\cdot\alpha_{Y_\sigma})=Pr(X_\varphi=Y_\sigma)\cdot\tfrac{1}{2}\sigma_A^2,$$
$$E(\alpha_{X_\varphi}\cdot\alpha_{Y_\varphi})=Pr(X_\varphi=Y_\varphi)\cdot\tfrac{1}{2}\sigma_A^2.$$

Also, $E(\delta_{X_{\sigma^{\prime}}Y_{\varphi}} \cdot \alpha_{Y_{\sigma}}) = Pr(X_{\sigma^{\prime}}=Y_{\sigma}, X_{\varphi}=Y_{\sigma^{\prime}}) \cdot \mathrm{Cov}(\alpha_Y \delta_X)$,

$E(\delta_{X_{\sigma^{\prime}}X_{\varphi}} \cdot \alpha_{Y_{\varphi}}) = Pr(X_{\sigma^{\prime}}=Y_{\varphi}, X_{\varphi}=Y_{\varphi}) \cdot \mathrm{Cov}(\alpha_Y \delta_X)$,

$E(\delta_{Y_{\sigma^{\prime}}Y_{\varphi}} \cdot \alpha_{X_{\sigma^{\prime}}}) = Pr(Y_{\sigma^{\prime}}=X_{\sigma^{\prime}}, Y_{\varphi}=X_{\sigma^{\prime}}) \cdot \mathrm{Cov}(\alpha_X \delta_Y)$,

$E(\delta_{Y_{\sigma^{\prime}}Y_{\varphi}} \cdot \alpha_{X_{\varphi}}) = Pr(Y_{\sigma^{\prime}}=X_{\varphi}, Y_{\varphi}=X_{\varphi}) \cdot \mathrm{Cov}(\alpha_X \delta_Y)$.

Finally, $E(\delta_{X_{\sigma^{\prime}}X_{\varphi}} \cdot \delta_{Y_{\sigma^{\prime}}Y_{\varphi}}) = Pr(X_{\sigma^{\prime}}=Y_{\sigma^{\prime}}, X_{\varphi}=Y_{\varphi}) \cdot \sigma_\delta^2$

or $Pr(X_{\varphi}=Y_{\sigma^{\prime}}, X_{\sigma^{\prime}}=Y_{\varphi}) \cdot \sigma_\delta^2$.

If the female parentage was somehow nonrandom, then $Pr(X_{\varphi} \cdot Y_{\varphi}) \neq 0$ and similarly for the others of the first four expectations. Summing yields $\sum_{i,j} Pr(X_i=Y_j) \cdot \frac{1}{2}\sigma_A^2$.

If X has a ♂ and ♀ parental relationship which itself is related to the male parentage of Y, then $E(\delta_{X_{\sigma^{\prime}}X_{\varphi}} \cdot \alpha_Y)$ is not zero and would have to be computed, but these kinds of relationships can temporarily be ignored if only nonrelatives are crossed. Therefore, the second set of four elements is assumed to be zero. If both ♂ and ♀ parentage of X and Y are identical or related then:

$$E(\delta_{X_{\sigma^{\prime}}X_{\varphi}} \cdot \delta_{Y_{\sigma^{\prime}}Y_{\varphi}}) = Pr(X_{\sigma^{\prime}}=Y_{\sigma^{\prime}}, X_{\varphi}=Y_{\varphi})\,\sigma_\delta^2 + Pr(X_{\varphi}=Y_{\sigma^{\prime}}, X_{\sigma^{\prime}}=Y_{\varphi})\,\sigma_\delta^2,$$

and these contribute to the existence of the last element.

Then for some common kinds of relationship, we can trace the various probabilities and determine the contributions of these genetic variances in terms of the first four and the last elements of the covariance of relatives. For example, if the female parent of X and Y was the same, then the only nonzero probability would be $Pr(X_{\varphi}=Y_{\varphi})$; its size would depend on how the choice of gametes is made in the production of eggs from the common mother. If the choice is random, then the probability is ½ that the same allele (either one) is chosen, and the only contribution of the genetic variance to the covariance of these half-sibs is $\frac{1}{4}\sigma_A^2$.

If both the ♂ and ♀ parents of X and Y were common, then $Pr(X_{\varphi}=Y_{\varphi}) = Pr(X_{\sigma^{\prime}}=Y_{\sigma^{\prime}}) = \frac{1}{2}$, and the probability that both are identical, $Pr(X_{\varphi}=Y_{\varphi}, X_{\sigma^{\prime}}=Y_{\sigma^{\prime}})$, is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, and the other probabilities are zero. Therefore, the genetic variance contribution to the covariance of full-sibs is $\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$.

For the case of parent-offspring covariances, if the maternal parent is the X individual and the offspring is Y, then $Pr(X_{\sigma^{\prime}}=Y_{\varphi}) = Pr(X_{\varphi}=Y_{\varphi}) = \frac{1}{2}$, and all other probabilities are zero. Then, the covariance of parent and offspring is $\frac{1}{2}\sigma_A^2$. The probability that a random allele from X is identical by descent to a random allele from Y is:

$$\sum_{i,j} \frac{Pr(X_i=Y_j)}{4},$$

which is Malécot's (1969) coefficient of relationship C_{XY}. Therefore, $2C_{XY}=\frac{1}{2}\sum_{i,j} Pr X_i=Y_j)$ can be used as the coefficient for the σ_A^2 contribution to the covariance of relatives.

If additional genetic loci affect the genetic variances and covariances among any relatives and if they are independent loci, then the probabilities of identity by descent for multiple loci are summed over the genetic variances at each locus. For the multiple-locus epistatic effects, the probabilities of joint identities by descent are products of the independent probabilities. In such cases, for any kinds of relatives which have the additive genetic variance coefficient of a, and the coefficient for σ_D^2 of d, the general covariance due to all types of genetic variances can be written as:

$$\text{Cov}=a\sigma_A^2+d\sigma_D^2+ad\sigma_{AD}^2+a^2\sigma_{AA}^2+d^2\sigma_{DD}^2+a^2d\sigma_{AAD}^2+\ldots$$

or in general, $\text{Cov}=\sum_{i\,j} a^i d^j \sigma^2_{A^i D^j}$.

As previously noted with respect to the definition of the genetic variances, inbreeding nullifies the independence assumptions and the derivations of the probabilities of drawing identical alleles. It is clear, for example, that if F is defined as the probability that the two alleles at a locus are identical by descent, the probability that two alleles in two gametes randomly drawn from an individual tree are identical is $\frac{1+F}{2}$ instead of $\frac{1}{2}$. Then, with a parental inbreeding coefficient of F, even with random choice of parents, and hence no inbreeding of the offspring, the a and d coefficients used to compute the covariances of relatives are increased by factors of $1+F$ and $(1+F)^2$, respectively. The problem remains, however, that the σ_A^2 and σ_D^2 themselves require specification with respect to the inbreeding generations they refer to.

The effects of linkage also can clearly affect the probabilities of some gametic combinations and hence the contributions of the epistatic gene effects and their summations in the additive variance. The manner in which they affect the covariance of relatives is not a simple derivable relationship (Cockerham 1956). Nonetheless, if we wish to define and estimate meaningful parameters, the broad effects of such factors as linkage and inbreeding must be considered.

It is also clear that hybrid populations will engender genetic variances and covariances among relatives with quite unique effects and probabilities of drawing various gametic contributions. The effects of dominance types of intralocus gene actions are unique, and all types of interlocus epistatic interactions would not only be unique but their frequencies would depend on the differences in gene frequencies among the populations and on the linkages disequilibria so induced (Stuber and Cockerham 1966). For our brief review purposes, all of these effects will be assumed

absent as we shall assume large, random-mating populations with independent loci.

If it is a reasonable approximation to assume independence, it is clear that we can create families of varying degrees of relationship and, hence, get a handle on the amounts of genetic variations that exist within the populations from which the samples are drawn. By drawing a sample of genotypes which presumably represents the various genotypic effects and their frequencies of occurrence in the population, we can only measure a total variance unless we can artificially construct known families to see how the genetic sources of variance affect the size of the covariance among relatives. The larger the genetic variance contributions to the total variance are, the larger will be the differences among family units, and the closer the relationship among family members, the larger also will be the differences among families. By making sets of different kinds of relatives, we can then partition the existing genetic variation according to the contribution of the genetic variance to the covariance of those relatives, and derive estimators for the genetic variance. We shall investigate the variety of mating forms which we can use when we consider genetic experimental designs and analyses, but it is instructive to derive one case in which a simple experiment provides an estimate of the additive genetic variance.

Consider that a random sample of females is drawn from the population and that each is fertilized by a large number of randomly chosen pollen grains from the general population. In such a case, we have female half-sib families, and the covariance among seedlings within their families is that of half-sibs. If it is also considered that the variance among these families (σ_f^2) is due to some females being AA; others AA'; and others $A'A'$, then σ_f^2 is some function of the genetic variance also. In fact, if the females are random samples from the population and their effects are defined as deviations from the general mean, then from a linear model of yield for two individuals,

$$X_{ij} = \mu + f_i + e_{ij}$$

$$Y_{kl} = \mu + f_k + e_{kl}$$

where $i,k = 1, 2 \ldots p$

$j,l = 1, 2 \ldots n,$

we can derive that $E(\bar{f}) = 0$ and $E(f_i^2) = \sigma_f^2$. It can also be observed that for two individuals, $E(f_i \cdot f_k)$ is zero if $i \neq k$ (by the randomness assumption) or $E(f_i \cdot f_k)$ is σ_f^2 if $i = k$ (if both have the same mother). Thus, the variance among female groups equals the covariance of individuals within groups and, in our case, would be expected to be $\frac{1}{4}\sigma_A^2$, plus any epistatic effects appropriate to half-sib relations. We can derive these effects also by considering the genetic variances we might expect from a population of half-sibs.

Consider the mating frequencies in table 5 in which females with genotypes AA are expected to occur with frequency q^2 in our population and in our random sample, and AA has phenotype a, AA' has phenotype au, and $A'A'$ has phenotype $-a$. The matings of AA females with AA males (which also exist in frequency q^2) occur at a frequency of $q^2 \times q^2$ if male and female were randomly and independently chosen. Since all offspring are AA, with average yield u, these progenies would yield q^4u to the family mean of AA females. For individuals of crosses of AA females with AA' males, the frequency is expected to be $q^2 \cdot 2q(1-q)$ with half the progeny being AA and half being AA'. Then, the contribution of these individuals to the AA female family mean would be $2q^3(1-q)\frac{1}{2}(u+au)$. For individuals of crosses of AA females with $A'A'$ males, the frequency is expected to be $q^2 \times (1-q)^2$ and all individuals would have an average phenotype of au and hence they would contribute $q^2(1-q)^2au$ to the AA female family mean. Within the AA maternal family, all genotypes contribute a collective frequency of q^2 and to a phenotypic mean of $q^2u+2q(1-q)[\frac{1}{2}(u+au)]+(1-q)^2au=qu+(1-q)au$, as shown in table 5. Similarly, the frequency and the expected means can be derived for AA' female families as frequency$=2q(1-q)$ and mean$=(\frac{1}{2})(2q-1)u+\frac{1}{2}au$, and for $A'A'$ female families as frequency$=(1-q)^2$, and mean$=-(1-q)u+q(au)$, and for all individuals the mean$=(2q-1)u+2q(1-q)au$. Then computing the variance among the family means as Σ family frequency $\times$ (family mean)^2 $-$ (grand mean)^2 yields:

$$q^2[qu+(1-q)au]^2+2q(1-q)(\tfrac{1}{4})[(2q-1)u+au]^2$$
$$+(1-q)^2[-(1-q)u+qau]^2-[(2q-1)u+2q(1-q)au]^2$$
$$=\frac{q(1-q)}{2}[1+(1-2q)a]^2u^2.$$

This value is exactly $\frac{1}{4}$ of the σ_A^2 we previously derived as the variance among average effects of alleles. Thus, only $\frac{1}{4}$ of the σ_A^2 is contributed to the covariance of half-sibs.

We might also notice that if mating was not at random, then the mating frequencies are not correctly computed and perhaps the parental genotypic frequencies are other than the expected q^2; $2q(1-q)$; and $(1-q)^2$. In such cases, as we have already remarked, the genetic variance itself is not simply defined, but changes with frequencies of genes, average effects of alleles, and the variance of average effects. These possibilities have been of some concern in experiments using open-pollinated tree seeds because effective pollination as well as ancestral relationship may be highly dependent on distance. While these conjectures seem reasonable (Wright 1962; Langner 1953; Sakai 1971) there is not enough evidence to indicate where or with what species this problem is serious. A further source of bias in open-pollination tests is the possibility that limited numbers of males may effectively

Table 5.—Mating frequency table

Matings		Mating frequency	Values of offspring			Family mean
Female genotype	Male genotype		AA	AA'	$A'A'$	
AA	AA	q^4	u	0	0	u
	AA'	$2q^3(1-q)$	$\frac{1}{2}(u)$	$\frac{1}{2}(au)$	0	$\frac{1}{2}(u+au)$
	$A'A'$	$q^2(1-q)^2$	0	au	0	au
AA female mean		q^2				$qu+(1-q)au$
AA'	AA	$2q^3(1-q)$	$\frac{1}{2}(u)$	$\frac{1}{2}(au)$	0	$\frac{1}{2}(u+au)$
	AA'	$4q^2(1-q)^2$	$\frac{1}{4}(u)$	$\frac{1}{2}(au)$	$\frac{1}{4}(-u)$	$\frac{1}{2}(au)$
	$A'A'$	$2q(1-q)^3$	0	$\frac{1}{2}(au)$	$\frac{1}{2}(-u)$	$\frac{1}{2}(-u+au)$
AA' female mean		$2q(1-q)$				$\frac{1}{2}(2q-1)u+\frac{1}{2}au$
$A'A'$	AA	$q^2(1-q)^2$	0	au	0	au
	AA'	$2q(1-q)^3$	0	$\frac{1}{2}(au)$	$\frac{1}{2}(-u)$	$\frac{1}{2}(-u+au)$
	$A'A'$	$(1-q)^4$	0	0	$-u$	$-u$
$A'A'$ female mean		$(1-q)^2$				$-(1-q)u+q(au)$
Grand means for all progeny						$(2q-1)u+2q(1-q)au$

pollinate any particular female in a given year. If this occurs, the variance among families which is more distinctive due to differences among male parents will increase (Namkoong 1965). We previously assumed a pollen mix from the general stand equally effective for all females, but we may realistically wish to consider the effects of such possible factors as few or single males producing families with close average relatedness, the males being related, and possibly even the female being related to whatever males may be effective. The overriding need seems to be for data to estimate the size of any possible biases. Until such data are available, the wisest course would seem to be to avoid relatedness and increase effectiveness of broad population egg and pollen samples by sampling different years or providing supplemental pollen dispersals by artificial means when possible and to proceed with such estimates as may be minimally biased.

MULTIVARIATE VARIANCES*

It is clear that genes, like most control factors, often affect more than one trait and that genes affecting different traits are often linked. Therefore, anything done to change one trait by manipulating genes will affect other traits. It therefore behooves us to consider that genetic covariances among traits provide information on the total variability and correlations that exist in forests. The only element of difference that genetic sources of covariance among traits creates is the possibility that the covariance is due to either correlated effects of the same genes, or to the existence of correlated frequencies among genes at loci which otherwise act independently. Both genetic sources of correlation among traits can cause very rapid and large changes in the correlation if selection is applied to the population or if relatively small populations are permitted to breed. Otherwise, we can treat the analysis of multivariate systems by standard means and can treat genetic sources of variance and covariance as simply another control variable in multivariate analyses of correlated traits.

Multivariate analysis in genetics has included selection index construction, cluster and distance analyses, and some attempts to simply reduce the total number of yield variates to a manageable number. In this section, we merely wish to develop the basic models and analysis as extension of the previously introduced concepts of variances and regression.

If a new equation is written for each of several Y variates with their respective effective loci represented, the covariance between traits may be expressed in terms of the correlated or pleiotropic effects of those commonly held genes and their allelic frequencies.

In the above notation, traits A and B with a commonly effective

*Graduate-level statistical training required for thorough understanding.

locus can be written as linear genetic models:

$$Y_{Aij}=\mu_A+\alpha_{Ai}+\alpha_{Aj}+\delta_{Aij}$$

$$Y_{Bij}=\mu_B+\alpha_{Bi}+\alpha_{Bj}+\delta_{Bij}.$$

The additive genetic variance in the individual traits is $2\Sigma p_i\alpha_i^2$. Here p_i are the allelic frequencies, α_i are the average allelic effects, and summation is over all alleles. The additive genetic covariance is:

$$2\Sigma_i\alpha_{Ai}\alpha_{Bi}.$$

Extensions to more inclusive models are direct. Most of the present work on multivariate genetic analysis is on this simple genetic basis. It is possible to obtain pseudopleiotropic effects in estimating additive genetic covariances without true genetic pleiotropy if linkage disequilibrium or other disequilibrium is present and causes an association of traits by correlating the frequency of alleles at different loci. However, if equilibrium conditions are assumed, the covariances of additive, dominance, and other effects are derivable for pairs of traits in covariance analyses just as the genetic variances are. The genetic covariance matrix is our multivariate analog of the simpler univariate genetic variances and has all of the sampling and interpretation problems of the univariate models extended into p dimensions.

Aside from distribution and hypothesis testing in multivariate analysis, the interest of geneticists lies in two main directions. One is towards reducing the number of variables to a more easily handled set. In these cases, the techniques of principal component and factor analyses have been pursued. The other direction of research is into the matrical representation of genetic effects, such as might be convenient for linkage studies or any genetic study extended to the multivariate case. In this direction also, studies of special interest for provenance research lie in determining the dimensionality of the space defined by species, hybrids, races, or provenances (for example, Namkoong 1967; Misra 1966). These latter studies usually employ the techniques of canonical analysis and use the vectors corresponding to the roots of $(B-\lambda W)=0$ equations to obtain scales on which to measure divergence or similarity. Recently, Rouvier (1966) has proposed a canonical analysis with a rotational transformation to the principal component factor of the W matrix.

Much provenance research involves the discernment of relations between environmental factors and yield factors. For example, after a local provenance test has been run, the breeder often wishes to estimate the relations between environmental variables at the seed source and performance in his plantation. The interest for population genetics lies in determining the extent of genetic segregation in allelic frequencies and whether substantial genetic variance exists within or between stands, or both. The extent to which variation in the several traits of interest is determined

by environmental factors indicates the relative strength of directional selection and migration versus drift and other random forces in determining allelic frequencies. The analysis of multiple regression in several traits simultaneously is therefore of value in interpreting genetic population structure. The genetic covariance matrix estimated after interpopulation effects are removed represents the multivariate analog of the simple genetic variance within populational subdivisions. One might wish to simplify interpretation by using canonical or component analysis, but estimates of the total regression and residual genetic covariance on all the traits should also be made.

The matrix of $\frac{p(p+1)}{2}$ genetic variances and covariances is therefore estimable, and general linear hypotheses (on additive or dominance effects in multivariate space, for example) can be tested by multivariate analogs of univariate analyses of variance tests. Thus, maximum likelihood testing on the dispersion matrix among provenances or among f half-sib families in p traits is performed by comparing the statistic:

$$-2nln[\ |W|\ /\ |W+B|\ ]$$

with χ^2 with $2(f-1)$ degrees of freedom,

where n = total sample size,

f = number of families,

W = dispersion matrix for error,

B = dispersion matrix for families,

$|W|$ and $|W+B|$ = generalized variances of their respective variates.

Several linear hypotheses can be tested by this criterion (Kendall and Stuart 1966) as well as by criteria based on the distribution of the roots of $|B-\lambda W|=0$ (Roy 1957).

Various kinds of value functions can be made up to provide simple measures of value by functionally incorporating the joint values of the several yield variates. The selection index is one kind of such value function that can be applied, is linear, and is determined such that it maximizes selection gain. Other criteria can be applied to nonlinear value functions (Namkoong 1970b) and will be discussed.

Standard definitions of gene effects and variances can thus be extended to the multivariate case. This analytical form may well be very important in forestry. In any case, the gene effects thus described and the relative allocation of differences (variances) between and within families permit estimation of the structure of variation in populations.

CHAPTER 8
ESTIMATING GENETIC PARAMETERS

While it is desirable to measure population means, variances, covariances, regressions, etc., because they are useful descriptors of population characteristics, the method of estimating these parameters is not immediately obvious. In this chapter, the concept and genetic use of variance components are developed as an extension of regression concepts. Estimation techniques for standard, balanced designs are described along with techniques for analyzing unbalanced experiments. The relationships between estimable experimental variance components and genetic variances are shown, and experimental designs suited to estimating genetic variances are then explored. The principal problem in estimation is to determine a reasonably good use of the sample data for accurate estimation of those parameters, at least on the average. However, the existence of variation in the population necessarily implies that any resampling or other new independent sampling of the population will give us a different set of data and therefore different estimates of any of the parameters. It is useful to know not only the best estimate of the parameter values, but also how much variation we might expect any new results to exhibit if new sample estimates were derived. For example, if breeding program decisions were to be based on the level of genetic variances and different estimates of the variances were available, the relative reliabilities of the estimators would be critical information. However, variations can be generated by many causal factors. Some may be controllable or measurable and adjusted for, while others would be uncontrolled and could cause unavoidable error in estimating means or variances.

Thus, there are two general means of reducing errors of estimation. If the sources of variation are identifiable and controllable, they may then be fixed or their contributions to the variation adjusted for. On the other hand, if the sources of variation are not controllable or their variances' contributions are to be estimated, then sampling among the units of variation may be increased to reduce error in estimating the variance parameters.

As an example of the first case, the variance of estimates of average wood specific gravity of slash pine was $793 \diagup 10^{-6}$, but much of this variation was due to differences among clones. When the clonal variations were removed, the residual error variation due to uncontrolled sources of environment and error was only

430×10^{-6} (Zobel and others 1962). If the objective is to estimate a tree's specific gravity precisely and clonal variations are extrinsic to this objective, then clonal differences add to the error of estimation. In this case, when genetic variations were removed, the mean specific gravity was more precisely estimable; more samples from within specific clones would provide greater precision. Thus, the variation in any parameter estimate is subject to how the sample is drawn and how the population is restricted. It is to be noted, however, that there almost always is a residual variance which cannot be adjusted for, and even adjusted statistics rarely estimate a parameter exactly.

As an example of the alternative case, however, the objective of the experiment may be to estimate the extent of variation among clones or to include clonal variation in estimates of total population means. In that case, more clones rather than fewer should be sampled to reduce the error of the total population mean as well as to estimate the variation due to clones. The concern has shifted from estimating means of a fixed set of units to one of estimating both means and variances of a more widely sampled set of units. It is then generally assumed that a random sample will provide an experimental set of units which will represent the types and proportions of effects as present in the wider population. The model and interest have thus shifted from the fixed effects of specific experimental or test entries to the random effects of a variable population.

It may be obvious for estimates of means that larger sample sizes increase precision within some restricted population. It is also true for estimates of variances that more samples of clones, families, or whatever factor causes variation will also increase precision in estimating the variance due to those factors. When parameters are sums of squared effects (i.e., variances) rather than sums of direct measures, the same results hold true with respect to error of estimation and its control. Thus, if sums and means are estimated with some variance, then sums of squares and mean squares are also estimated with some imprecision. In a very large experiment with loblolly pine, Stonecypher (1966) obtained a direct measurement of the variance in estimated variances by analyzing different sample blocks separately and showed that the variance estimates differed. In that case, some of the differences were due to variations in sites and years, but a large portion was due simply to sampling variations in drawing different sample replicate blocks. Whenever there is variation in the basic data, all derived estimates of parameters such as means, variances, or higher moments will exhibit variation. By analyzing the sampling variance we can help ourselves in two ways. First, we can determine the reliability of the estimation statistics under the conditions of the experiment and possibly under greater or lesser sampling restrictions. Second, we can determine what factors affect the sizes of the errors of estimation and therefore can plan future experiments to provide predetermined levels of precision.

It is beyond the scope of this publication to construct the distributions of stochastic processes. There are many texts available which describe moments and estimators of moments for the commonly assumed distributions. Some characteristics of error distributions, parameter estimates, and the variance of the estimators will simply be asserted here. The discussion will generally be confined to simple, linear models, but some solutions for more general conditions will be indicated.

In the multiple regression concepts previously outlined, the independent variables x_i were assumed to have some average proportionate effect (b) on the size of the yield variate Y_{ij}. Hence:

$$Y_{ij}=b_0+b_{1j}X_{1j}+b_{2j}X_{2j}\,b_{2j}X_{2j}+\ldots e_{ij}.$$

The total sums of squares in Y was seen to have been reduced by accounting for the regression effects (or could be reduced by adjusting for the regression), by an amount $\underline{b}'(X'\underline{Y})$. The sums of squares thus derived as being accounted for by the regression could also have been written as $\underline{b}'(X'X)\underline{b}$. We could just as reasonably state the relationship between the Y and X variables as the existence of an average regression effect in Y for each X variable chosen. If the various X variables are not controlled or specifically chosen in the experiment or if the X levels in the experiment are to depend only on the frequency of their occurrence in nature, then the variation in the X or b effects would itself be a variance statistic of interest. The emphasis in analyzing the relationships changes. In simple regression, a single X variable has many levels and the objective is to estimate an average regression response for a given range in X. In simple analysis of variance, a single X variable (for example, a family i) has a single response level (b_i) and the objective is to estimate the variations in Y caused by samples of many different b_i effects. For example, family 1 may have an average deviation effect of $+5$, and family 2 $a-5$, etc. Then individual trees would have Y variate measures of the mean, plus or minus the family effect, plus an error. Two trees from family 1 would have effects added as:

$$Y_{11}=\mu+b_1X_1+0+e_{11}$$

$$Y_{12}=\mu+b_1X_1+0+e_{12}$$

Family 2 may have several trees:

$$Y_{21}=\mu+0+b_2X_2+e_{21}$$

$$Y_{22}=\mu+0+b_2X_2+e_{22}$$

$$Y_{23}=\mu+0+b_2X_2+e_{23}$$

If the size of the family effect (for example, ± 5) is taken as the b coefficient, then $X_1=1$ whenever family 1 is measured, and zero otherwise. Similarly, $X_2=1$ only if the tree measured is family 2 and is zero otherwise. The above set of Y measures would then carry a model:

$$
\begin{aligned}
Y_{11} &= \mu + b_1 X_1 + \quad 0 \ldots + e_{11} \\
Y_{12} &= \mu + b_1 X_1 + \quad 0 \ldots + e_{12} \\
Y_{21} &= \mu + \quad 0 + b_2 X_2 \ldots + e_{21} \\
Y_{22} &= \mu + \quad 0 + b_2 X_2 \ldots + e_{22} \\
Y_{23} &= \mu + \quad 0 + b_2 X_2 \ldots + e_{23} \\
&\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
&\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\
&\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots
\end{aligned}
$$

$$
\underline{Y} = \begin{pmatrix} 1 & 1 & 0 & \ldots \\ 1 & 1 & 0 & \ldots \\ 1 & 0 & 1 & \ldots \\ 1 & 0 & 1 & \ldots \\ 1 & 0 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} \begin{pmatrix} \mu \\ b_1 \\ b_2 \\ \ldots \\ \ldots \\ \ldots \\ \ldots \\ \ldots \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{23} \\ \ldots \\ \ldots \\ \ldots \end{pmatrix}
$$

$$
Y = \qquad (X) \qquad\qquad (\underline{b}) \quad + \quad (\underline{e})
$$

The form is the same as for the simple regression except that the X's are counting indices and are known in the experiment. The effects of the X's on the Y are determined by the b's, and these are expected to reflect some sample of effects from the population the sample was drawn from. Then, the sum of squares due to variations in accounting for the regression effects would be:

$$
\begin{aligned}
E[\underline{b}'(X'\underline{Y})] &= E[(X\underline{b}+\underline{e})'X(X'X)^{-1}(X'Xb+X'\underline{e})] \\
&= \mathrm{E}(\underline{b}'X'X\underline{b}+\underline{e}'\underline{e}) \\
&= E(\underline{b}'X'X\underline{b}) + (df)\sigma^2.
\end{aligned}
$$

Then, the expected value of the mean square due to regression effects is:

$$
\frac{E[\underline{b}'(X'\underline{Y})]}{df} = \sigma^2 + \frac{E(\underline{b}'X'X\underline{b})}{df}.
$$

This form of the mean square due to regression effects now requires some concept of what those effects are and how they were sampled in the population in order to interpret the term $E(\underline{b}'X'X\underline{b})$. It is useful to define the effects arbitrarily as causing deviations around the general mean. Then the mean of the b effects would be zero. In addition, if the population is assumed to have been randomly sampled, then the covariance between the randomly sampled effects would be zero. If it is further assumed that the variance throughout the population thus sampled was the same, that is, the population was not subdivided into segments with different means or variances, then $E(b_i-b)^2$=variance among regression effects$=\sigma_b^2$. The definition of effects as deviations requires that $E(b)=0$, and the assumption of randomness requires that $E(b_ib_j)=0$ if $i\neq j$.

Now, it can be seen from the above definitions of the matrix X that

$$X'X=\begin{pmatrix} n & 2 & 3 & \dots \\ 2 & 2 & 0 & \dots \\ 3 & 0 & 3 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

$$\text{that } (X'X)\underline{b}=\begin{pmatrix} n & 2 & 3 & \dots \\ 2 & 2 & 0 & \dots \\ 3 & 0 & 3 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}\begin{pmatrix} \mu \\ b_1 \\ b_2 \\ \dots \end{pmatrix}$$

$$=\begin{pmatrix} n\mu+2b_1+3b_2+\dots \\ 2\mu+2b_1+\ 0\ +\dots \\ 3\mu+\ 0\ +3b_2+\dots \\ \dots\ \dots\ \dots\ \dots \end{pmatrix}$$

$$\text{that } b'(X'X)b=(\mu,\ b_1,\ b_2\dots)\begin{pmatrix} n\mu+2b_1+3b_2\dots \\ 2\mu+2b_1\quad 0\ \dots \\ 3\mu+\ 0\ +3b_2\dots \\ \dots\ \dots\ \dots\dots \end{pmatrix}$$

$$\begin{aligned} = \ & n\mu^2\ +2\mu b_1+3\mu b_2+\dots \\ & +2\mu b_1+2b_1^2+\ 0\ \ +\dots \\ & +3\mu b_2+\ 0\ \ +3b_2^2+\dots \\ & +\dots \end{aligned}$$

If the above assumptions are valid, then

$$E(\underline{b}'X'X\underline{b}) = n\mu^2 + 2E(b_1^2) + 3E(b_2^2) + \ldots$$
$$= n\mu^2 + 2\sigma_b^2 + 3\sigma_b^2 + \ldots,$$

and if the correction factor for the mean is subtracted, this term includes only $\sigma_b^2(\Sigma X_i^2)$ where the ΣX_i^2 coefficient is only the number of independent times each of the b_i^2 elements are included. In this case, it would be $(2-1)$ for b_1, $(3-1)$ for b_2, etc. Therefore, the expected value of the mean square due to regression is $\sigma^2 + \sigma_b^2(\Sigma X_i^2)$, and our only problem is to count the number of independent σ_b^2 elements there are (i.e., ΣX_i^2) in what has been constructed as the mean square. Whatever the effects are which we try to estimate as a contributor to the population variance, the form of the analysis is the same. Our great interest is in genetically related sources of variation which can be of many different kinds.

Family differences or fertility variations can be treated as sources of variance and can come in several forms. Hence, they would have to be interpreted in terms of the kinds of effects and variances that they measure. Thus, fertility-caused variations may be quite different if we measure nitrogen rather than iron levels in forests, and genetically caused differences are quite different if we measure differences among full-sib families rather than half-sib families. For the moment, however, consider that a single effect like families is sampled from a large population and that the variance in yield due to family differences (σ_f^2) is to be estimated. We shall try to compose squared sums so that we can estimate the components of variation due to error (σ^2) and due to the variation among the regression or family effects (σ_b^2 or σ_f^2).

ESTIMATING VARIANCE COMPONENTS IN ANALYSES OF VARIANCE

For the several sources of variance which we wish to estimate in an experiment, we can compose several analyses of variance to estimate sizes of the components of variance of those sources.

In general, we should consider that variances can be estimated in many different ways. For example, we might construct different combinations of observations which, when squared, give different variance functions and which may then give estimates of the contribution of each of the component sources of variance. In particular, for unbalanced experiments, variances can be estimated efficiently by constructing sums of squares different from those that would be constructed for testing significance of treatment effects. However, for balanced experiments, it can be shown that the usual kinds of analyses of variance require mean squares which are, in fact, unbiased estimators of the components of variance and that those estimators have least sampling variance of all possible quadratic forms. A familiar example is the randomized block experiment, where $Y_{ijk} = \mu + b_i + f_j + bf_{ij} + e_{ijk}$, with r blocks

($i=1, 2 \ldots r$), f families ($j=1, 2 \ldots f$) and n seedlings ($k=1, 2 \ldots n$) per family plot. The analysis of variance (ANOVA) is shown in table 6 in which a dot indicates summation over the subscript for which it is substituted. Then:

$$\begin{bmatrix} EMS \text{ families} \\ EMS \text{ plot error} \\ EMS \text{ within} \end{bmatrix} = \begin{bmatrix} 1 & n & nr \\ 1 & n & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_w^2 \\ \sigma_p^2 \\ \sigma_f^2 \end{bmatrix}$$

Each variance component can then be estimated since there are three linear equations with only the three unknown components.

Even for unbalanced experiments, a good general procedure to follow is to write out the linear model of yield for each experimental unit, and to then determine the sums, squared sums, and differences required in the usual kinds of ANOVA. We must then compute the expected values of the model components when they are summed and squared as required by the ANOVA formulations for obtaining sums of squares. This can always be done regardless of either the balance of the experiment or the particular sum of squares computed. It only requires that whatever experimental observations are summed and squared, we also sum and square the corresponding model elements in the same way and hence determine the expectations of the sums of squares in terms of the elements in the models. Such procedures are explicitly traced by Anderson and Bancroft (1952, chs. 17 and 18), Searle (1971, chs. 9 to 11), and Graybill (1961, ch. 16) for some common types of experimental designs. If the experiment is unbalanced, the traditional types of sums of squares can always be computed to give mean squares which would often unfortunately contain all of the variance components. Since none of the mean squares would contain clean estimates of any components, the solution would require the simultaneous estimation of all components. Squillace and others (1967) used this technique to estimate several variance components of height growth in an unbalanced western white pine experiment. They derived nine sums of squares as if the data were balanced and found the coefficients for nine variance components so derived. Thus, the 9×1 column vector of mean squares ($\overline{MS}$) was equated to a 9×9 matrix of coefficients (A) multiplied by the 9×1 column vector of variance components (σ_i^2). Then, since $\overline{MS} = (A)\sigma_i^2$, the nine variance components were estimated by $(A)^{-1}\overline{MS} = \hat{\sigma}_i^2$. This is a readily usable way to estimate the components, but it involves very high errors of estimation.

A simpler method for calculating a set of independent sums of squares, which also provides the expected values of those sums of squares in terms of the variance components, is the Abbreviated Doolittle method. The method is well described elsewhere and requires no review here (Anderson and Bancroft 1952).

Table 6.—Analysis of variance for a randomized block experiment

Source of variance	*df*	Sum of squares	Expected mean squares
Blocks	$(r-1)$	$\sum_i \frac{Yi..^2}{fn} - \frac{Y...^2}{rfn} = SSB$	
Families	$(f-1)$	$\sum_j \frac{Y.j.^2}{rn} - \frac{Y...^2}{rfn} = SSF$	$\sigma_w^2 + n\sigma_p^2 + nr\sigma_f^2$
Plot error	$(r-1)(f-1)$	$\sum_{ij} \frac{Yij.^2}{n} - SSB - SSF + \frac{Y...^2}{rfn} = SSE$	$\sigma_w^2 + n\sigma_p^2$
Within plot error	$rf(n-1)$	$\sum_{ijk} Yijk^2 - \sum_{ij} \frac{Yij.^2}{n}$	σ_w^2

UNBALANCED DESIGN ANALYSES*

For our purposes, the greatest utility of the Abbreviated Doolittle procedure lies in composing sets of independent sums of squares for unbalanced experiments and determining their expected values in terms of squared components. We can illustrate its use in a simple experiment where seedlings from open-pollinated females have been planted in a randomized block but in which some plots are missing due to causes which are independent of the measured trait. In the following example, the grand sum (G) and each block (B_i) and family (F_j) sum can be seen to contain the indicated amounts of each μ, b_i, and f_j effect. If the data were balanced, the sum of squares for families (treatments) could be computed as $\frac{1}{r}\sum_j (F_j - \frac{G}{f})^2$, or since $\sum_i B_i = G$, the sum of squares for families equals $\frac{1}{r}\sum_j (F_j - \frac{1}{f}\sum_i B_i)^2$.

Block	Families				Block sums
	1	2	3	4	
1	Y_{11}	Y_{21}	Y_{31}	Y_{41}	$B_1 = Y_{\cdot 1}$
2	Y_{12}	Y_{22}	Y_{32}	Y_{42}	$B_2 = Y_{\cdot 2}$
3	Y_{13}	Y_{23}	Y_{33}	Y_{43}	$B_3 = Y_{\cdot 3}$
Family sums	$F_1 = Y_{1\cdot}$	$F_2 = Y_{2\cdot}$	$F_3 = Y_{3\cdot}$	$F_4 = Y_{4\cdot}$	$G = Y_{\cdot}$

Effects included in yield sums								
μ	b_1	b_2	b_3	f_1	f_2	f_3	f_4	Yield sums
12	4	4	4	3	3	3	3	G
4	4	0	0	1	1	1	1	B_1
4	0	4	0	1	1	1	1	B_2
4	0	0	4	1	1	1	1	B_3
3	1	1	1	3	0	0	0	F_1
3	1	1	1	0	3	0	0	F_2
3	1	1	1	0	0	3	0	F_3
3	1	1	1	0	0	0	3	F_4

Each F_j sum contains 3 μ elements and one each of b_i effects in addition to three of its own f_j elements, as indicated in the above table which is essentially the X matrix of coefficients. Suppose however, that family 3 is missing from block 1 and family 2 is missing from block 3. By determining the content of each B_i/f, and subtracting it from F_j, the X matrix of coefficient becomes:

*Graduate-level statistical training required for thorough understanding.

Effects included in yield sums

μ	b_1	b_2	b_3	f_1	f_2	f_3	f_4	Yield sums
10	3	3	3	3	2	2	3	G
3	3	0	0	1	1	0	1	B_1
4	0	4	0	1	1	1	1	B_2
3	0	0	3	1	0	1	1	B_3
3	1	1	1	3	0	0	0	F_1
2	1	1	0	0	2	0	0	F_2
2	0	1	1	0	0	2	0	F_3
3	1	1	1	0	0	0	3	F_4

However, it is still possible to adjust the treatment sums as to obtain sums (F^*) which are clear of μ and b_i effects as follows:

μ	b_1	b_2	b_3	f_1	f_2	f_3	f_4	Adjusted family sums
0	0	0	0	25/12	−7/12	−7/12	−11/12	$F_1-[1/3B_1+1/4B_2+1/3B_3]=F_1^*$
0	0	0	0	−7/12	17/12	−3/12	−7/12	$F_2-[1/3B_1+1/4B_2]=F_2^*$
0	0	0	0	−7/12	−3/12	17/12	−7/12	$F_3-[1/4B_2+1/3B_3]=F_3^*$
0	0	0	0	−11/12	−7/12	−7/12	25/12	$F_4-[1/3B_1+1/4B_2+1/3B_3]=F_4^*$

It is now necessary to derive independent sums for the sums of squares to be additive, but since the data are unbalanced, the above sums must be further adjusted. The Abbreviated Doolittle procedure may be followed a further step to provide the sums of squares as follows:

f_1	f_2	f_3	f_4	Adjusted family sums
25/12	−7/12	−7/12	−11/12	F_1^*
1	−7/25	−7/25	−11/25	$F_1^* \cdot 12/25$
	376/300	−124/300	−252/300	$F_2^*-7/25\ F_1^*=F_2^{**}$
	1	−124/376	−252/376	$F_2^{**} \cdot 300/376$
		126/1128	−126/1128	$F_3^*-7/25\ (F_1^*)$
			−1	$-124/376\ F_2^{**}=F_3^{**}$
		1		$F_3^{**} \cdot 1128/126$
		0	0	0

By sweeping out all but family effects, the adjusted sums of squares can be computed from the model effects in the left-hand side of the Abbreviated Doolittle matrix. In our example, the expected value of the sums of squares for the family effects is:

$$2.0833f_1^2+1.4167f_2^2+1.4167f_3^2+2.0833f_4^2=3.5\ \sigma_f^2$$

since $E(f_if_j)=0$, if $i\neq j$, but $E(f_i^2)=\sigma_f^2$.

We have thus created a sum of squares unconfounded with other

main effects (μ, b_i) which can then be interpreted according to the meaning of the $\Sigma\, a_i f_i^2 = \Sigma\, X_i^2\, \sigma_f^2$. The Abbreviated Doolittle forward solution in effect transforms the X matrix into a matrix Z in which rows are orthogonal linear functions of the original X rows. Thus, $X=ZB$ and hence $X'X=B'Z'ZB$. Since Z contains orthogonal rows, $Z'Z$ is a diagonal matrix (D), and $X'X=B'DB$; and $(X'X)\underline{b}=X'\underline{Y}$ is therefore transformed into $B'DB\underline{b}=B'Z'\underline{Y}$. If B^{-1} exists, then $DB\underline{b}=Z'\underline{Y}$, and in the Abbreviated Doolittle forward solution, we find the matrix $A=DB$ and $DB\underline{b}=Z'\underline{Y}=A\underline{b}$. Thus, by transforming the X into ZB, we can see that the corresponding b is transformed into $B\underline{b}=\underline{b}^*$. The sum of squares due to the regression is invariant under these transformations and can then be seen to be:

$$\begin{aligned} \underline{b}'X'\underline{Y} &= \underline{b}^{*\prime}\ (B')^{-1}B'Z'\underline{Y} \\ &= \underline{b}^{*\prime}Z'\underline{Y} \\ \text{or} \qquad &= \underline{b}'B'A\underline{b} = \underline{b}'B'DB\underline{b} = \underline{b}'X'X\underline{b}. \end{aligned}$$

As given in the earlier notation, this is the expected value of the sum of squares due to regression (SSR) for the regression effects. $E(e'e)$ must be added to this to complete sum of squares due to regression. The Abbreviated Doolittle forward solution provides the A and B matrices in the form:

$$\begin{array}{cccc|c} & & (X'X) & & X'\underline{Y} \\ \hline A_{11} & A_{12} & A_{13} & A_{14} & \cdots \\ 1 & B_{12} & B_{13} & B_{14} & \cdots \\ \hline & A_{22} & A_{23} & A_{24} & \cdots \\ & 1 & B_{23} & B_{24} & \cdots \\ \hline & & A_{33} & A_{34} & \cdots \\ \hline & & 1 & B_{34} & \cdots \end{array}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & \cdots \\ & A_{22} & A_{23} & A_{24} & \cdots \\ & & A_{33} & A_{34} & \cdots \\ & & & A_{44} & \cdots \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & B_{12} & B_{13} & B_{14} & \cdots \\ & 1 & B_{23} & B_{24} & \cdots \\ & & 1 & B_{34} & \cdots \\ & & & 1 & \cdots \end{bmatrix}$$

It can also be observed that the b regression coefficients have been changed, and that estimates of the new effects, b^*, can be converted back to the originally defined ones by the inverse transformation. This is computationally simple in the backward solution of the Abbreviated Doolittle. Since the sums of squares are invariant under these transformations, however, we can interpret the meaning of the $\Sigma a_i f_i^2$ in terms of variances of these f_i or b^* effects. If we sampled at random, and all the f_i are independent, then $E(f_i f_j)=0$, $i \neq j$. If in addition, the f_i are deviations from some general mean, then $\Sigma f_i=0$, $\bar{f}=0$. Hence, $E(f_i^2)=E(f_i-\bar{f})=\sigma_f^2$, and the number of such elements in each squared sum is computed in the Abbreviated Doolittle by the $B'A=\sum_i A_{h_i} B_{h_i}$ for each of the i effects in the hth squared sum. The number of such squared sums that yield any σ_f^2 is the number of degrees of freedom.

In general, it is not possible to adjust main effects for interactions. Therefore, while one can adjust any main effects for all other main effects simply by listing those desired "clean" effects last, interactions involving the main effect will be included in its sums of squares. If nonindependence among the f_i is assumed, or any difference among the $E(f_i^2)$ exists, then these effects too can be traced by completely writing out the $B'A$ products.

In unbalanced data, the general objective of the various computing procedures such as the Abbreviated Doolittle is to adjust various sums for other extrameans effects. Unless some of the effects contain interactions, we can view the problem as one of transforming the A matrix to a partitioned upper-triangular matrix in the equation:

$$\left[\begin{array}{c|c} A_{mm} & A_{mq} \\ \hline A'_{mq} & A_{qq} \end{array}\right] \left[\begin{array}{c} b_m \\ \hline b_q \end{array}\right] = \left[\begin{array}{c} g_m \\ \hline g_q \end{array}\right]$$

A direct method requires that certain inverses exist or that generalized inverses be found. Then:

$$\left[\begin{array}{c|c} I_m & 0 \\ \hline -A_{mq}A_{mm}^{-1} & I_q \end{array}\right] \left[\begin{array}{c|c} A_{mm} & A_{mq} \\ \hline A'_{mq} & A_{qq} \end{array}\right] \left[\begin{array}{c} b_m \\ \hline b_q \end{array}\right] = \left[\begin{array}{c|c} I_m & 0 \\ \hline -A_{mq}^{-1}A_{mm}^{-1} & I_q \end{array}\right] \left[\begin{array}{c} g_m \\ \hline I_q \end{array}\right]$$

where $A^*_{qq} = A_{qq} - A_{mq}' A_{mm}^{-1} A_{mq}$

and $g^*_q = g_q - A_{mq}' A_{mm}^{-1} g_m$.

Other transformations can similarly be made to obtain zeros in the lower-left partition, but all require some direct inversion of submatrices. The great advantage of the Abbreviated Doolittle method is that the inversions do not have to be made directly. Since almost all ANOVA will have (A) matrices with several

singularities, direct inversion is very difficult even with generalized inverse programs. Thus, when missing plots or otherwise unbalanced data exist and linear dependencies are created in forms difficult to detect, the Abbreviated Doolittle provides sums of squares with all dependencies removed.

A third possibility in computing sums of squares and their expectations when unbalance exists is to simply compute the sums of squares as if no plots were missing. Then the expectations of the sums of squares can be determined simply by repeating the summing and squaring operations on the model components included in the appropriate yield variables. While not an elegant procedure, this one can be used if all other procedures fail.

For balanced experiments the ANOVA's are usually easily determined, and various algorithms are available for finding the appropriate expectations of the mean squares. Many methods have been described for determining appropriate ANOVA's for complicated replication, treatment factorial, and nested designs and their expectations under assumptions of fixed, mixed, or variance component models. They are not reviewed here.

DISTRIBUTIONS OF VARIANCE COMPONENTS*

While it is clear that unbiased estimates of the variance components can be obtained, it is also clear that a resampling of the original population would yield different estimates. Like any other estimator with sampling error, the error distribution is used for determining reliability of estimates as well as for designing good experiments. It can be shown that if elements drawn from an $N(0,1)$ distribution are squared, the distribution of the squared elements is a X^2 and that squaring elements from an $N(0, \sigma^2)$ yields variates with a $X^2 \cdot \sigma^2$ distribution. In the ANOVA, the effects of any of the sources of variances are corrected for the mean. Hence, they have a zero expectation and the variance of those sums is usually identical to the expected mean square (EMS). Therefore, these sums are distributed $\sim N(0, EMS)$ and the sum of squares is distributed $\sim X^2 \cdot EMS$. The variance of the sum of squares is $(EMS)^2 \cdot$ (variance of the X^2). Therefore, to compute the variance of the mean square, we require only the variance of the X^2 which is 2 (df). The variance of the mean square is

$$\frac{2 \cdot df(EMS)^2}{df^2} = \frac{2(EMS)^2}{df}.$$

Since the variance components are linear functions of the mean squares, the variance of those linear functions would determine the variances of the components. Thus, if

$$\hat{\sigma}_1^2 = \frac{MS_1 - MS_2}{k},$$

*Graduate-level statistical training required for thorough understanding.

as is usually the case, then

$$V(\hat{\sigma}_1^2) = \frac{1}{k^2}\left[V(MS_1) + V(MS_2) - 2 \text{ Cov } (MS_1, MS_2) \right].$$

If the mean squares are orthogonal, the covariances are all zero, and

$$V(\hat{\sigma}_1^2) = \frac{2}{k^2}\left[\frac{(EMS_1)^2}{df_1+2} + \frac{(EMS_2)^2}{df_2+2} \right].$$

It has been empirically determined that the addition of 2 to the differences in the denominators gives better fits. The square root of the variance is the standard error of the estimated variance component, which is frequently used as a measure of the significance of the component. Thus, while the distribution of the variance component is not a X^2 (since the mean squares are different), the variance is easily computed and various ways to estimate confidence intervals are available (Anderson and Bancroft 1952, ch. 22; Graybill 1961, ch. 17; Searle 1971, ch. 9). Whenever any unbalances exist in the analysis, it can easily be seen in the Abbreviated Doolittle that the individual squared sums do not represent identical estimates of the same $X^2\,\sigma^2$ distribution and, hence, that the mean squares are not $X^2\,\sigma^2$ variates. Nevertheless, the variances of each of the independent contrasts can still be estimated, and if orthogonal sums of squares are computed as provided in such procedures as the Abbreviated Doolittle, the different mean squares will be uncorrelated.

Regardless of the method used to obtain the sums of squares, it is always possible to determine not only the expected values, but also their variances and whatever covariances exist due to imbalance and nonorthogonality of the sums of squares. One method of computing is to write the sum of squares for each source of variation in quadratic form: $\underline{Y}'Q_T\underline{Y}$ where $\underline{Y}$ is the vector of all of the observations and Q_T is a matrix of coefficients which gives the appropriate weighting for the observations in the sum of squares for the T source of variance.

For example, if an experiment contained four treatments (i) with three random replicates (j) each, the sum of squares for treatments (SST) would be:

$$\text{SST} = \frac{\sum_i 4\left(\sum_{j=1}^{3} Y_{ij}\right)^2}{12} - \frac{\left(\sum\sum Y_{ij}\right)^2}{12}$$

$$= 1/12[4(Y_{11}+Y_{12}+Y_{13})^2 + 4(Y_{21}+Y_{22}+Y_{23})^2 + 4(Y_{31}+Y_{32}+Y_{33})^2 + 4(Y_{41}+Y_{42}+Y_{43})^2 - (Y_{11}+Y_{12}+\ldots+Y_{43})^2]$$

$$= (Y_{11}, Y_{12} \ldots Y_{43})$$

$$\begin{bmatrix} 4 & 4 & 4 & -1 & -1 & -1 & -1 & -1 & \ldots \\ 4 & 4 & 4 & -1 & -1 & -1 & -1 & -1 & \ldots \\ 4 & 4 & 4 & -1 & -1 & -1 & -1 & -1 & \ldots \\ -1 & -1 & -1 & 4 & 4 & 4 & -1 & -1 & \ldots \\ -1 & -1 & -1 & 4 & 4 & 4 & -1 & -1 & \ldots \\ -1 & -1 & -1 & 4 & 4 & 4 & -1 & -1 & \ldots \\ -1 & -1 & -1 & -1 & -1 & -1 & 4 & 4 & \ldots \\ -1 & -1 & -1 & -1 & -1 & -1 & 4 & 4 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix} \begin{bmatrix} Y_{11} \\ Y_{12} \\ . \\ . \\ . \\ . \\ . \\ . \\ Y_{43} \end{bmatrix}$$

$$= \underline{Y}'Q_T\underline{Y}.$$

A matrix of covariances over i, j, k, and e, $\mathrm{Cov}(Y_{ij}, Y_{kl})$, can then also be constructed in which each expected cross product is derived in terms of the model components. For example:

if $\quad Y_{ij}=\mu+t_i+e_{j(i)}$

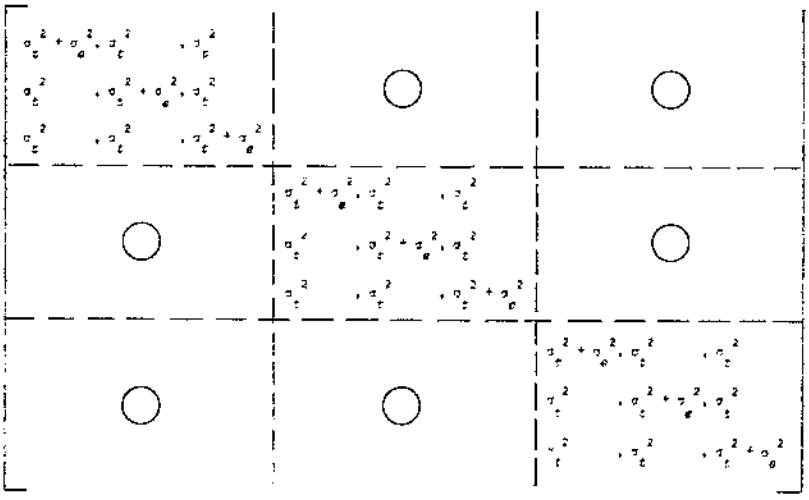
then $\quad E(Y_{11} \cdot Y_{11})-E(Y_{11})^2=\sigma_t^2+\sigma_e^2$

$\quad E(Y_{12} \cdot Y_{13})-E(Y_{12})E(Y_{13})=\sigma_t^2$

and $\quad E(Y_{11} \cdot Y_{21})-E(Y_{11})E(Y_{21})=0.$

Then the covariance matrix (V) would have the form:

$$V =$$



It can then be shown that $E(\mathrm{SST}) = VQ_t$.

It can also be shown that $V(\mathrm{SST}) = tr(VQ_t \cdot VQ_t)$

and that $\mathrm{Cov}\ (\mathrm{SST}, \mathrm{SSR}) = tr(VQ_t \cdot VQ_e)$

where tr signifies the trace of the argument matrix.

Thus, for any sum of squares, the expectations, variances, and covariances can always be found though this might be tedious.

Estimates and the variances for variance components can then also always be derived, even if the distribution is unknown.

For many statistics of interest to geneticists, various functions of the variance components are constructed and the variance of these constructed statistics is also often desired. Thus, while estimates of certain genetic variances are sometimes sufficient information, ratios of the components in heritabilities are often also desired. If simple functions of mean squares such as Hanson (1963) derives can be used, then approximate, noncentral F distributions will do reasonably well to determine variances and confidence intervals. If the functions are not simply constructed, as is often the case in forest genetics, an appropriate asymptotic variance, as derived by Kendall and Stuart (1963, ch. 10), can often be used. If we take the complicated function of the mean squares or any other variables $X_1, X_2 \ldots$ to be $g(X_1, X_2 \ldots)$, and the expected value (mean) of each of the mean squares or other variables to be $\theta_1, \theta_2 \ldots$; then the variance of the function $g(X_1, X_2 \ldots)$ is approximately:

$$V[g(X_1,X_2 \ldots)] = \sum_{i,j}\left[\frac{\delta g}{\delta \theta_i}\frac{\delta g}{\delta \theta_j}\mathrm{Cov}(X_i,X_j)\right].$$

This relationship holds true as long as the second moments of the θ's are small relative to the means. The variance approximation can be extended to the case of the approximate covariance between two functions say g and h:

$$\mathrm{Cov}[g(X_1,X_2 \ldots), h(X_1,X_2 \ldots)] = \sum_{i,j}\left[\frac{\delta g}{\delta \theta_i}\frac{\delta h}{\delta \theta_j}\mathrm{Cov}(X_i,X_j)\right]$$

In particular, the variance of a ratio: $g=\dfrac{(X_1)}{(X_2)}$ is:

$$V(g) = \frac{\mathrm{Var}(X_1)}{\theta_2{}^2} + \frac{\theta_1{}^2\mathrm{Var}(X_2)}{\theta_2{}^4} - \frac{2\theta_1\mathrm{Cov}(X_1,X_2)}{\theta_2{}^3}$$

$$= \left[\frac{E(X_1)}{E(X_2)}\right]^2\left[\frac{\mathrm{Var}(X_1)}{[E(X_1)]^2} + \frac{\mathrm{Var}(X_2)}{[E(X_2)]^2} - \frac{2\mathrm{Cov}(X_1,X_2)}{E(X_1)\cdot E(X_2)}\right]$$

This is the form used by Osborne and Paterson (1952) and most heavily used by Namkoong and others (1969) to compute variances of heritabilities of wood quality traits. For example, assume the analysis of variance was:

Mean square	df	Expected mean square
MSM	m	$\sigma_e{}^2+n\sigma_f{}^2+ns\sigma_m{}^2$
MSF	f	$\sigma_e{}^2+n\sigma_f{}^2$
MSE	e	$\sigma_e{}^2$

and $h^2 = \dfrac{\sigma_m{}^2}{\sigma_m{}^2+\sigma_f{}^2+\sigma_e{}^2} = \dfrac{\sigma_m{}^2}{\sigma_T{}^2}$.

Then our estimate of the numerator of h^2 is $(MSM-MSF)/ns$.

The variance of the numerator is simply the variance of a difference of two $\frac{X^2\sigma^2}{df}$ kinds of variables which have no covariance. Hence, in the notation of the approximate variance function:

$$\mathrm{Var}(X_1)=2\left(\frac{1}{ns}\right)^2\left[\frac{MSM^2}{m}+\frac{MSF^2}{f}\right].$$

We can estimate the denominator of h^2 $(\sigma_T^2=\sigma_e^2+\sigma_f^2+\sigma_m^2)$ also as a linear function of the mean squares as:

$$\frac{MSM+(s-1)MSF+s(n-1)MSE}{ns}.$$

The variance of the denominator is also the sum of variances of elements, each of which is known and between which no covariance exists. Thus,

$$\mathrm{Var}(X_2)=\frac{2}{n^2s^2}\left[\frac{MSM^2}{m}+\frac{(s-1)^2MSF^2}{f}+\frac{s^2(n-1)^2MSE^2}{e}\right]$$

Since the covariance between balanced mean squares is zero, the covariance between $\hat{\sigma}_m^2$ and $\hat{\sigma}_T^2$ is simply

$$\frac{1}{n^2s^2}V(MSM)-\frac{(s-1)V(MSF)}{n^2s^2}$$

or $$\mathrm{Cov}\,(\hat{\sigma}_m^2,\hat{\sigma}_T^2)=\frac{2}{n^2s^2}\left[\frac{MSM^2}{m}-\frac{(s-1)MSF^2}{f}\right]$$

Then, since $E\ (X_1)$ is simply σ_m^2 and $E\ (X_2)=\sigma_T^2$, we substitute estimates into the $V(g)$ function and compute the sampling variance.

Thus, for almost any kind of experiment, approximate covariances of functions of sums of squares can be estimated. If the geneticist is fortunate enough to have balanced experiments to work with, distributions of the sums of squares and sums of cross products (from the analysis of covariance) are known. If the yield variates are all distributed as multivariate normal variables, $\sim N(\mu, \Sigma)$, then the distribution of the mean squares and mean cross products is called the Wishart distribution with two parameters $\left[\frac{\Sigma}{df}, df\right]$, where Σ is the matrix of sums of squares and cross products, and df is the degrees of freedom appropriate to the source of variance designated. In general, for any sum of cross products, A_{ij}, the covariance between any two sums of cross products is

$$E(A_{ij}-df\sigma_{ij})(A_{kl}-df\sigma_{kl})=df(\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}).$$

Thus, for the variance of a sum of cross products, when the

variance implies $i=k$ and $j=l$,

$$V(A_{ij}) = df(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}),$$

where σ_{ij} is the covariance of i and j and σ_{ii} is the variance of i.

For the variance of a sum of squares, we have $i=j=k=l$ and $V(A_{ii}) = df \cdot 2\sigma_{ii}^2$, where σ_{ii} is the mean square expectation for trait i. Then $V(SS_i) = 2df(MS_i^2)$ and $V(MS_i) = \frac{2}{df}(MS_i^2)$, as before.

DESIGNING GENETICS EXPERIMENTS

While it is clear that results of almost any kind of replicated experiments can be analyzed, it is also clear that the nature of future forestry experiments can be enhanced by appropriate allocation of materials among the sources of variance. If certain genetic components are important to estimate with precision, then obviously the degrees of freedom will partly control the variance of those estimates and should be maximized. Depending on the objectives of the experiments, different allocations of effort would maximize the benefit/cost ratio. In the extensive studies on variation in wood quality, Goggans (1961) appropriately allocated considerable effort to estimate variances due to several hierarchies of sampling within families, trees, sections, annual rings, and part of annual rings, but he necessarily sampled genetically distinct families lightly. Once the sampling variances were estimated, however, interest in estimating the family variances claimed higher priority, and he recommended sampling more families with a reduced amount of within-family sampling. This procedure has subsequently been followed in the North Carolina State University-Industry Cooperative Tree Improvement Program. In such programs where cost factors can be unified in a simple function of the numbers of samples at each of the sampling levels and benefit can be measured as an inverse function of the estimator variance, an optimum sampling system can be derived. If the variance of the estimator is independent of the parameter being estimated, as when means or regressions are estimated, the cost/benefit ratio can be minimized fairly directly when the conditions affecting the cost of sampling are known (Marcuse 1949). However, when estimating variance components, it can be seen that the size of the component affects the size of the mean square and, therefore, affects the variance of its own estimator. The variance can then be expected to increase with the size of the component, and designs and allocations would have to be compared on some contrived value function for all levels of the component.

A still further complicating factor in considering optimizing an experiment is the common desire to estimate more than one component with reasonable precision. In genetics experiments, the error component is often almost as important to estimate as the additive genetic variance. The dominance genetic variance may

also be of some interest. Since most experiments in forestry require considerable time and space, most are established for a variety of objectives including the measurement of several traits. Therefore, it is reasonable to assume that the prudent forester will be estimating several covariance components and will want to optimize his experimental allocations with respect to some criterion of goodness for his various objectives. We must therefore consider an appropriate value function as well as variances and covariances of estimators in experiments likely to be useful in forest genetics. In the previous chapter, the covariance of relatives was taken as a function of the genetic variances. In this section, variance components due to family effects are taken as functions of the covariance of relatives. Hence, the direct relations between estimated family variance and genetic variances are established.

If we are to consider the variances and covariances, we must first briefly review the kinds of estimators used for the genetic variances. The commonly used mating schemes provide a few mean squares which are functions of the genetic variances. Whereas in mean or regression estimation problems the investigator could choose combinations of environmental variables to minimize errors of estimates, the geneticist chooses to construct different kinds of families and controls the number of families and family members. Since the variance components are estimated from second-order statistics which roughly follow a $\chi^2\sigma^2$ distribution, the design variables which the geneticist can choose are the degrees of freedom and the composition of the expected mean squares.

As noted in the previous chapter, the degree to which family members are closely related is in some sense proportional to the degree to which the families differ. Thus, measures of variances among families obtained in the ANOVA are interpreted in terms of the covariances of members within those familes. To extend the simple designs which have already been discussed in the preceding chapter, consider an experiment in which both male and female parental identities are known and are experimentally structured so that each female is crossed to a different set of males. This is a hierarchal or nested design, designated as A/B by Cockerham (1963), in which each male services only one female, but each female (f_i) is served by several males $m_{j(i)}$. We may diagram the crossing scheme as:

Hierarchal (A/B) mating design

		Male trees									
		E	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*
	A	X	X	X							
Female	*B*				X	X	X				
trees	*C*							X	X	X	
	D										X . . .

The linear model for progeny trees assuming a completely randomized experimental design for r progenies of each mating

$(e_{k(ji)})$ is:

$$Y_{ijk} = \mu + f_i + m_{j(i)} + e_{k(ji)}$$

$$Y_{pqr} = \mu + f_p + m_{q(p)} + e_{r(qp)}$$

where $i, p = 1, 2 \ldots f,$

$j, q = 1, 2 \ldots m,$

$k, r = 1, 2 \ldots r.$

The ANOVA for this experiment, as found in texts including Steele and Torrie (1960), Cockerham (1963), and Becker (1967), is:

Source of variance	*df*	Mean square	Expected mean square
Females	$(f-1)$	$MS(F)$	$\sigma_e^2 + r\sigma_m^2 + rm\sigma_f^2$
Males/females	$f(m-1)$	$MS(M/F)$	$\sigma_e^2 + r\sigma_m^2$
Error	$fm(r-1)$	MSE	σ_e^2

Since there are three kinds of relationships among the seedlings, we can define covariances of these three relatives in terms of the variances. Full sibs exist when male and female parents are identical, and hence $i=p$ and $j=q$. Then for two individuals so related, Y_{ijk} and Y_{pqr}, their covariance is:

$$\begin{aligned}\text{Cov}\ (Y_{ijk}, Y_{pqr}) = {} & E[\mu + f_i + m_{j(i)} + e_{k(ji)}]\ [\mu + f_p + m_{q(p)} + e_{r(qp)}] \\ & - E[\mu + f_i + m_{j(i)} + e_{k(ji)}]\ [\mu + f_p + m_{q(p)} + e_{r(qp)}].\end{aligned}$$

We define each effect as a deviation around a mean so that the individual progeny effects are deviations from the full-sib family mean, the male effects are deviations around the female half-sib family mean, and the female effects are deviations around the general experimental mean. Therefore, for this covariance:

$$\begin{aligned}\text{Cov}(Y_{ijk}, Y_{pqr}) = {} & E^*(f_{i(i)}f_{(p)}) + E^*(f_i m_{q(p)}) + E\ (f_i e_{r(qp)}) \\ & + E^*\ (m_{j(i)}M_{q(p)}) + E^*(m_{j(i)}f_p) + E\,(m_{j(i)}e_{r(qp)}) \\ & + E^*\,(e_{k(ji)}e_{r(qp)}) + E\,(e_{k(ji)}f_p) + E(e_{k(ji)}m_{q(p)}).\end{aligned}$$

Since all effects are deviations, their expected values are zero, and we are concerned only with these nine elements. If we can assume that individual progeny trees are randomly assigned to experimental units and their deviations from their family means are not affected in any way by their male or female parentage, then all of the remaining expectations which contain cross products of $e_{r(qp)}$ or $e_{k(ji)}$ elements are also zero. The remaining five expectations (*) may be nonzero, depending on the manner in which the families were constructed, and are the components which generally determine the utility of the design. In this case, males do not serve as females and if the choice and assignment of males into single

female groups are made at random, then

$$E(f_i m_{q(p)}) = E(m_{j(i)} f_p) = 0.$$

Of the remaining three, consider the case that Y_{ijk} and Y_{pqr} are related as half-sibs and hence $i=p$, but $j \neq q$, and $k \neq r$. Then

$$E(f_i f_p) = E(f_i^2) = E(f_i - \bar{f})^2 = \sigma_f^2$$

and $$E(m_{j(i)} m_{q(p)}) = E(e_{k(ji)} e_{r(qp)}) = 0,$$

since males and progenies are not identical and are randomly chosen. Therefore, if Y_{ijk} and Y_{pqr} are half-sibs, the covariance between them is σ_f^2. Next, consider the second kind of relationship between two individuals which can exist in our experiment and allow them to be full-sibs. In this case, two distinct progeny trees have the same female ($i=p$) and male ($j=q$) parents but $k \neq r$. Then, of the three remaining expectations,

$$E(f_i f_p) = E(f_i^2) = \sigma_f^2,$$

$$E(m_{j(i)} m_{q(p)}) = E(m_{j(i)}^2) = \sigma_m^2,$$

and $$E(e_{k(ji)} e_{r(qp)}) = 0.$$

Therefore, the covariance between full-sibs $= \sigma_f^2 + \sigma_m^2$.

Finally, consider that the individual seedling's covariance with itself is taken. In this case, the same parentage exists ($i=p$, $j=q$) and the same individual deviation exists ($k=r$). Then the expectations of this covariance include

$$E(f_i f_p) = E(f_i^2) = \sigma_f^2,$$

$$E(m_{j(i)} m_{q(p)} = E(m_{j(i)}^2) = \sigma_m^2,$$

$$E(e_{k(ji)} e_{r(qp)}) = E(e_{k(ji)}^2) = \sigma_e^2.$$

Therefore, the covariance between an individual and itself $= \sigma_f^2 + \sigma_m^2 + \sigma_e^2$.

We have thus derived covariances among relatives in terms of the variance components estimated in the ANOVA of our experiment:

$$\text{Cov}(HS) = \sigma_f^2$$

$$\text{Cov}(FS) = \sigma_f^2 + \sigma_m^2$$

$$\text{Cov (individual)} = \sigma_f^2 + \sigma_m^2 + \sigma_e^2.$$

As developed in the preceding chapter, we can always determine the genetic variance contributions to each of the covariances among relatives by determining coancestries among the relatives. For our case, if we assume no inbreeding and no relatedness of the

parents, we derive the genetic variances of these covariances as:

$$\text{Cov}(HS) = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \ldots$$

$$\text{Cov}(FS) = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \ldots$$

$$\text{Cov}(\text{indiv}) = \sigma_A^2 + \sigma_D^2 + \ldots \sigma^2_{\text{environment}}$$

$$= \sigma_G^2 + \sigma_{\text{env}}^2.$$

Therefore, we can directly identify the design components in terms of the genetic variances they estimate as:

$$\sigma_f^2 = \text{Cov}(HS) = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \ldots$$

$$\sigma_m^2 = \text{Cov}(FS) - \sigma_f^2$$

$$= \text{Cov}(FS) - \text{Cov}(HS) = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \ldots$$

$$\sigma_e^2 = \text{Cov}(\text{indiv}) - \text{Cov}(FS) = \tfrac{1}{2}\sigma_A^2 + \tfrac{3}{4}\sigma_D^2 + \ldots + \sigma_{\text{env}}^2.$$

To summarize the analysis and interpretation of this design—variously called the hierarchal or nested mating design or the North Carolina Design I—we may list:

Source of variance	*df*	Mean square	Expected mean square
Females	$f-1$	*MSF*	$\sigma_e^2 + r\sigma_m^2 + rm\sigma_f^2$
Males/females	$f(m-1)$	*MSM*	$\sigma_e^2 + r\sigma_m^2$
Error	$mf(r-1)$	*MSE*	σ_e^2

where:

$$\sigma_f^2 = \text{Cov}(HS)$$

$$\sigma_m^2 = \text{Cov}(FS) - \text{Cov}(HS)$$

$$\sigma_e^2 = \text{Cov}(\text{indiv}) - \text{Cov}(FS)$$

$$= \sigma_{\text{env}}^2 + \sigma_G^2 - \text{Cov}(FS).$$

A second design commonly used to obtain similar kinds of genetic variance estimators is one in which the full factorial combinations of all males are crossed with all females. This arrangement has been termed the "factorial mating design," the "North Carolina Design II," or more rarely, a "diallel design" (Hanover and Barnes 1962). This mating scheme may be diagramed as:

		Male trees					
		E	*F*	*G*	*H*	*I*	*J*
Female trees	*A*	X	X	X	X	X	X
	B	X	X	X	X	X	X
	C	X	X	X	X	X	X
	D	X	X	X	X	X	X . . .

The linear models for two progeny trees in a completely randomized experiment are:

$$Y_{ijk}=\mu+m_i+f_j+mf_{ij}+e_{ijk}$$
$$Y_{pqr}+\mu+m_p+f_q+mf_{pq}+e_{npr}.$$

The analysis can be outlined as:

Source of variance	df	Mean square	Expected mean square
Females	$f-1$	MSF	$\sigma_e^2+r\sigma_{mf}^2+rm\sigma_f^2$
Males	$m-1$	MSM	$\sigma_e^2+r\sigma_{mf}^2+rf\sigma_m^2$
Males × females	$(m-1)(f-1)$	$MSMF$	$\sigma_e^2+r\sigma_{mf}^2$
Error	$mf(r-1)$	MSE	σ_e^2

$$\sigma_f^2 = \sigma_m^2 = \mathrm{Cov}(HS)$$
$$\sigma_{mf}^2 = \mathrm{Cov}(FS) - 2\ \mathrm{Cov}(HS)$$
$$\sigma_e^2 = \mathrm{Cov}(\mathrm{indiv}) - \mathrm{Cov}(FS)$$
$$= \sigma_{env}^2 + \sigma_G^2 - \mathrm{Cov}(FS).$$

We can derive the covariance of relatives in terms of the design components of variance much as was done for the nested design. If the experiment was properly conducted with respect to randomization of sampling, mating, and planting, and parental sexual identities were kept distinct, then only the $E(m_im_p)$, $E(f_jf_q)$, $E(mf_{ij}mf_{pq})$, and $E(e_{ijk}e_{pqr})$ can be nonzero. If Y_{ijk} and Y_{pqr} were half-sibs of the same male parent, $E(m_im_p)=E(m_i^2)=\sigma_m^2$, and all others are zero. If Y_{ijk} and Y_{pqr} were maternal half-sibs, $E(f_jf_q)=\sigma_f^2$, and all others are zero. If the two trees are different individuals of the same full-sib family, $E(m_im_p)=\sigma_m^2$, $E(f_jf_q)=\sigma_f^2$, and $E(mf_{ij}mf_{pq})=\sigma_{mf}^2$, and only the last component is zero. Then if the two trees were identical, the covariance includes σ_f^2, σ_m^2, σ_{mf}^2, and σ_e^2. From these identities,

$$\mathrm{Cov}(HS) = \sigma_f^2 = \sigma_m^2$$
$$\mathrm{Cov}(FS) = \sigma_f^2 + \sigma_m^2 + \sigma_{mf}^2$$
$$\mathrm{Cov}(\mathrm{indiv}) = \sigma_f^2 + \sigma_m^2 + \sigma_{mf}^2 + \sigma_e^2,$$

the tabulated expectations can be derived and the genetic variance contributions computed.

A third design, the diallel, (Griffing 1956) can also provide similar estimators when the choice of mating patterns is again limited to progenies from controlled crossing among parents which exist in a general, random-mating, unstructured population. Parents are assumed to be capable of functioning as both male and female and sometimes are capable of self-fertilization. The mating scheme for a modified diallel without selfs or reciprocals can be

diagramed as:

		Male trees					
		A	B	C	D	E	F
	A		X	X	X	X	X
	B			X	X	X	X
Female	C				X	X	X
trees	D					X	X
	E						X
	F						

The linear models for two progeny trees in a completely randomized design are virtually identical to the factorial mating design but in the more commonly used notation are:

$$Y_{ijh}=\mu+g_i+g_j+s_{ij}+e_{ijh}$$

$$Y_{pqr}=\mu+g_p+g_q+s_{pq}+e_{pqr}$$

in which g_i effects are general combining abilities or average performance deviations when the ith parent is crossed with a sample of the whole population, and the s_{ij} effects are specific combining abilities or the deviation of the cross of i by j parents from the expected average of the general combining abilities of the parents. While the mating design is itself unbalanced, the ANOVA for many diallel arrangements can be performed in such a way as to give clean mean squares for all of the components. Thus, for the general case in which there are s crosses for each of q parent trees serving as both male and female, the ANOVA is:

Source of variance	df	Mean square	Expected mean square
Among parents (gen. comb. ability)	$q-1$	$MSGCA$	$\sigma_e^2+r\sigma_s^2+\frac{r(q-2)s}{q-1}\sigma_g^2$
Interaction (spec. comb. ability)	$\frac{q(s-2)}{2}$	$MSSCA$	$\sigma_e^2+r\sigma_s^2$
Error	$\frac{qs(r-1)}{2}$	MSE	σ_e^2

$$\sigma_g^2=\mathrm{Cov}(HS)$$

$$\sigma_s^2=\mathrm{Cov}(FS)-2\ \mathrm{Cov}(HS)$$

$$\sigma_e^2=\sigma_{env}^2+\sigma_G^2-\mathrm{Cov}(FS).$$

The derivation of the identity between the covariance of relatives and the design components is similar to that of the factorial design. For half-sibs, the expected covariance contains only σ_g^2. For full-sibs, the expected covariance contains the σ_g^2 from both parents plus the σ_s^2 interaction component. For the same individual, the expected covariance contains $2\ \sigma_g^2+\sigma_s^2+\sigma_e^2$, and the in-

verse relations can then be derived.

Various partial diallel designs and blockings are given by Braaten (1965) with complete computational formulas and expectations.

In all of these designs, simple modifications of plot structure can give additional data on replication, plot error, and within-plot variances. If r randomized complete blocks are used and the analyses are performed on plot means, the following changes have to be made in the analyses:

(1) To the ANOVA add the lines "replications" and "$r-1$" under the columns headed "source of variance" and "*df*."

(2) Change the error *df* to $(mf-1)(r-1)$ for the nested and factorial designs, and $\frac{(qs-2)}{2}(r-1)$ for the diallel design.

Otherwise, no changes are required. If samples of the variance among trees within plots are obtained, then the composition of σ_e^2 can be broken down into components of between- and within-plot error variances:

$$(\sigma_e^2)^* = \frac{\sigma_w^2}{k} + \sigma_p^2,$$

where k is the harmonic mean of numbers of trees per plot. If the analyses were done on plot sums,

$$(\sigma_e^2)^* = \sigma_w^2 + k\sigma_p^2,$$

and the coefficients for the other variance components require multiplication by k. In either case, σ_p^2 now measures a plot mean error variance and $\sigma_w^2+\sigma_p^2$ has the same composition as we formerly composed for σ_e^2. Thus,

$$\sigma_p^2 = \frac{\sigma_e^2}{k} \text{ and } \sigma_w^2 = \frac{k-1}{k}\sigma_e^2.$$

The genetic components of within-family variance are thus shared, $1/k$ in the σ_p^2 part, and $\frac{k-1}{k}$ in the σ_w^2 part. Since many tree experiments are planned with multiple-tree plots and the usual experience is for some measurements and trees to be missing, the use of plot mean analyses for traits not affected by spacing is common (Stonecypher 1966). For traits which are affected by differential mortality or density, adjustments of the data for the spacing effects should be made before analysis. If plot means are analyzed, the usual procedure is to sample several plots to estimate the variance among trees within plots (σ_w^2), and by computing k and σ_e^2, to then fully determine σ_w^2 and σ_p^2.

All of these designs may also be augmented by the inclusion of reciprocal crosses and selfs and the extension of the linear

models to account for these effects. Two types of effects are often defined in terms of contrasts between performance of trees when serving as male versus female parent. A general maternal effect is defined as the contrast or difference which exists when the tree is crossed with a sample of the whole population, first as a male and second as a female parent. A specific reciprocal effect is then defined as the contrast between full-sib families which differ only by the sexual order of the parents. These augmentations are often carried in diallel experiments but can be used with any other mating design as long as the parental genotypes can function both ways.

The use of subblocking to reduce the error variation within complete replications is often recommended for estimating means and may find considerable use in forestry (Snyder 1966). However, for estimating variances, blocking will not generally be useful unless very large errors within replications are otherwise unavoidable.

The four mating designs are diagramed in figure 15.

The above designs are the ones most immediately useful in forestry since we generally start with a presumably unstructured population and we wish to estimate as many genetic variances as possible. For further details of the analyses, the reader is referred to the surveys available in Cockerham (1963), Gardner (1963), or Becker (1967). For the analysis of the nested and factorial designs, see Comstock and Robinson (1948), and for the diallel see Griffing (1956), and for partial and blocked partial diallels, see Braaten (1965). If balanced experiments are created, the different designs can be easily compared with respect to the best allocations of crossing efforts among numbers of males, females, and numbers of crosses per parent as well as the general goodness of the designs over different levels of the genetic components. In balanced experiments, the sums of squares computed by standard methods are independent.

In all of these cases, the Cov(HS) estimator is used to estimate $\sigma_A^2/4$ since it has no dominance variance. If no epistatic variances are present, or are ignored, the estimator is simply derived. The Cov (FS) which has both σ_A^2 and σ_D^2 is used to estimate σ_D^2 after any adjustments for the σ_A^2 are made as may be necessary.

If the family genetic variations are generated by multinomial distributions, their means may often approximate a normal distribution and hence the variance of mean squares would approximately follow a $X^2\sigma^2$ distribution. For traits with few genes operating and with few individuals per family, the approximation holds less well. For our assumed quantitative traits, however, the normality assumptions will be closely approximated, especially if means are used and skew corrections applied.

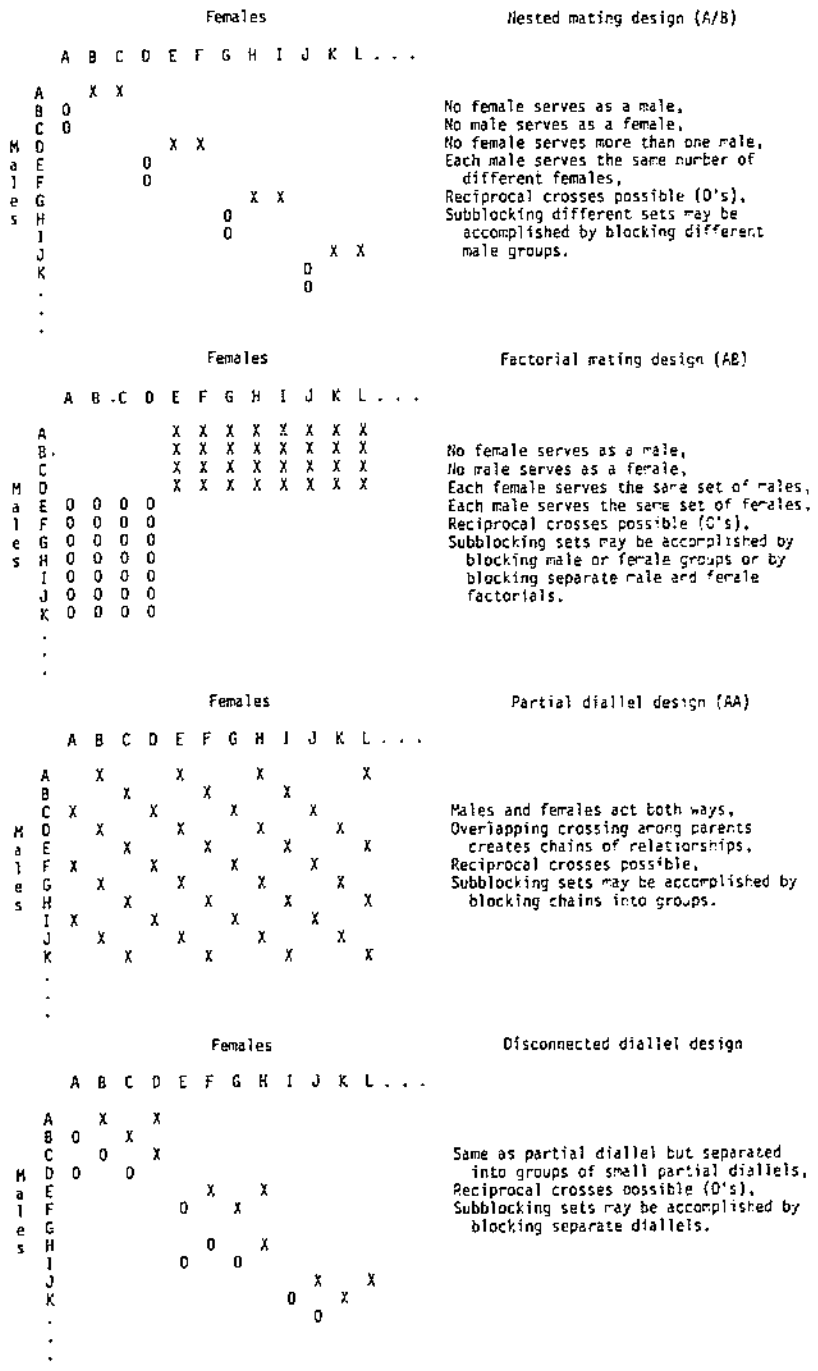


Figure 15.—Schematic diagram of four mating designs.

ERRORS OF ESTIMATING GENETIC VARIANCES

Using the $X^2\sigma^2$ distributions of the sums of squares for the nested and factorial designs, and the variance of the sums of squares of the diallel as traceable by its individual squared sums, the variances of the additive genetic variance can be derived. Allowing $\hat{\sigma}_A^2 = 4\hat{\sigma}_f^2$ for the nested design, and using a pooled estimate of $4\sigma_f^2$ and $4\sigma_m^2$ to estimate σ_A^2 in the factorial, and $4\hat{\sigma}_g^2 = \hat{\sigma}_A^2$, the variances are:

Nested design $$V(\hat{\sigma}_A^2) = \frac{32}{r^2m^2} \times \left[\frac{MSF^2}{f-1} + \frac{MS(M/F)^2}{f(m-1)} \right]$$

Factorial design $$V(\hat{\sigma}_A^2) = \frac{32}{r^2[f(m-1)+(f-1)]^2}$$

$$\left[MSF^2(f-1) + MSM^2(m-1) + \frac{(m+f-2)^2}{(m-1)(f-1)} MSMF^2 \right]$$

Partial diallel design $$V(\hat{\sigma}_A^2) = \frac{32}{r^2} \frac{MSSCA^2(q-1)^2}{(q-2)^2s^2}$$

$$\left[\frac{2}{q(s-2)} + \frac{1}{q-1} + \frac{2rs(q-2)}{(q-1)^2MSSCA} \hat{\sigma}_g^2 + \frac{[q+(q-4)s]sr^2}{(q-1)^2MSSCA^2} \hat{\sigma}_g^4 \right]$$

A pooling of sums of squares with different expectations produces a non-$X^2\sigma^2$ variate. The variances are computed as non-X^2 variances.

Similarly, the dominance genetic variances can be estimated by a linear function of the sums of squares and those variances can be estimated. Only in case of the nested design are we required to use more than two sums of squares, since that estimator contains both σ_A^2 and σ_D^2 and hence requires that we estimate σ_A^2 separately. These sampling variances are:

Nested design $$V(\hat{\sigma}_D^2) = \frac{32}{r^2} \times$$

$$\left[\frac{(1+m^2)MSM^2}{m^2f(m-1)} + \frac{MSF^2}{m^2(f-1)} + \frac{MSE^2}{(r-1)(mf-1)} \right]$$

Factorial design $$V(\hat{\sigma}_D^2) = \frac{32}{r^2} \left[\frac{MSMF^2}{(m-1)(f-1)} + \frac{MSE^2}{(r-1)(f-1)} \right]$$

$$V(\hat{\sigma}_D^2) = \frac{64}{r^2} \left[\frac{MSSCA^2}{q(s-2)} + \frac{MSE^2}{(r-1)(qs-2)} \right]$$

Partial diallel design $$V(\hat{\sigma}_D^2) = \frac{64}{r^2} \left[\frac{MSSCA^2}{q(s-2)} + \frac{MSE^2}{(r-1)(qs-2)} \right]$$

Finally, the σ_e^2 is estimated directly from the error mean square and its sampling variance is simply the variance of that mean square. However,if the error is partitioned into o_{tc}^2, and o_p^2, or

other components, other appropriate mean square functions would have to be derived.

It is clear that the variances of the estimated components vary according to the size of the genetic components themselves, as well as how the experimental effort is allocated among numbers of parents, numbers of crosses, crossing patterns, and numbers of trees per plot. While there would be considerable variation in actual cost efficiencies according to species, ease of making crosses and experimental plantings, and how an agency can schedule such activities, it is instructive to compare sampling errors of experiments with the same total number of crosses.

For 100 crosses with a randomized block planting design of 8 trees per plot and 10 replications, an analysis of variance with estimates of σ_A^2 can be reconstructed. For any given experiment, the error in estimating σ_A^2 is also a multiple of the environmental sources of variation (σ_{env}^2 above), and allowing this to equal 1 puts the comparisons of equal-sized experiments on the same basis. When $\sigma_A^2=1$, for example, the use of 25 females and 4 males per female in a nested design yields a variance of the estimates of only 0.14. However, if the 100 crosses were made by using say 4 females and 25 males per female, the error would be 0.71. At lower levels of σ_A^2, the female variance and its mean square are lower, and the error variance for estimating σ_A^2 is therefore also lower. The higher the σ_A^2 is, the larger also is its error of estimation. The relationship of the size of the component to its error of estimation is thus roughly constant over wide ranges of σ_A^2, and any design that is good at moderate heritabilities of around 0.5 is generally good between heritabilities of 0.2 to 0.9. Only at relatively low values of σ_A^2 do the relative efficiencies of the allocation of materials change very much. If we were to chart the variance of estimates of σ_A^2, $V(\sigma_A^2)$, as a ratio of σ_A^2 itself for all possible values of heritability ($\sigma_A^2/(\sigma_A^2+\sigma_E^2)$), the curves for the 4 males and 25 females per male nested design are shown in figure 16. The curve for the allocation of 25 males and 4 females per male would generally be lower, except at low heritabilities, when the curves cross (fig. 17). At low heritabilities, the error and female mean squares are more nearly equal to the male mean square and, since they influence the error for estimating σ_A^2, also require precision in estimation. Since precision requires high degrees of freedom, relatively more full-sib families than half-sib families have to be sampled.

While the error variance for estimating σ_A^2 tends to decrease rapidly when σ_A^2 itself decreases at very low values of σ_A^2, the curves display a curious reversal of direction. For any mating design, there is a point in heritability below which the decline in $V(\sigma_A^2)$ is more than matched by the decline in σ_A^2, and hence the curve rises for lower heritability. If we wished to design an experiment which would be satisfactory for any level of σ_A^2 and we

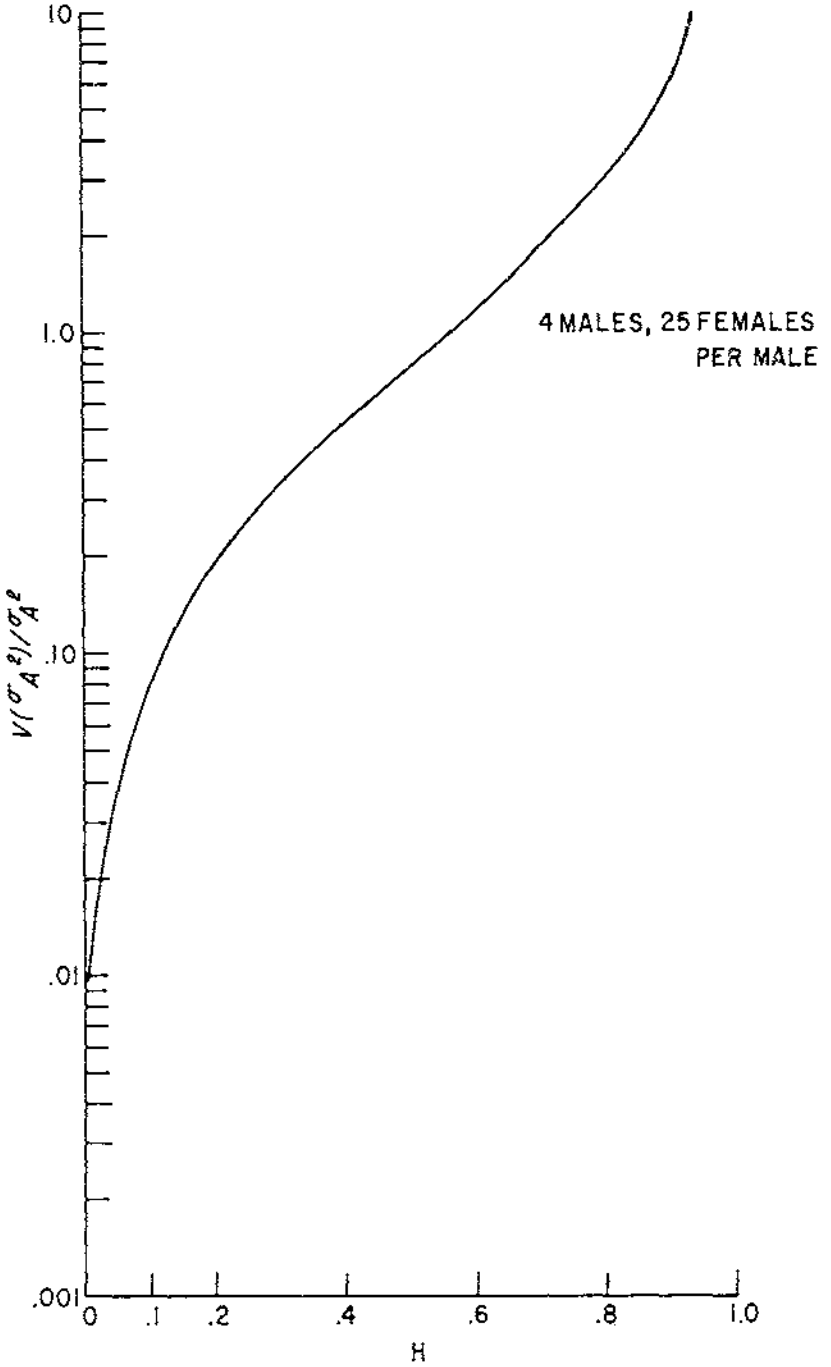


Figure 16.—Efficiency of allocating 4 males and 25 females per male in a nested mating design over all heritability levels.

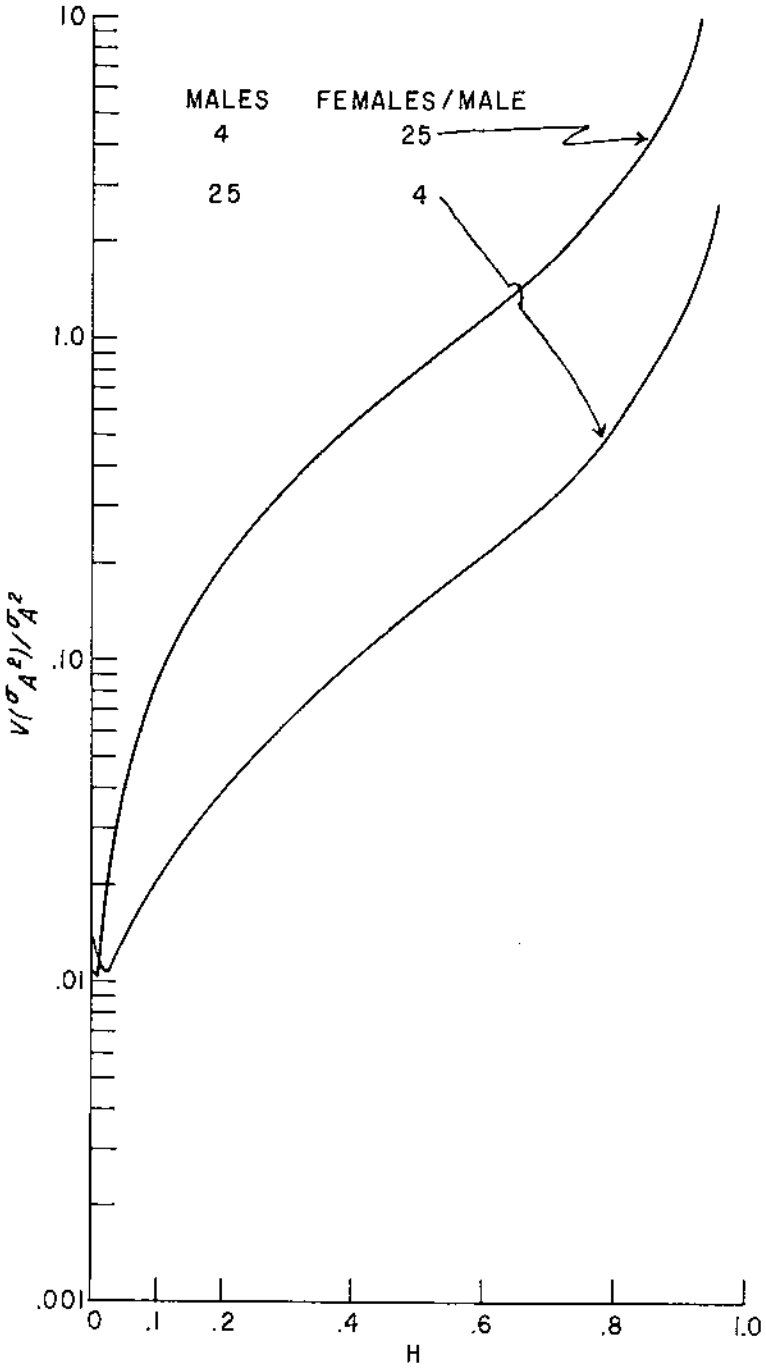


Figure 17.—Comparative efficiency of allocating 4 males and 25 females per male versus 25 males and 4 females per male in a nested mating design.

use a criterion of satisfaction like the coefficient of variation

$$\left(\mathrm{CV}=\frac{\sqrt{V(\sigma_A^2)}}{\sigma_A^2}\right),$$

the above change in direction implies that no design can be completely satisfactory. Bogyo (1964) has shown that astronomically larger experiments would be necessary if CV=0.05 is uniformly required for low σ_A^2. However, other criteria of satisfaction may serve as well at low σ_A^2. One suggestion is that at low σ_A^2 our concern is to not estimate a high σ_A^2, and hence if $\sigma_A^2=0.001$, or 0.01, or 0.02, a standard error of estimate of 0.05 may be quite acceptable. Therefore a reasonable criterion to suggest is that at heritabilities above 0.05, a CV=0.05 or less be used, and below 0.05, a constant variance be used as a maximum critical curve. Imposing these criteria on the above comparison of male:female allocations show that the 25 males and 4 females/male are satisfactory everywhere, while the other is only satisfactory at low σ_A^2 (fig. 18).

Other designs and allocations may be compared. Among the better allocations of the nested design (A/B), 16 males and 6 females/male and 50 males and 2 females/male can be compared with the factorial design (AB) with allocations of 6 males and 16 females or 10 males and 10 females, and compared with the diallel (AA) with q=83 parents and $s=6$ crosses per parent or with $q=66$ parents and $s=3$ crosses per parent (fig. 19).

Similar analyses and interpretations of the estimates and variances on σ_D^2 can also be made (fig. 20). For these estimates the full-sib covariance estimator is most critical. All designs require good estimations of the error component, which is largely a function of replication numbers for any constant number of crosses. Thus, the choice of parental allocation affects error variance only if the crossing pattern is so costly or time consuming or, conversely, is so cheap and easy as to affect the number of replications which can be planted. The nested design is the only one requiring an estimate of σ_A^2 to estimate σ_D^2. Hence, it is affected directly by the precision of estimating the half-sib variance.

It can again be observed that as σ_D^2 rises, the $V(\sigma_D^2)$ also rises, and that there exist allocations of parents and half-sib versus full-sib family members which are reasonably good over a wide range in σ_D^2. The actual choice of design will again depend on relative operational costs and specific criteria of goodness, but there still seems to be evidence that some design allocations can find wide favor. Unfortunately, the optimum design for minimizing the errors of estimating σ_A^2 are not the same as for estimating σ_D^2. Therefore, optimum levels of allocation for estimating σ_A^2 and σ_D^2 will be in some conflict.

Several other judgments must be made before selecting a design. These involve such problems as the use of several designs simultaneously for the multiplicity of purposes usually intended for any experiment, and the effects of inevitable plot mortality or

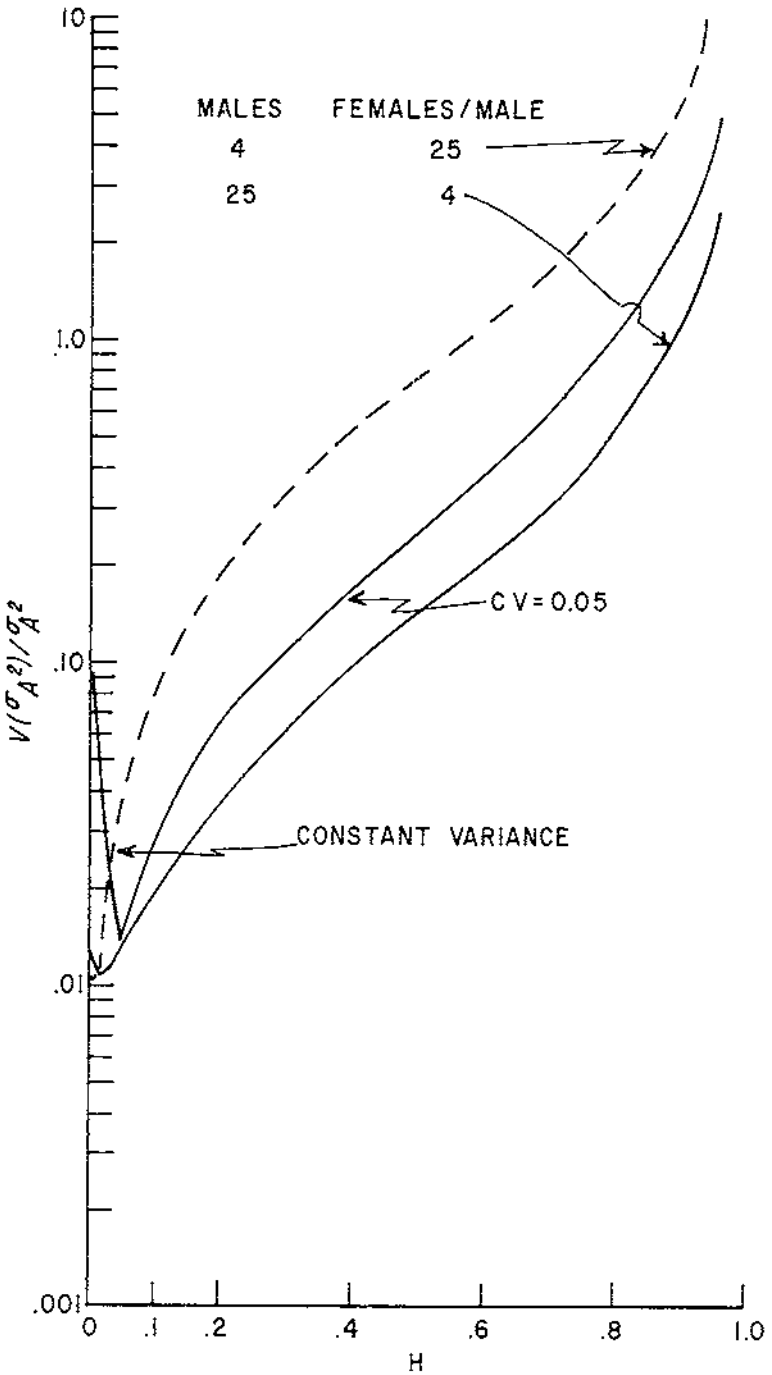


Figure 18.—Efficiency of allocating 4 males and 25 females per male, and 25 males and 4 females per male, relative to a coefficient of variation (*CV*) of 0.05 and a constant variance at low heritability.

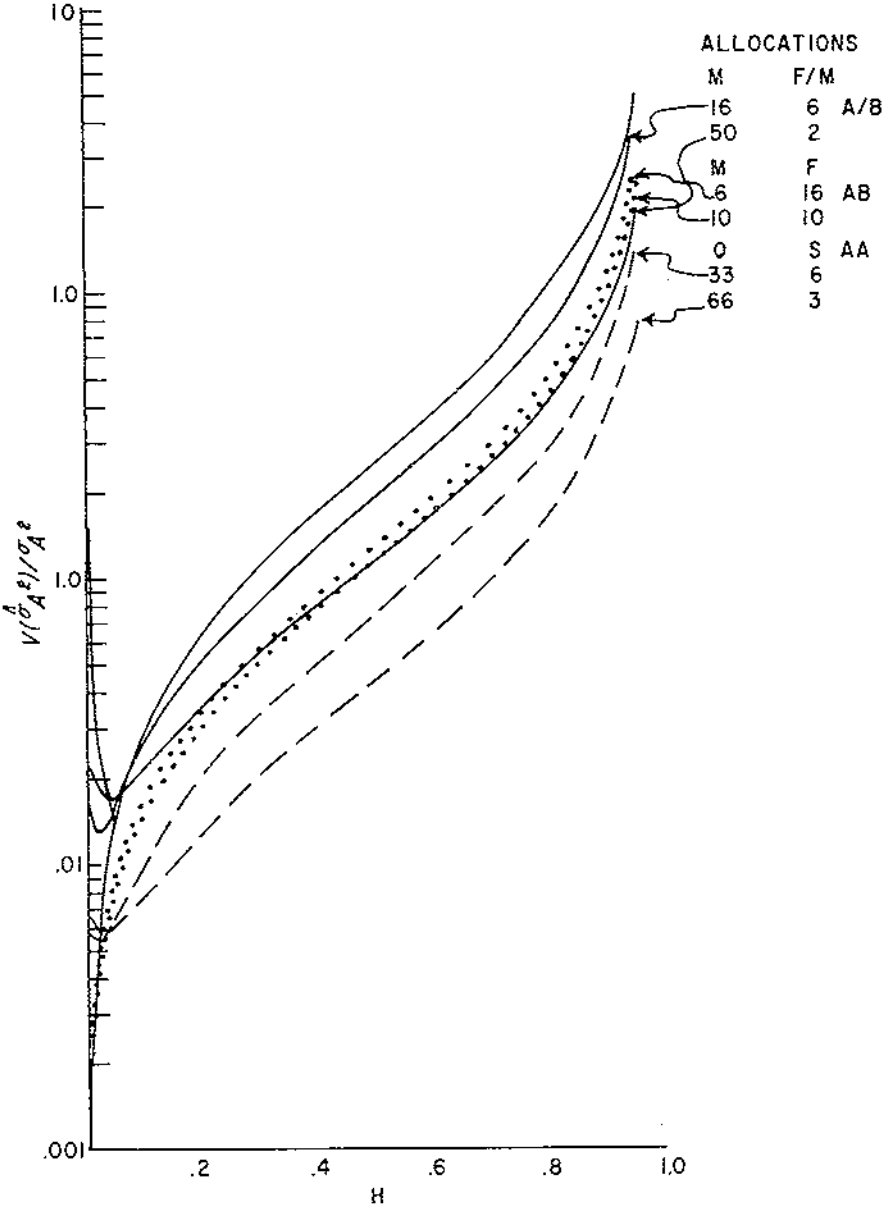


Figure 19.—Relative design efficiencies $(V(\sigma_A^2)/\sigma_A^2)$ for estimating σ_A^2 when $\sigma_B^2=\sigma_A^2$, $c=100$.

missing crosses. The problem of missing trees in plots has already been discussed, but the problem of missing plots involves many analytically significant decisions. We have seen that missing plots will inevitably cause some difficulty in determining the distribution and hence the sampling variance of the sums of squares esti-

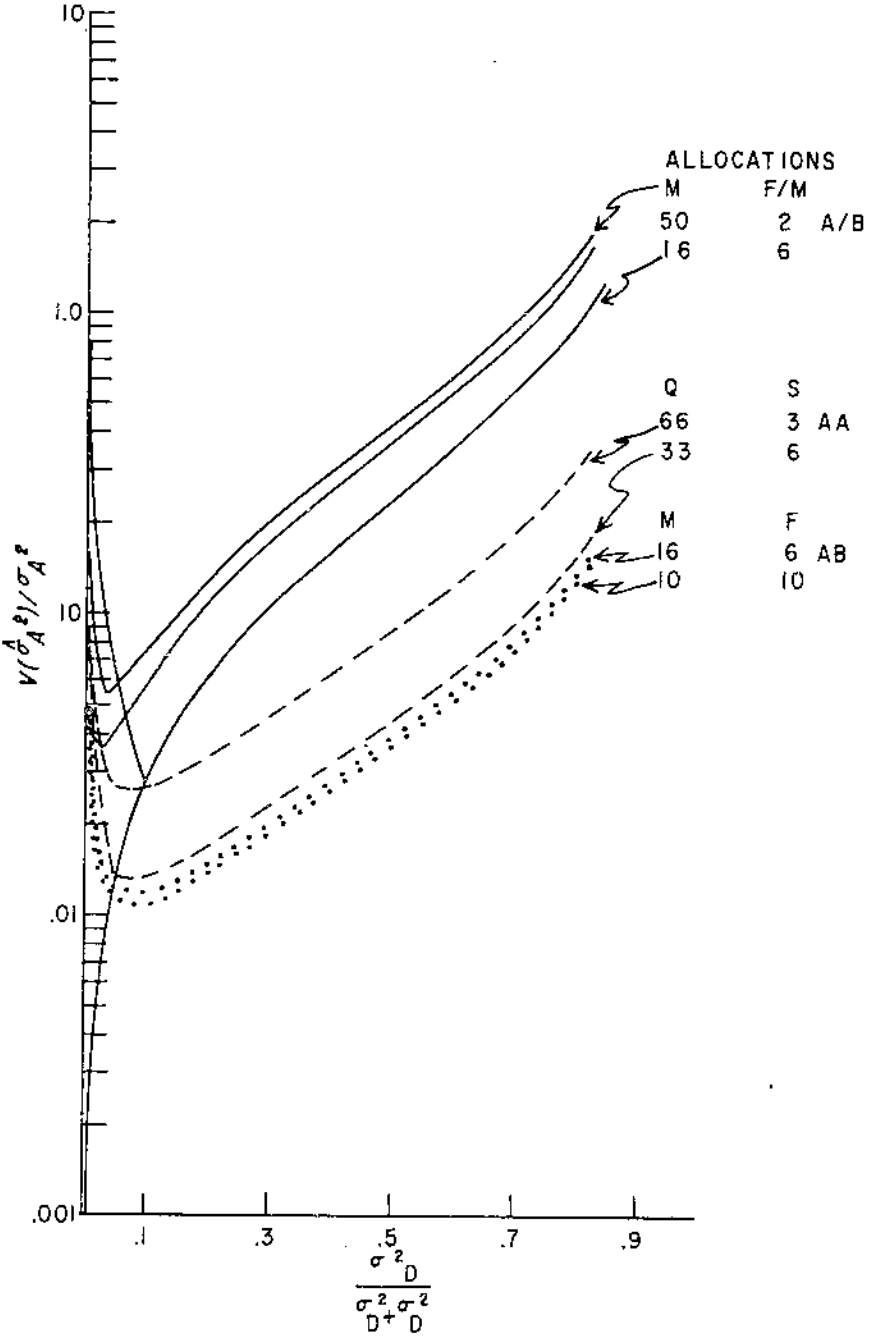


Figure 20.—Relative design efficiencies $(V(\sigma_D^2)/\sigma_D^2)$ for estimating σ_D^2 when $\sigma_D^2=\sigma_A^2$, $c=100$.

mators. Some crossing programs will be more likely than others to yield unbalance by their difficulty of execution. Regardless of the method used to obtain sums of squares as previously described, the variance components, and hence the genetic variances, can be derived as linear functions of various sums of squares, some with higher error than others. However, the variances of linear functions of elements can also be derived if the elements have variances. This procedure requires that the covariances among the sums of squares in unbalanced experiments must also be accounted for. As long as large computers are available, most analyses and estimates of sampling errors can be derived or other quadratic estimators derived. For agencies without such facilities, it might be desirable to choose designs which are relatively unaffected by missing plots. Where computing facilities are lacking, the diallel cannot be generally recommended since both the nested and factorial designs can be more easily handled for any missing plots. With computing facilities, however, other design criteria can be more significant.

Designs must often be chosen on the basis of only partly known data because the forest geneticist must often deal with unknown difficulties, costs, times for making crosses, and actual levels of the variances. Each species that he deals with has different pollination and seedling production problems and exhibits different error and genetic variances for each trait. For any one species, the geneticist may want to estimate means and several components for several traits from a reasonable population sample. Thus, given the capability for estimating variances and means, but with multiple demands for estimators, either some compromises on efficiency or several different experiments will have to be run.

OTHER ESTIMATORS

Several sources of data on genetic variances can often be obtained or planned to jointly provide estimators of the components. Thus, several other kinds of relatives and estimators, such as parent-offspring regressions, clonal variances, and eventually, various kinds of cousin relatives, may be made available. It is not necessary that we use only the three mean squares from the above mating design ANOVA's to estimate three components, since other kinds of estimators are also available. Perhaps the simplest estimator requiring no relationships at all is that developed by Shrikhande (1957) and used for forest trees by Sakai and Hatakeyama (1963). In this method, variances due to systematic soil gradients and to a random source of variation are estimated. The random variation is assumed to include genetic and some environmental covariances and is taken as an estimate of the total genetic variance. While some bias and large sampling errors exist (Namkoong and Squillace 1970), some estimates of a total genetic variance have been derived with considerable economy.

More commonly used and only slightly more complicated are

estimates derived from parent-offspring regressions. The relationship structure we exploit in this case involves only two kinds of relatives—parent and offspring or unrelated. The sib structures we first studied involve the three relationships; full-sib, half-sib, and unrelated. Shrikhande's (1957) structure involves no relationships. Whenever any progeny and their parents can be measured, the regression can be interpreted genetically in terms of the covariance between parent and offspring and the variance of either or both sets. As long as the parents are noninbred, unrelated, and randomly mated, the covariance of parent and offspring is $\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_{AA}^2+\ldots$, as derived in the previous chapter. When both parents are known and their average is taken for a midparent value, its covariance with the offspring is the same. Similarly, if a random set of trees is mated to a common parent, the parent-offspring covariance within common parent sets also has the same genetic variance expectation. It often occurs that the progeny trees are not truly planted at random but rather in some plot and replication design which may not be at all similar to the environments sampled by the parents. Nevertheless, progeny means can be related to parental values and hence, covariances, parent and progeny variances, and the regressions and correlations of parent and offspring can all be estimated by simple standard procedures (Becker 1967).

The numerator of the regression coefficient is simply the covariance between two variables where one is a measure of potential performance, and the other is the progeny performance. The denominator of the regression is the variance of the independent variable. The parent is the independent variable if the parent is known and the progeny performance is to be predicted or if selection on the basis of parental phenotypes is made and the expected response in the progeny is desired. The offspring can be the independent variate if they are known and measured and the response of relatives such as parents or other sibs is to be predicted.

In forest genetics experiments, it is not always clear whether the parent or offspring is considered to be the independent variate. In some genetics experiments, it seems that the parent is assumed to be the independent variate, and the variance among parents is the denominator of the regression of offspring means on parental values. The use of midparent values reduces this parental variance by one-half. Since parental genotypes may be more immediately useful in forestry, and reliable data from plantings in more representative environments are on offspring, the offspring data can frequently form the independent variable. The regression estimates will depend on the variance of offspring means that are often derived from some replicated design and hence can be some function of several design variance components. Hence, the variance of the estimated means can be quite different according to whether the parent or offspring means are being estimated.

A very useful feature of the parent-offspring covariance is that we can also determine the covariance between juvenile and mature characteristics. This parameter is necessary for predicting the gain to be made in mature growth performance on the basis of juvenile tests or observations. If juvenile performance is desired, as for early survival and competition, and juvenile measures are taken, then the additive genetic covariance between the measured trees and estimated breeding values is $r\sigma_A^2{}_{(j)}$, where r is the coefficient of relationship between the measured trees and the trees whose value is being estimated, and $\sigma_A^2{}_{(j)}$ is the additive genetic variance of the juvenile trees. Similarly, if mature performance is being estimated by mature relatives, the additive genetic covariance is $r\sigma_A^2{}_{(m)}$. But if one estimates mature performance from juvenile relatives, or juvenile performance from mature relatives, the covariance is r Cov $[A(j),A(m)]$ the product of the r and additive genetic covariance between the juvenile and mature traits. If r is known, then the covariance is estimable and the efficacy of selecting on the basis of correlated traits can be determined. The covariance, in this case, is not a genetic variance and should not be used as such to estimate heritabilities unless the same trait is being measured in the relatives or unless one wishes to obtain a lower bound estimate for the larger genetic variance.

When the covariance actually estimates a genetic variance, then the variance of either single trees or replicated plots can be used as the variance denominator of a regression coefficient. It sometimes occurs that age or environmental effects cause large differences in average performances which might induce scale differences such that the variances are not comparable in the different materials. In such cases, it is a common procedure to standardize the units of measure by dividing the x variable by σ_x and the y by σ_y[9] (Frey and Horner 1957). It is also sometimes done by scaling one of the variables (say x) to the other (say y) by multiplying it by σ_y/σ_x. Then, the variances are comparable and the covariance is multiplied by σ_y/σ_x. When such rescaling is justified and if the variance denominator of b is σ_y^2, the regression can be rescaled and becomes the correlation coefficient. If the variance denominator is σ_x^2, then the regression should be rescaled to have a multiplier of σ_y/σ_x^2. Rescaling should be done with caution, and the denominator for the regression should be selected for the specific purposes for which the regression is to be used.

The sampling error of the parent-offspring covariance estimator is

$$\frac{1}{df}\left[\text{Cov}^2 + V(\text{parents}) \times V(\text{offspring})\right],$$

which can clearly be quite cheaply reduced by simply increasing the number of parent-offspring pairs. If the genetic variance is estimated by doubling the parent-offspring covariance, the sam-

pling variance is increased by a factor of 4.

It is often desired to also estimate the parent-offspring regression coefficient, which can be computed as in any other regression problem with the same sampling error as in any other regression problem. The great attraction of computing the regression is that it can be interpreted in terms of a ratio of a portion of the genetic variance to the total variance among the parents or offspring. Any such function which is a ratio of some portion of the genetic variance divided by some function of genetic and environmental variances is called a heritability. Clearly, the portion of the genetic variance in the numerator and the content and structure of the total variance in the denominator determine the meaning and value of heritability. While a full discussion of different forms of heritability is included in chapter 3, we can briefly note that any procedure which involves estimating the numerator and denominator separately requires us to estimate its variance as the variance of a ratio of two random variables. Thus, the simpler estimate of heritability as a regression of parent and offspring is not only very simple to compute, but its errors are easily reducible and the error distribution is well known. By using a regression estimator, it is also possible to bias the sample of the independent variate by selecting extreme elements to reduce the error of estimate on the regression. While this does not provide valid estimates of either the genetic variance of the population in the numerator or the phenotypic variance of the population in the denominator, it does provide an unbiased estimate of the regression with small error (Hill 1971).

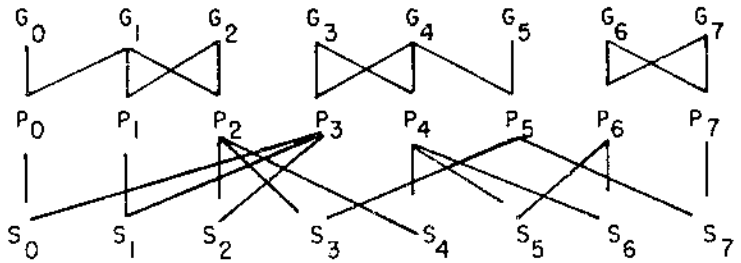
HIGHER-ORDER RELATIVES*

Thus, simpler designs that involve fewer different kinds of family relationships than the ANOVA full-sib, half-sib, and no relationship can give estimates of one or two genetic variances. The mating designs among unrelated parents provide three variance components for estimating the myriad of genetic variances, and it is clear that the more different kinds of relationship constructed, the more different kinds of genetic variances can be estimated. As long as each variance component, such as the half-sib covariance, is a different, independent, linear function of the several genetic variances, including the different epistatic types, then the additional design components give us estimates of more genetic components. Since full-sib covariances contain both additive and dominance genetic variances, and half-sib covariances do not contain dominance variances, they allow us to separate additive from dominance variances. However, they still contain various additive types of epistatic variances which cannot be separated without additional kinds of relationships.

As trees of known parental origin mature, the opportunity for creating cousin types of relationships also emerges. Since grand-

*Graduate-level statistical training required for thorough understanding.

parental identities are now known, individuals which have common grandparents can be created and identified. For example, in the following diagram,



if all G_i are unrelated, all P and S individuals are not inbred but do have a variety of relationships. Within the P generation, P_1 and P_2 are full-sibs, as are the pairs P_3 and P_4, and P_6 and P_7. P_0 is a half-sib of P_1 and P_2, and P_5 is a half-sib of both P_3 and P_4. Within the S generation, S_5 and S_6 are full-sibs and S_4 and S_8 are half-sibs. First-cousin relations exist for S_2 and S_5, while S_1 and S_4 are double-first cousins since both sets of grandparents are common. S_2 and S_4 appear to also be double-first cousins but additionally have a common P_2 parent and hence may alternatively be considered as half-sibs with the alternate parents related as full-sibs. A slightly weaker relationship of S_2 and S_3 exists since the alternate parents are related as half-sibs. An even weaker relation exists between S_1 and S_3 since their P_1 and P_2 parents are full-sibs while their P_3 and P_5 parents are half-sibs. Weaker still is the relation between S_0 and S_3 since their P_0 and P_2 parents are half-sibs and their P_3 and P_5 parents also are half-sibs.

If it is also reasonable to assume that the gene frequencies and genetic variances do not change over the generations, then a series of parent-offspring, grandparent-offspring, aunt-niece, and aunt-half-niece types of relations can also be identified. Therefore, a very much expanded set of relationships can yield many new equations for genetic variances among relatives and hence allow for estimating variances due to a variety of inherited effects. Eisen (1967) lists most of the above cousin types of relations and describes designs to estimate additional genetic and nongenetic variances.

It is also possible that most epistatic variances can safely be assumed to be negligible components, and hence that the additional information on relatives could be used to do a better job of estimating a restricted set of genetic variances. When the number of independent equations relating to the experimental design components of variance to the genetic components of covariance equals the number of genetic components, then one can directly derive estimates of the genetic components.

With a set of mean squares, covariances, or other estimators ($\underline{v}$), the design components of variance $\underline{d}$ can be estimated by

$\underline{v}=C\underline{d}$, where C is a square matrix of coefficients and $\underline{d}$ is the vector of design components. Since the design components can also be related to an equal number of genetic components, $\underline{d}=D\underline{g}$, where D is a square matrix of coefficients and $\underline{g}$ is the vector of genetic variance components. Then $\underline{v}=CD\underline{g}$, or letting $M=CD$, then $\underline{v}=M\underline{g}$.

If there are more genetic components than estimating functions of the design, we can only estimate combinations of the genetic components, for example, Cov (half-sibs) $=\frac{1}{4}\sigma_A{}^2+\frac{1}{16}\sigma_{AA}{}^3+\ldots$. Otherwise, we must reduce the model and estimate only those components we wish to assume to really exist, for example, assume $\sigma_{AA}{}^2=0$.

If there are fewer genetic components than estimating functions, we can choose a method to provide a good estimator of each component. A simple method would simply average any independent estimates, but this would give equal weighting to all estimates weak and strong and might not use all the information in the data. A logical procedure is to consider the mean squares, covariances, and variances as linear functions of the genetic variances which are essentially constant, regression-like coefficients. Then, each computed mean square or variance would have an expected, unique combination of the genetic and environmental variances plus some component of error variation. This model is essentially that of a linear regression in which the dependent variate is the mean square, the constant regression coefficients are the genetic and environmental variances, the independent variables are the set of coefficients which determine the contribution of each genetic component to the mean square, and the error term is the deviation or variation of the actually computed mean square from its expectation.

As with any such regression problem, the least squares estimation for the regression coefficients (genetic and environmental components) can be derived. For the above model we can write:

$$\underline{v}=M\underline{g}+\underline{e}$$

when $\underline{v}$ = vector of mean squares, covariances, etc.,

$\underline{g}$ = vector of genetic and environmental variance components,

M = matrix of coefficients relating the expected value of $\underline{v}$ to its $\underline{g}$ components,

$\underline{e}$ = vector of errors around each mean square.

An unbiased and unweighted estimate of $\underline{g}$ then is:

$$\hat{\underline{g}}=(M'M)^{-1}M'\underline{v}.$$

When more estimators than $\hat{\underline{g}}$ components exist, Nasoetion

(1965) has shown that it is better to estimate the $\hat{\underline{g}}$ directly from the $\underline{v}$ rather than to first estimate the experimental design components from the $\underline{v}$ and then to estimate the $\hat{\underline{g}}$.

However, each of the mean squares, variances, and covariances in $\underline{v}$ usually has different errors and some may, in fact, be correlated. The complete weighted least squares solution would then require weighting by V, the variance-covariance matrix of the $\underline{v}$ vector, and the weighted least squares estimate would be:

$$\hat{\underline{g}} = (M'V^{-1}M)^{-1}M'V^{-1}\underline{v}.$$

An additional problem created by the use of the V matrix is that we must now estimate its elements, and the problem is that the best estimates of those variances and covariances require estimates of $\hat{\underline{g}}$. That is, the genetic and environmental components $\underline{g}$ are needed to provide good estimates of V, and the V is needed to provide good estimates of $\underline{g}$. Hayman (1960a) therefore recommends an iterative procedure whereby initial estimates of $\underline{g}$ are used to provide initial estimates of $\underline{v}$, which are then used according to procedures formulated in chapter 7 to estimate $\underline{g}$, etc. This iteration is continued until the $\underline{g}$ does not change significantly from one trial to the next.

It is reasonable at this point to question the utility of such procedures or of generating large and complicated experiments. Is the additional precision worth the effort, or would it be wiser to simply create small two-factor designs in each subsequent generation? Is it necessary to estimate the contribution of epistatic components? When no estimates are available, the value of extra information is much higher than when some estimates and experience have accumulated. When data are scarce and there are few designed experiments, then any kind of relative can aid precision considerably. Therefore, in the beginning stages of programs, the use of a variety of estimators to estimate many components is reasonable. As populations develop, however, primary concern will be devoted to changes in the genetic components and more attention then may be given to fewer and simpler designs. Therefore, in these times of newly developing forestry programs, the information from sib designs, parental and clonal variances and covariances with offspring, and eventually grandparental, cousin, and nephew types of relatives, will be useful. If the time ever comes that foresters can consider alternate mating designs for estimation purposes, then the question of allocating resources among the multiplicity of relatives will require optimization. It seems clear, however, that those relatives with large contrasts in the contribution of the several genetic variances will generally be favored.

Considering the array of relatives which might be generated and the great proliferation of plots and consequent unbalance in allocating degrees of freedom, some consideration should be given to unbalanced designs with a more equal distribution of degrees of freedom among the various kinds of relatives and sources of variance. While a set of designs has not been specifically developed for genetic experiments, the utility of unbalanced designs for variance component estimation is established for industrial experiments (Goldsmith and Gaylor 1970). Since most forestry experiments become unbalanced by mortality or pollination failures anyway, deliberately designed unbalance may not cause any extra work and may afford experimental economies.

For estimating variances, it is therefore desirable to consider designs which do not necessarily satisfy the requirements of testing treatment means very well but which do estimate variances efficiently even if unplanned imbalances occur. For example, Goldsmith and Gaylor (1970) examine several combinations of nested designs, which could easily be implemented in forest genetic experiments. For their particular restrictions, they find that the balanced design is good for low heritabilities if one wishes to minimize the unweighted sum of error in estimating the three design components. However, for higher heritabilities, unbalanced sets can be optimal. Estimation of genetic components has not been thoroughly examined as yet, nor have many other criteria of relative value of estimating them. Thus, other combinations of unbalanced nested designs require investigation for robustness. In addition, if some random-plot loss can be expected, then designs should be examined for the possible configurations of unbalance such losses may induce. As for other designs, such as the factorial or diallel, only preliminary investigations on the general utility of unbalanced sets have been conducted (Mostafa 1967; Gaylor and Anderson 1960) and further examination is required to determine their value in forest genetics.

If we can accommodate possible imbalances in design configurations, then efficiency of estimation of the variance components should also be considered. For any given set of data, it is possible to construct almost an infinite number of sums of squares as functions of the variance components, and, for unbalanced experiments, it is not immediately obvious which sum of squares would minimize the variance of estimators of the variance components. Some investigations (Rao 1971) indicate that other sum of squares functions than those traditionally used in analyses of variance can considerably improve estimates of variance components while maintaining unbiasedness of the estimators. However, the unbiasedness requirement may also impose restrictions on estimating functions such that even more efficient estimators may be available. Since forestry experiments are so costly, foresters should consider both unbalanced designs and minimum variance estimators for their analyses.

SPECIAL FORESTRY PROBLEMS

All these designs can be modified in several ways to give information on additional parameters of environmental sources of variation, and genotype-by-environmental interactions. Among the various kinds of environmental effects, it is often desirable to determine the sizes of random variations among what the forester would consider to be roughly comparable sites within geographic planting zones. Often, it is also desirable to estimate variances among smaller replicates within plantation-sized areas, among plots within replicates, and even among individual trees within plots. To satsify such demands, the simplest experiment would replicate the entire genotypic array over as many sites and plantations with as many plots as feasible. In any large-scale silvicultural experiment, testing locational or site differences or soil, spacing, or any treatment differences, the insertion of any kinds of relatives as split subplots may often be feasible. It might occasionally be cheaper to allow families to be major plots and cultural effects to be subplots, but the former case is more likely to occur. Thus, partially balanced factorial arrangements of the environmental variables should provide some important efficiencies in the number of plots required of each family. While it may be impossible to sample all combinations of all levels of the important site factors, it may be possible to sample enough variations to estimate the response surface for each family or replicated genotype. The form and magnitude of genotype-environment interactions could then be examined.

However, even reducing the number of replicated plots required of each family by judicious sampling of site factor combinations may often not be adequate for the crosses with relatively few seedlings. In such cases, some efficiencies in combining genetic and cultural variation in the same experiment might be achieved by at least partial confounding of one with the other. Unintentional mortality and differential planting of families will also cause some confounding, but the deliberate planning of unbalance may be feasible. Partial confounding for mean estimation is a well-established technique, but for variance component estimation it is not well developed in biological experiments. The objective again is to affect the allocation of degrees of freedom to those mean squares required to estimate the important variance components. Hence, by locating most families on some environments and a few on many environments, the usual excess of degrees of freedom on the family-by-environment interaction mean square is avoided while more effort can be made to sample either more families or environments. Nevertheless, the families not represented on some sites cannot be adequately evaluated on those sites if the interaction is high. The only recourse available for estimating means or response surfaces would be to model the interaction forms so that families that behave similarly are grouped into homogeneous reaction classes and their response

estimated relative to their behavior within these classes. It would, therefore, be desirable to combine complete environmental sampling of all families for response estimation on some minimal set of sites which span meaningful site variations, with partial genotype sampling of more complete site samples.

Another complicating factor in design is that most experiments are established for several purposes. At least several traits are usually measured for variances and for the covariances among traits attributable to various genetic and environmental effects. A suitable design for one purpose or for a trait with a high genetic component may not be suitable for another purpose or trait. Hence, designs usually have to be chosen to provide reasonably precise estimates of variance components for each trait and for their covariances. Criteria such as minimizing the error volume or minimax criteria may be better than the traditional criteria of minimizing average error. In general, however, the simplest, most robust design will be of greatest utility. Thus, designs which can be subdivided into balanced subsets can be easily analyzed within subsets. Also, designs that do not require the use of poorly estimated covariates for adjusting means should be favored to avoid the additional estimation errors so generated.

An additional type of design and estimation problem exists when genotypic competitive effects are to be estimated. Two experimental strategies might be adopted according to whether estimates of specific interactions between families or a general level of competitive interaction is to be estimated. If pair-wise competition estimates are to be tested, the parameterization and estimation techniques, outlined by Byrd and others (1965), seem appropriate and generally applicable. The parameters described by Schutz and Usanis (1969) and estimated by Schutz and Brim (1971) for soybeans are further developments describing and measuring competitive effects between noninterbreeding elements. More direct utility for populational effects with intermating populations is achieved by Griffing's (1967) parameterization. However, estimation is difficult unless methods such as developed by Sakai and others (1968) can be applied. The genetic consequences of competition effects, however, have been investigated by Huhn (1970c).

For estimation purposes, trees also present unique experimental problems in space and time, especially if plot thinning is expected. Blocking and planning for spacing for the duration of the experiment require care to assure that reasonable plot sizes are maintained for the spacing and other environmental conditions required for analysis of larger trees. While spacing experiments themselves might be of interest with respect to genotypic variations in density response (Namkoong 1966), the more common consideration is to assure that genotypes are measured under the environments planned for specific ages. If growth curves or other time-dependent responses are to be analyzed, extra care is required to assure that plot integrity is maintained. Similarly, correlation

analyses among juvenile and mature tree characteristics will require plots at a variety of ages and will require some plot continuity. The relative and absolute sizes of variance components for single traits can generally be expected to change with time. Some may, in fact, tend to disappear while others maintain their same relative size according to the form of genetic control of growth (Namkoong and others 1972).

The duration of tree life also creates a dimension of uncertainty of future environments. Future environments will surely be different from present ones, and variations will undoubtedly be controlled by some events not yet discernible. The problem for the forest geneticist is to determine a reasonable array of environments for which his estimations will be valid. Therefore, sampling a wide array of controlled as well as uncontrolled variables is desirable to define the kinds of interactions large enough to worry about as well as the average performance over uncontrolled environmental variations. Initially, the geneticist may often reasonably guess that location and site differences are larger than the time-trend differences he can sample and hence that genotypic interactions with site variations may be most important. Certain environmental effects that occur periodically, such as disease and insect epidemics, however, may require special sampling of periodic environmental events.

Otherwise, estimation problems for forest trees are not significantly different from those for any other organism. The principal problems are as stated in the beginning of this chapter. Perhaps the greatest problem is how to assure the continuity of experimental administration so that the value of well-designed experiments is not compromised by future neglect or change in personnel. However, since changes in personnel and organization are to be expected, greater reliance should be placed on simple, easily analyzable designs. Considerable sagacity is required to plan for an uncertain future so that the parameters of future value can be well estimated.

CHAPTER 9
POPULATION GENETICS

The importance and pervasiveness of genetic effects and sizable variances in forests are well established. The origins and evolutionary utility of these variances in the evolution of tree populations, however, are not clear. If we are to control the future evolution of tree species, it behooves us to know not only the status of existing genetic variations but how they originated and may be maintained. It might then be easier to direct evolution for any given set of objectives while we are developing our understanding of the dynamics of natural forest evolution. If we accept the concept of tree populations with significantly changing genetic effects and variations, then we must determine the nature of the forces which disturb any stable uniformity. We shall try to understand how variations might originate and be propagated and how they affect the resultant distributions of gene effects. We first consider how the forces might act independently and then consider how their joint actions might operate. We shall consider the forces as they operate in large random-mating populations without linkage disequilibrium and shall therefore assume that deterministic models are adequate and genetic and zygotic frequencies are determined by gene frequencies. Finally, expansions of the analyses to cases in which mating is restricted and stochastic variations are significant will then be considered for their effects on the evolutionary processes.

MUTATION

The basic force which provides alternative alleles to the population is mutation. However, single mutations are rare and would not occur frequently enough to be a major source of variation in a population without persistent recurrence. While this persistence depends on the mutability of the alleles and their frequency in the population, it could have cumulative effects on allelic frequency. Not all changes in the molecular structure of the DNA material result in important differences. Some changes have no direct effect on amino acid structure and some have no effect on function even if an amino acid sequence is disturbed. Thus, we might expect molecular changes in DNA to occur at a higher rate than what is actually observable as mutants that might have any observable selective effect (Kimura 1969). At the functional cistron level of gene locus definition, the per locus, per generation

rate of mutation has commonly been estimated at around 10^{-5}. While most studies have been conducted on major genes, there is little apparent difference in the mutation rate of polygenes. That is, the polygene which has a less visible, individual effect that may be partly masked by environmental and other genetic effects has been estimated to have mutation rates of about the same magnitude on a per locus per generation basis (Mukai 1969). The two types of genes are also similar in the existence of variation in mutability (Russell and others 1963), and the average mutation rates are similar. On a more readily observable basis than on the traditional per locus basis, Mukai (1969) estimates the mutation rate for polygenes per second chromosome of Drosophila per generation to be around 0.14, and for lethal mutations on the same basis to be around 0.006. On a different but perhaps more useful basis for plant breeders, Russell and others (1963) estimated the mutation rate for several quantitative types of characters in corn populations and found an average per trait, per gamete, per generation mutation rate of 0.028.

Thus, for any trait that has a large number of loci, persistent mutation rates of even 10^{-5} or 10^{-6} per locus, per gamete, per generation can have some of its variation generated by mutations alone. It is clear that recurrent mutation from one allele (a_1) to another (a_2) would eventually either move the population to homozygosity for a_2, or the mutation back from a_2 to a_1 would produce a gene frequency equilibrium. Of less practical interest, but great fun, is the analysis of the fate of nonrecurrent mutations which we ignore.

To study the effect of mutations, we first ignore the confounding effects of selection and consider how mutants may affect the population's genetic means and variances if all products are selectively equivalent and populations are large enough that sampling error can be ignored. In this case, the initial frequency of an allele q_0 and its mutation rate μ to another allele strictly determine the frequency of the allele in the next generation, the frequency being decreased by μq_0. Hence, $q_1=q_0-\mu q_0=q_0(1-\mu)$ and $q_2=q_1-\mu q_1=q_1(1-\mu)=q_0(1-\mu)^2$, etc. Then, $q_t=q_0(1-\mu)^t$.

To simplify this expression, an approximation can be substituted to give a useful form to the equation. This approximation depends on the condition that μ is very small and hence that μ^2 and all higher powers of μ are almost zero. If that is true, then the identity

$$e^{-\mu}=1-\mu+\frac{\mu^2}{2!}-\frac{\mu^3}{3!}+\ldots$$

can be approximated by $e^{-\mu}\approx 1-\mu$,

and therefore $$(1-\mu)^t\approx(e^{-\mu})^t=e^{-\mu t}.$$

Hence, we can trace, with close approximation, the change in gene frequency from any initial q_0 to the frequency at any time

t, q_t, by $q_t \approx q_0 e^{-\mu t}$. Alternatively, we may wish to approximate the progress (or regress) of gene frequency under continuous time or completely overlapping generations and define the rate of change in gene frequency $\frac{dq}{dt}$ at a point in time as $-\mu q$. Then:

$$\frac{dq}{dt} = -\mu q$$

or $$\frac{dq}{q} = -\mu dt.$$

Integrating the equation yields:

$$\log q = C - \mu t$$

or $$q_t = q_0 e^{-\mu t}.$$

In either case, the eventual result of persistent mutations is that q approaches zero as the alternative allele eventually takes over the population. If we wish to consider that back mutation occurs at a rate of γ, then the gene frequency decrease of $-\mu q$ is offset by the increase $\gamma(1-q)$ from the other allele which exists at frequency $1-q$. Then:

$$q_1 = q_0 - \mu q_0 + \gamma(1-q_0)$$
$$= \gamma + q_0(1-\mu-\gamma),$$

and $$q_t = \gamma + q_{t-1}(1-\mu-\gamma),$$

and we can derive that $q_t = \gamma t + q_0(1-\mu-\gamma)^t$.

It is clear that there must eventually be some equilibrium between the alternate alleles and hence that $q_t = q_{t-1}$ after some large time period. At that time, the equilibrium frequency q_e must clearly be:

$$q_e = \gamma + q_e(1-\mu-\gamma)$$

or $$q_e = \frac{\gamma}{\mu+\gamma}.$$

We can also derive that for any intermediate generation i between 0 and equilibrium,

$$q_i - q_e = (1-\mu-\gamma)^i(q_0 - q_e)$$

or approximately $q_t - q_e = e^{-(\mu+\gamma)t}(q_0 - q_e)$.

For continuous time and overlapping generations, we can use the rate of gene frequency change $\frac{dq}{dt}$ from:

$$q_1 - q_e = (1-\mu-\gamma)(q_0 - q_e)$$

or approximately:

$$q_1 - q_e = (q_0 - q_e) - (\mu+\gamma)(q_0 - q_e)$$

or $\quad q_1 - q_0 = -(\mu+\gamma)(q_0 - q_e)$

and $\quad \frac{dq}{dt} \approx -(\mu+\gamma)(q - q_e).$

Solving this equation by integration yields:

$$q - q_e = e^{-(\mu+\gamma)t}(q_0 - q_e),$$

as approximated above.

In addition to back mutation, selection against new mutants can also prevent a complete change to the new allele and can usually be expected to be a potent force in suppressing the frequency of mutant alleles. Among the molecular variants which may exist at a locus, many may have the same average fitness-endowing qualities. We can generally expect modifications in alleles that exist at substantial gene frequency to lead to decreases in average fitness if the modifications affect differences at the cistron level. In fact, among any set of alleles, regardless of their mutation history, frequency will be controlled by the average selective values of their zygotes. Thus, for any two alleles, A and A', with frequency q and $1-q$, respectively, three zygotic fitnesses may exist; W_2 for AA, W_1 for AA', and W_0 for $A'A'$. For large populations in random mating, the relative zygotic frequencies are q^2 for AA, $2q(1-q)$ for AA', and $(1-q)^2$ for $A'A'$. If selection operates according to the relative fitnesses of the zygotes by reducing their contributions to the next generation, the new frequency of gene A will be:

$$q_s = [W_2 q^2 + (1/2) W_1 2q(1-q)] / \overline{w},$$

where $\overline{w}$, a scaling factor of the average fitness, is:

$$\overline{w} = W_2 q^2 + W_1 2q(1-q) + W_0 (1-q)^2.$$

A simple parameterization of the W's may help interpretation. Let $W_2 = 1$, $W_1 = 1-h$, and $W_0 = 1-s$ so that the contrast between homozygotes is measured by s, and h determines the level of dominance $(0 < h < s)$ or overdominance $(h < 0)$. Then by substituting these W values into the above equation, the effect of selection produces a gene frequency q_s of

$$\frac{q^2 + (1-h)q(1-q)}{1 - 2q(1-q)h - (1-q)^2 s},$$

and the one generation change in gene frequency then is

$$\Delta q_s = q_s - q = \frac{q(1-q)[qh - (s-h)(1-q)]}{\overline{w}}.$$

We have thus constructed a model for changing gene frequencies for any s and h values. We can now consider that selection is for A (therefore $s < 0$), but that mutation occurs from A to A' at frequency μ, and that simple mutation changes frequency of

A by $\Delta q_m = -\mu q$. Since the net change in gene frequency is the sum of $\Delta q_s + \Delta q_m$, and the net change will eventually reach an equilibrium, $\Delta q_s + \Delta q_m = 0$, then $\Delta q_s = -\Delta q_m$, and hence

$$\Delta q_s = \mu q_e,$$

or
$$\frac{\Delta q_s}{q_e} = \mu = \frac{(1-q_e)\,[q_e h + (s-h)\,(1-q_e)\,]}{\overline{w}_e},$$

at the equilibrium frequencies of $q\,(q_e)$ and at those frequencies

$$\overline{w} = \overline{w}_e.$$

Some special cases are illuminating to indicate the effect of dominance on the value of q_e. If gene actions of the new mutants are not masked by any dominance of the original A alleles, then $h = (1/2)\,s$, all factors are positive, and using these values in the above equation, the equilibrium gene frequency of $A' = 1 - q_e = \frac{\mu}{h}\overline{w}_e$. If h is small and μ is small, then $\overline{w}_e$ is almost 1 and $1 - q_e \approx \frac{\mu}{h}$ or $\frac{2\mu}{s}$. On the other hand, new mutants are often recessive, making h almost zero, and then the above equation solves for $1 - q_e = \sqrt{(\mu / s)}$. More generally, we can derive approximations when μ is very small with respect to s or h, and both s and h are close to zero themselves, then $\overline{w}_e \approx 1$. Then, also:

$$\mu = (1-q_e)\,[h q_e + (s-h)\,(1-q_e)\,]$$
$$= (1-q_e)\,[h + (s-2h)\,(1-q_e)\,].$$

This is a quadratic equation for $(1-q_e)$ and can be solved by finding

$$1 - q_e = -\frac{h + \sqrt{h^2 + 4\mu(s-2h)}}{2(s-2h)},$$

which is approximately $1 - q_e \approx \frac{\mu}{h}$. This approximation is a useful rule of thumb. It is clear that mutations do occur among forest trees, and may even occur at higher frequencies among some species (Sorensen 1969). Since trees generally differentiate sexual organs from their outer branches' terminal meristems and since some species are quite susceptible to radiation (Mergen 1963) and temperature shocks (Eriksson and others 1972), it might be relatively easy to accumulate mutations in the germ plasm. The effects of selection on these introduced alleles are complicated by any agencies that restrict the randomness with which gametes are distributed—either through mating patterns or through recombinations among other genetic loci.

MIGRATION

The simple effects of migration can be similarly modeled as a source of new alleles which are fed into the local population and therefore affect the gene frequency. For large populations with random mating, the encroachment of mutants or migrants must recur persistently to have much effect, but when they do they can have cumulative effects on means and variances even if the new alleles have small effect in any one generation. The effect of low migration into a population from some constant source is indistinguishable from mutation and is often lumped in the heading of "mutation" as a source of new variation. High migration rates would diminish the value of the approximations and may operate differently with respect to selection. Entire genome substitution in gametes (for example, pollen flight) or zygotes (for example, seed dispersal) is the form of migration and can occur in random patterns or may flow from high- to low-density areas of allelic concentrations. Such clustering of genes and zygotes creates correlations among loci on the one hand, and in the probabilities of mating among relatives on the other. An adequate treatment of migration effects therefore requires consideration of the breeding structure of populations which we postpone to a later section of this chapter. The simple treatment of migration rates as another form of mutation, however, can serve as a reasonable working model.

In forests, seeds and pollen can migrate by the action of wind, water, and animals. For the gametes, eggs do not generally migrate, but pollen is often carried long distances and can heavily influence migration rates of foreign alleles.

While we cannot treat multiple-locus problems adequately, it is important to point out that when genes migrate into a population together, as in whole genome substitution, independent treatment of several loci is inadequate. To see one feature of the phenomenon we can determine that after one generation of random mating, each locus is expected to produce zygotes in their expected Hardy-Weinberg equilibria frequencies: $p^2(AA):2p(1-p)(AA'):(1-p)^2(A'A')$ for whatever their new (p) frequencies are. We can simply demonstrate this by considering that random mating with gene frequency p necessarily yields the expected zygotic frequencies. Regardless of the nonequilibrium zygotic frequencies in one generation, the gene frequencies determine the next generation's zygotic frequencies at the Hardy-Weinberg equilibrium, as long as population size is large and mating is at random. However, when alleles enter the population together or are selected for differently in combination rather than independently (epistatically), then the frequency of $AABB$ zygotes is distorted with respect to the eight other genotypes, while the A and B loci still occur with their expected distribution.

These considerations are most significant if selection on gene pairs or on multiple loci is not independent, and we considered those effects in detail under epistatic selection theory. For the

moment, however, it is clear that gene migrations and subsequent selection can become vastly complicated by considering multiple loci. Another factor which we have thus far assumed absent is the variation caused by having subdivided or small population sizes such that sampling error causes gene frequencies to depart from expectations.

THE SPECIAL RESTRICTIONS OF SAMPLING ERRORS IN SMALL POPULATIONS

For most species of forest trees, populations are not homogeneously distributed and mating is restricted by such external factors as distance and by such internal factors as phenological mechanisms. Even if we consider that we have studied only a small part of the history of forest tree matings, the general concept of population regeneration must consider the correlation between mating frequency and geographical or phenological closeness. Among trees with little pollen or seed dispersal, such as yellow-poplar, we might even expect that population islands have developed in isolated coves. With large distances between cove populations for many generations and little indication of past populations which might have significantly bridged between present populational lineages, little gene exchange is likely to have occurred. In addition, periodic reductions and explosions of populations can be expected to increase homogeneity within population segregates unless the explosions were sufficiently large and frequent to induce enough exchange among populations to offset loss of alleles when populations were small and few progenitors regenerated the local lineages. Since reducing the number of independent or unrelated families is equivalent to increasing the average relatedness among trees within populations, genetic differences among populations are generated. Unless the differences are later affected by selection, we would observe populational differences similar to the family differences observed in estimating experiments. The higher the relationship among trees in a population or family, the greater the differences will be among populations of families according to the size of the genetic effects we choose to measure. In experimentation, of course, we can separate environmental effects and can create known, regular relationships. In contrast, natural populations have environmental and genetic effects confounded, and the degree of relatedness and inbreeding among parents is generally unknown. In addition to confounding observed differences among stands, selection may either increase or decrease the genetic differences themselves either by selecting for different environmental responses or by selecting for the same homeostatic response to different environments. The latter phenomenon is strikingly observed in yellow-poplar in the contrast between a lightly selected trait, leaf morphology, versus a trait directly affecting fitness, seedling height growth. While growth rate exhibited about the same amount of genetic variance among

stands as among trees within stands, leaf morphology exhibited much higher variances among stands than among trees within stands (Kellison 1970). In contrast, slash pine is a more uniformly distributed species which exhibits about the same levels of genetic variance among and within stands for the growth and vigor traits as for leaf morphology (Squillace 1966c).

The nature and effects of nonrandom mating and any subsequent inbreeding, therefore, must be considered to thoroughly understand the development of forest tree populations. Since future populations may be subjected to even more restricted lineages by mating only specially selected individuals, the effects of inbreeding systems must also be understood to control future population evolution.

INBREEDING

Consider that populations have diverged with respect to some trait so that the gene frequency varies among populations. Regardless of the reasons for the divergence, we can regard each population as perhaps randomly mating within the neighborhood and hence being in Hardy-Weinberg equilibrium with respect to its particular local gene frequency. Thus, each population has a p_i gene frequency and $p_i^2(AA):2p^2(1-p_i)\ (AA'):(1-p_i)^2A'A'$ zygotic frequencies. However, there would also exist a variance in gene frequency among populations, σ_p^2. Because there is a variance among populations, an excess of homozygotes in the general population is generated regardless of any random mating within subpopulations. Within each subpopulation, the frequency of AA homozygotes is p_i^2, and over all subpopulations, the frequency is therefore Σp_i^2, which implies an average over n subpopulations of $(1/n)\Sigma p_i^2$. Over all of the subpopulations, there is also an average frequency, $\bar{p}$, which is the mean for the whole species and from which we could estimate the expected frequency of AA homozygotes if mating was truly random, as $\bar{p}^2$. The average of the squares is not generally equal to the average squared; that is,

$$(1/n)\Sigma p_i^2 \neq \left(\frac{\Sigma p_i}{n}\right)^2$$

Then, since $\sigma_p^2 = (1/n)\Sigma p_i^2 - \bar{p}^2$, the observed frequency of AA can be expressed as $(1/n)\Sigma p_i^2 = \bar{p}^2 + \sigma_p^2$. In other words, the observed frequency of homozygotes exceeds the random-mating frequency by an amount equal to σ_p^2. This is Wahlund's principle of the subdivision of large populations and can be extended to the frequency of $A'A'$ as $(1-\bar{p})^2+\sigma_p^2$, and to the diminished frequency of AA' as $2\bar{p}(1-\bar{p})-2\sigma_p^2$. Thus, inbreeding is reflected in an excess of homozygotes at the expense of heterozygotes and exists when subpopulations differ regardless of any random mating within the subdivisions. We have thus also implied that inbreeding is a general phenomenon which exists whenever the definition of

population is made broad enough to include genetic differences among subdivisions. Inbreeding and random mating are thus defined relative to some base population which must be well defined. Regardless of the size or history of the populations, Hardy-Weinberg equilibrium within even very small populations simply means that the zygotes are found at the frequencies expected from the local gene frequency and independent gametic association. Without gene frequency differences, we cannot detect mating patterns except by direct observation of gametic unions. If for any reason differences exist among populations or groups in gene frequency, an excess of homozygotes exists, and inbreeding is said to exist. We can parameterize the phenomenon by considering the array of male and female gametes with the same average gene frequency p and constructing a mating table as follows:

Female	Male		Total
	A	A'	
A	$\bar{p}^2+E$	$\bar{p}(1-\bar{p})-E$	$\bar{p}$
A'	$\bar{p}(1-\bar{p})-E$	$(1-\bar{p})^2+E$	$1-\bar{p}$
Total	$\bar{p}$	$1-\bar{p}$	1

Mating $A\times A$ occurs at frequency $\bar{p}\times\bar{p}$ plus some deviation, E. Since the average gene frequency for A remains constant at $\bar{p}$, the same deviation E must occur with opposite sign for the frequency of heterozygote formation when A' is the other gamete. Hence, all cells can be parameterized for their expected frequency $\pm E$. To construct a variance and covariance of frequencies, we can use a dummy variable t which takes the value 1 when A occurs and 0 when A' occurs. Then,

$$\mu_t = 1\cdot\bar{p}+0\cdot(1-\bar{p})=\bar{p},$$

and $$\sigma_t^2 = 1\cdot\bar{p}-\bar{p}^2=\bar{p}(1-\bar{p})$$

are the marginal means and variances, and for two samples of mates, t and t'

$$\sigma_{tt'} = E(tt') - [E(t)][E(t')]$$
$$= \bar{p}^2+E-\bar{p}^2=E$$

and the correlation, $$\rho=\frac{E}{\bar{p}(1-\bar{p})}.$$

Hence, $$E=\rho\,\bar{p}(1-\bar{p}),$$

where ρ is the correlation between uniting gametes. It has also been defined as F, the inbreeding coefficient (Malécot 1969).

We can now see that the frequency of AA homozygotes is increased over $\bar{p}^2$ by $F\bar{p}(1-\bar{p})$ as is the frequency of $A'A'$, while the heterozygote frequency is reduced from $2\bar{p}(1-\bar{p})$ by an amount equal to $2F\bar{p}(1-\bar{p})$. F is a measure of relatedness within mating groups relative to some base population mating entirely at random.

Mechanisms which can generate variations in zygotic frequencies by sampling different alleles or allelic frequencies can thus clearly generate measurable variations among subpopulations or family units.

CORRELATIONS AMONG RELATIVES

It is useful to quantify relationships within groups to describe the inbreeding structure of breeding populations and to analyze the genetic variances caused by family differences. A more complete treatment than given thus far on coefficients of relatedness is needed.

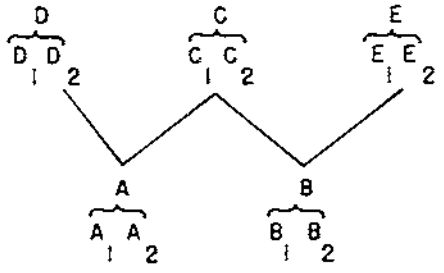
Two measures of relatedness are commonly defined, a retrospective coefficient of relationship among individuals and a more prospective coefficient of inbreeding for individuals which result if relatives are mated. Malécot (1969) defines relationships in terms of probabilities of alleles in two individuals being identical derivatives from some common ancestor. With this coancestry parameter, various degrees of relationships can be expressed and can then be used to examine the effects on inbreeding and homozygosity which are associated but not identical with coancestry.

The coancestry of two trees is the probability that a randomly drawn allele of one tree is identical by descent to a randomly drawn allele of the second tree. Identity by descent exists if the two alleles are copies of an allele that was possessed by some single progenitor and carried through whatever lineages may exist. While this implies that the alleles are alike in molecular composition, different allelic ancestral histories can still be identical in the state of their molecules. Thus, we require the existence of a base population in which no relatedness is assumed to exist in order to define degrees of relationship.

From any such base population without relationships among individuals, probabilities of allelic identities by descent are computed to get the coancestry. Thus, for trees A and B with alleles A_1A_2 and B_1B_2, the coancestry between A and B (f_{AB}) is computed as the sum of probabilities that A_i and B_i are chosen at random ($Pr(A_i, B_j)$) and are identical by descent $Pr(A_i \equiv B_j)$. Thus,

$$f_{AB} = \sum_{i,j} Pr(A_i, B_j)\,(Pr(A_i \equiv B_j)) = \tfrac{1}{4}\,Pr(A_1 \equiv B_1) + \tfrac{1}{4}Pr(A_1 \equiv B_2) + \tfrac{1}{4}Pr(A_2 \equiv B_1) + \tfrac{1}{4}Pr(A_2 \equiv B_2).$$

If A and B are half-sibs from a common parent tree C and the unrelated alternate parents are D and E, we can diagram their relationships as:



$Pr(A_1 \equiv B_1)$ depends solely on A_1 and B_1 being copies of the same gene from C, either C_1 or C_2, and this is the probability that $C_1 \equiv A_1$ and $C_1 \equiv B_1$; or that $C_2 \equiv A_1$ and $C_2 \equiv B_1$; or that $C_1 \equiv A_1$, $C_2 \equiv B_1$, and $C_1 \equiv C_2$; or finally that $C_2 \equiv A_1$, $C_1 \equiv B_1$, and $C_1 \equiv C_2$. Since A_1 has probability $1/2$ of being from C instead of D, and the probability is $1/2$ that it is a copy of C_1 instead of C_2, then $Pr(C_1 \equiv A_1) = 1/4$, and similarly $Pr(C_1 \equiv B_1) = 1/4$, therefore, $Pr(C_1 \equiv A_1$, and $C_1 \equiv B_1) = 1/16$. Also, $Pr(C_2 \equiv A_1$, $C_2 \equiv D_1) = 1/16$, and for the last two cases, designating f_c as the probability that $C_1 \equiv C_2$ (the inbreeding of parent C), $Pr(C_2 \equiv A_1$, $C_1 \equiv B_1$, $C_1 \equiv C_2) = 1/16\ f_c$ and $Pr(C_1 \equiv A_1$, $C_2 \equiv B_1$, $C_1 \equiv C_2) = 1/16 f_c$. Then $Pr(A_1 \equiv B_1) = \frac{1+f_c}{8}$. Using similar derivations for $Pr(A_1 \equiv B_2)$, $Pr(A_2 \equiv B_1)$, $Pr(A_2 \equiv B_2)$ gives our coancestry of A and B as:

$$f_{AB} = 1/4\left[\frac{4(1+f_c)}{8}\right] = \frac{1+f_c}{8}.$$

For other relatives, or those in which some relationships may exist among precedent parents, the stepwise paths may become rather intricate but, fortunately, some algorithms reduce the tedium (Cockerham 1971; Harris 1964; Li 1955; Malécot 1969).

The coefficient of relationship therefore expresses the degree to which alleles have some probability of identity among individuals. Clearly, if relatives are mated, then the probability of having alleles identical by descent in the offspring is the same as the coefficient of coancestry among its parents. Hence, in the above example, if A and B mate to produce an offspring X, the inbreeding coefficient of X,

$$f_x = f_{AB} = \frac{1+f_c}{8}.$$

The coancestry of an individual with itself is the probability of drawing the same allele twice ($1/4$ for each allele) or drawing alternate alleles which may be identical (f), and hence

$$f_{AA} = \frac{1+f_A}{2}.$$

We can also derive that, in general, the coancestry of two individuals is the average of the two coancestries between one tree and the parents of the other, or the average of four coancestries between the parents of both, etc. In Wright's (1922) traditional

notation, the covariance between individuals A and B is $2f_{AB}$ and the coefficient of relationship is

$$r_{AB}=\frac{2f_{AB}}{\sqrt{1+f_A}\sqrt{1+f_B}}.$$

In the notation of chapters 7 and 8, the coefficient used for contributions of genetic variances to the phenotypic covariances among relations is $2f_{AB}$.

The probability measures of relationship and inbreeding can be extended to groups of individuals within which some relationships may exist and between which some identical forebears may also have existed (Cockerham 1967). Coancestries among groups may then be computed and would be the same as the inbreeding of progenies from crosses between the groups. Extensions to more than one degree of hierarchal relatedness and analyses of variances in gene frequencies among the various levels of the hierarchy are also possible (Cockerham 1973).

In populations in which relationships exist, the associations among genotypes often are not the same as if truly complete random mating in large populations existed. One of the problems induced by the lack of complete and continuous random mating is that variations in gene frequency are induced and the genetic variance can change.

We have already formulated the degree to which increases in relationship among mates increases the probability of identity by descent and have examined the rate of increase in homozygosity (or identity in molecular form) with increased inbreeding as:

$$P_{AA}=p_A^2+Fp_A(1-p_A)=Fp_A+(1-F)p_A^2.$$

It is also interesting to notice that the difference in the frequency of homozygotes caused by inbreeding can be written as:

$$p_A^2+Fp_A(1-p_A)-p_A^2=Fp_A(1-p_A).$$

Hence, the difference in homozygosity is a linear function of F at any given gene frequency. In particular, at low gene frequencies, the percentage change in homozygosity can be very large since the above difference, taken as a ratio of the noninbred frequency of homozygotes

$$[p_A^2+Fp_A(1-p_A)-p_A^2]\ /\ p_A^2,$$

is $F(1-p_A)\ /\ p_A$ and can be very large.

The increase in homozygosity and its effects on genotypic frequencies also affect the variance in gene frequency among replicate populations.

Phenotypically, the phenomenon can be observed in the increased variance of an additive gene-action locus as inbreeding polarizes the population into the two homozygotes. If the zygotic values are scaled as $2(AA):1(AA'):0(A'A')$ with $\bar{p}_t$ average gene frequency, then within any subpopulation which is randomly

mated with gene frequency p_i, the mean is $2p_i^2+2p_i(1-p_i)$ and the variance is $4p_i^2+2p_i(1-p_i)-4p_i^2=2p_i(1-p_i)$. However, over all populations, the zygotic frequencies are $\bar{p}^2+F\bar{p}(1-\bar{p}):2\bar{p}(1-\bar{p})(1-F):(1-\bar{p})^2+F\bar{p}(1-\bar{p})$. As F approaches 1, the heterozygotes disappear and each line becomes AA or $A'A'$. For any intermediate value of F, the total genetic mean is again $2\bar{p}$ but the variance is

$$4\left[\bar{p}_A^2+F\bar{p}_A(1-\bar{p}_A)\right]+2\bar{p}_A(1-\bar{p}_A)(1-F)-4\bar{p}_A^2$$
$$=2(1+F)\bar{p}_A(1-\bar{p}_A).$$

If no inbreeding existed, then $p_i=\bar{p}$, but with inbreeding the genetic variance is $(1+F)$ times the noninbred variance. Hence, at $F=1$, inbreeding can double the additive genetic variance.

In the sense that gene frequency variations occur among small subpopulations, inbreeding can also be expressed in terms of the fact that the total population size does not give us an accurate idea of how many individuals may be considered to be random mating. Even if mating is at random, the relationships will inevitably accumulate and result in some levels of inbreeding. To compare inbreeding populations, it is often useful to use as a measure the number of individuals which, in an idealized, completely random mating population, would produce the gene frequency variations or inbreeding displayed by the actual number of mating individuals. Two such conceptual population sizes, both called the effective number, have been used to indicate the increases in homozygosity due to limited numbers in any population on the one hand, and to indicate the variance in gene frequency among finite populations on the other hand.

The first term is called the inbreeding effective number and is less than the census number of parents because a finite number of parents will produce offspring with different degrees of relationship among them and unequal representation of parental genes increases the average degree of relationship. Thus, out of N monoecious parents, homozygosis is increased when the $2N$ gametes randomly associate and create a probability of $\frac{1}{2}N$ that they come from the same individual. Heterozygosis is therefore decreased by a factor of $(1-\frac{1}{2}N)$ each generation. Even completely random mating within a finite population therefore has an expected rate of increase in inbreeding. If families are created, then probabilities of random alleles being alike by descent may be increased above that expected by random mating simply by having different family sizes. Thus, if average family size produced is $\bar{k}$, then N parent trees could produce $N\bar{k}$ trees and then, $\frac{N\bar{k}(N\bar{k}-1)}{2}$ pairs of mates. Of all possible such pairings, each family would produce $\frac{k_i(k_i-1)}{2}$ pairs in which both parents came

from the same family. In an idealized population, these pairings within families would occur only $1/N$ times and by so defining an effective N_e,

$$1/N_e = \frac{\Sigma k_i(k_i-1)}{N\bar{k}(N\bar{k}-1)}.$$

Since $\Sigma k_i^2 - \bar{k}^2 = \sigma_k^2$, we can rewrite this equation as

$$N_e = \frac{N\bar{k}(N\bar{k}-1)}{(N-1)\sigma_k^2 + N\bar{k}^2 - N\bar{k}}$$

If $\sigma_k^2 = k$, as would be the case if k was Poisson distributed, then $N_e = N$. If k_i was controlled so that $\sigma_k^2 = 0$, then $N_e = 2N$.

Another way in which the census member would not accurately reflect rates of loss of heterozygotes under conditions of restricted parental matings occurs when unequal numbers of dioecious males and females exist. In such cases

$$N_e = \frac{4N_m N_f}{N_m + N_f}.$$

If $N_m = N_f = N/2$, then $N_e = N$. Other effects, such as overlapping generations, differential reproductive rates among age classes, etc., would also decrease N_e relative to N (Giesel 1971).

The alternative measure of effective numbers is also an abstraction related to the census number that would exist in an idealized situation for the variance in gene frequency to be as expected. The concern in this case is that the progeny population is expected to display variations among their subpopulations according to a binomial distribution; $\sigma_p^2 = \frac{\bar{p}(1-\bar{p})}{2N}$ for subpopulations of size N. However, due to the same kinds of influences that make N a biased number for predicting inbreeding, N is not a good number to use to predict σ_p^2. For the case in which family sizes may vary, the N_e required to satisfy $\sigma_p^2 = \frac{\bar{p}(1-\bar{p})}{2N_e}$ is:

$$N_e = \frac{2N}{\frac{\sigma_k^2}{\bar{k}}(k+F) + (k-F)}.$$

Again, F is the departure of zygotic frequencies from Hardy-Weinberg equilibrium frequencies (Crow and Kimura 1970).

If populations undergo sequential, temporal variations in N, then the variance of p will also change over generations. In order to determine an N_e that would give an average estimate of how σ_p^2 was generated for n such generations, we can estimate the sampling variance for the sequence as:

$$\sigma_p^2 = \frac{\bar{p}(1-\bar{p})}{n}\left[\frac{1}{2N_1} + \frac{1}{2N_2} + \ldots + \frac{1}{2N_n}\right]$$

and define N_e as satisfying $\sigma_p^2 = \frac{\bar{p}(1-\bar{p})}{2N_e}$. Then equating the two σ_p^2 estimates yields $N_e \frac{n}{\sum_i \frac{1}{N_i}}$, the harmonic mean of the various N_i's. The harmonic mean, being more strongly affected by low numbers than is the arithmetic mean, gives lower values than the arithmetic mean, and therefore indicates that the effective, idealized population size is strongly affected by bottlenecks of low N_i.

Thus, regardless of present population sizes, it is quite possible by a variety of means that the effective population sizes may be small, homozygosity relatively high, and our forests largely made up of partially isolated populations with different gene frequencies. With populational subdivisions, we are required to consider how multiple subpopulations may have evolved. To formalize consideration of these effects as well as to study the progress of inbreeding in possible breeding programs, regular systems of mating have been studied for the rate of increase in F, or homozygosity (Li 1955; Crow and Kimura 1970; Wright 1969). Clearly, systems that limit the number of parents per group most severely restrict mating and induce rapid increases in inbreeding. If mating can be controlled by natural or artificial means, however, several options are available by which the species or population may either avoid inbreeding for an initial period of time or allow some early generational inbreeding and perhaps reduce the rate of increase in F in the longer run.

For a finite set of parents, it is possible to avoid inbreeding by simply assuring that matings occur among individuals with no common ancestors until forced by the limitations of finite initial population sizes. At that time, matings among only the most distant cousins might be permitted. The number of generations without inbreeding and the closeness of relationships are both dependent on initial population size. Systems that permit no inbreeding in the early generations force the average relationship among individuals to build up rapidly. In these systems, inbreeding is avoided and population size is maintained by crossing among an ever-widening set of ancestries. These crosses cause the average relationsnip among the units of the breeding population to increase rapidly.

Alternatively, some controlled inbreeding may be used in each generation to more slowly and more uniformly allow the inbreeding to increase. Such systems as proposed by Kimura and Crow (1963), for example, cross trees $A \times B$, $B \times C$, and $C \times D$ in one generation, then $(A \times B) \times (B \times C)$, and $(B \times C) \times (C \times D)$ in the next, then $[(A \times B) \times (B \times C)] \times [(B \times C) \times (C \times D)]$ in the next.

etc. The relatively early onset of inbreeding is offset in later generations by slower increases in average relatedness.

For any regular mating system the relationships and inbreeding coefficients can be traced and the accumulation of homozygosity determined. Algebraic solutions of recursion relations can then be sought to determine rates of increases in homozygosity (Malécot 1969; Fisher 1965; Hayman and Mather 1953) by solving difference equations or by solving for the roots of the mating transition matrices. In particular, if any effects such as selection have deterministic effects on genotypic composition, the joint effects of selection and inbreeding can readily be traced.

If inbreeding is considered to occur within small subdivisions of populations, then the average rate of loss of heterozygotes will depend on the subdivision sizes, gene frequency and the balancing effects of selection, new mutations, and new migrant genes, as well as the forms in which the inbreeding occurs.

PREDICTING INBREEDING*

For deterministic models, the progress of inbreeding can be directly analyzed by considering that in each generation, the effects of assortative mating, selection, etc., can be modeled linearly to give relations between mating frequencies from one generation to the next. This is essentially the same method of analysis as used to predict age profiles in chapter 6. Here, it shows how recurrent systems can be analyzed for their long-term behavior. The frequency of various mating types can be written in a vector form $\underline{X}$; for example:

$$\underline{X} = \begin{Bmatrix} f(AA \times AA) \\ f(AA \times Aa) \\ f(Aa \times Aa) \\ f(Aa \times aa) \\ f(aa \times aa) \\ f(AA \times aa) \end{Bmatrix}$$

The form of mating and selection then determines the frequencies with which these matings in generation 0 will be allowed to leave progenies and regenerate matings in generation 1. Presumably, relative frequencies will differ among generations. Thus, with matings only within full-sib families, say the $AA \times AA$, only AA types are left, and hence all future matings are of $AA \times AA$ type. The $AA \times Aa$ type, however, generates equal numbers of AA and Aa genotypes and hence would generate mating types

*Graduate-level statistical training required for thorough understanding.

$AA \times AA$ ¼ of the time, $AA \times Aa$ ½ of the time, and $Aa \times Aa$ ¼ of the time. The process can be followed for each mating type to give us a matrix (A) of coefficients relating the matings in generation 0 to the matings in generation 1 as:

Generation 0	Generation 1					
	$AA \times AA$	$AA \times Aa$	$Aa \times Aa$	$Aa \times aa$	$aa \times aa$	$AA \times aa$
$AA \times AA$	1					
$AA \times Aa$	¼	½	¼			
$Aa \times Aa$	1/16	¼	¼	¼	1/16	⅛
$Aa \times aa$			¼	½	¼	
$aa \times aa$					1	
$AA \times aa$			1			

If selection was against AA genotypes such that the selective ratio of $AA:Aa$ was $1-s:1$, then the frequency of new matings as generated from the old would have to be modified such that $AA \times AA$ would occur $(1-s)^2$ of its former frequency, and $AA \times Aa$ matings would generate $AA \times AA$ matings $\frac{(1-s)^2}{4}$ of the time; $AA \times Aa \frac{(1-s)}{2}$ of the time; and $Aa \times Aa$ ¼ of the time. The remainder of the table would have to be similarly adjusted.

We can then determine the progress of inbreeding since

$$\underline{x}^{[1]} = A\underline{x}^{[0]}$$

or generally $\underline{x}^{[t+1]} = A\underline{x}^{[t]} = A^t\underline{x}^{[0]}$.

From matrix algebra we know that for any real, nonsingular matrix, A, there is a real matrix, U, and a diagonal matrix, D, such that $UDU^{-1} = A$. Therefore, for any power of A, say A^t, we can see that $A^t = UD^tU^{-1}$. Now let us try to find out what U and D are by first noting that if proportions among the mating frequencies $\underline{x}$ ever reach stabilities (which they will do for real, nonsingular A matrices), eventually, $\underline{x}^{[t+1]} = \lambda\underline{x}^{[t]}$, where λ is a constant of proportionality.

Since

$$A\underline{x}^{[t]} = \underline{x}^{[t+1]} = \lambda\underline{x}^{[t]},$$

$$A\underline{x}^{[t]} = \lambda\underline{x}^{[t]},$$

and $$(A - \lambda I)\underline{x}^{[t]} = 0,$$

where we can now see that λ is an eigenroot of the matrix A with an associated eigenvector $\underline{x}^{[t]}$. In fact, there are r such roots with associated vectors (where r is the rank of A) so that we can write

$$A\underline{x}_i = \lambda_i\underline{x}_i \text{ for each root } \lambda_i,\ i = 1, 2 \ldots .$$

Hence, we can put the various roots together,

$$A(\underline{x}_1, \underline{x}_2, \underline{x}_3 \ldots) = \lambda_1\underline{x}_1, \lambda_2\underline{x}_2, \lambda_3\underline{x}_3 \ldots$$

and letting the matrix $\underline{x}_1, \underline{x}_2, \underline{x}_3 \ldots = U$,

$$AU = U \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \ddots \\ 0 & & & \ddots \end{bmatrix}$$

or $$A = U(\Lambda)U^{-1}$$

where Λ is the diagonal matrix of λ_i's,

or $$A^t = U(\Lambda)^t U^{-1}$$

and

$$(\Lambda)^t = \begin{bmatrix} \lambda_1{}^t & & & \\ & \lambda_2{}^t & & \\ & & \lambda_3{}^t & \\ & & & \ddots \end{bmatrix}$$

Then $$x^{[t]} = U(\Lambda)^t U^{-1}\underline{x}^{[0]}$$

and we can now see that as t increases, the dominating effect that the largest λ_i and its $\underline{x}_i$ vector will have on the composition of $x^{[t]}$. Regardless of t, however, the technique is useful in determining the status of mating types at any time t, and the decay of heterozygosity to whatever stabilities may exist.

SELECTION

Selection itself of course has deterministic results which can lead to homozygosity or to the existence of intermediate gene frequency equilibria. In the simplest cases we only consider one locus, two alleles, and assume large populations without inbreeding, mutation, or migration effects. The results of simple, one-locus selection on genotypic and phenotypic performance were outlined in chapter 2 and require no further review. The principal problem discussed in achieving the state of stability or fixation determined by the genes was the accidents of sampling. However, any reasonable model of population behavior must also account for more complicated models of how environmental or breeding selection may vary and how genes may interact.

MULTIPLE ENVIRONMENT SELECTION

Several other kinds of effects can also lead to intermediate equilibria even with these simple models. For example, consider the case where variation in the selection effects over different parts of a population or over different generations is a better

model of the actual vagaries of population growth than a model with single constant-selection coefficients. We distinguish these variations from those in which all organisms must face the same kinds of variations in each life stage in a single generation and which can then be studied as if some average fitness of genotypes existed for the given pattern of variation. These latter types are fine-grained variations common to all individuals in contrast to coarse-grained variations in which individuals exist largely in one or another variant of the environment. Coarse-grained variations may be modeled in several ways. For example, Haldane and Jayakar (1963) show that if the frequency of environments which favor first one then another genotype change within certain limits and if the varying selection coefficients also exist within certain limits, that relative genotypic frequencies will also remain within certain intermediate limits. Even when the arithmetic average of all environments favors one genotype, if the alternate environments occur often enough that the geometric mean of selection coefficients favors the other, then selection will maintain an equilibrium. This condition might occur, for example, if a site is generally favorable for one form of growth but occasionally is very poor for that one form and relatively good for the alternate growth form. Thus, rare but severe extremes such as untimely frosts or fires may be enough to maintain alleles that are not advantageous in common types of competition but that may endow genotypes with exceptional resistance or reproductive capacity on those rare occasions when needed. In addition, Li (1967), Prout (1968), and Levene (1953) have theoretically shown that random mating of genotypes which undergo selection in different environments can lead to stable equilibria even if neither environment alone would lead to stability. For example, if trees intermate freely, some located in a site favoring AA and others in a site favoring $A'A'$, it is possible that the geometric mean of r, $\overline{r}_g$, is maximized at an intermediate gene frequency.

While the above simple models may accurately describe how some intermediate allelic frequencies are maintained by heterotic effects, it is also true that if many loci of this sort existed, the great majority of seedlings would contain many deleterious, homozygous loci. It is possible, however, that the detrimental effects of mutations and any consequent segregations of deleterious homozygotes can be modified by more exact models of how selection on fitness actually operates (Wallace 1968) and by synergistic gene actions (Crow 1968). Among forest trees, we lack much data on single gene actions, but since forestry has been a field of applied ecology, our knowledge of the detailed ecology and selective forces in forests can be used to investigate those sources of genetic variation.

The factor affecting gene frequency distributions which has received most interest is differential reproduction of genotypes such that the descendants of one genotype are more likely to have

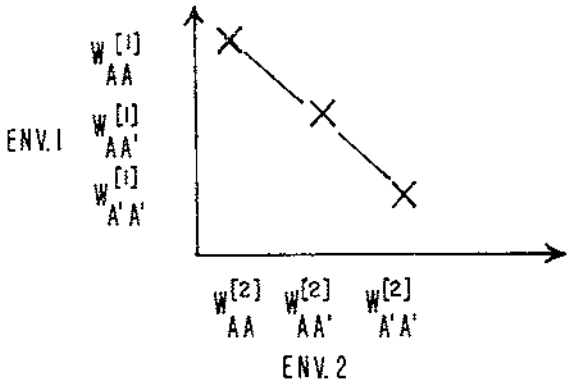
higher representation in future forests than now. In an earlier chapter, selection effects of relative success were described in terms of the joint effects of individual birth and survival processes, and average success was measured in terms of the largest eigenvalue of the expected Leslie matrix. Selective effects can also be exercised in the gametic stage of the life cycle since gene actions and differential success can also occur among gametes, but this is usually an exceptional type of selection for organisms in which the zygotic stage dominates the life cycle in both size and duration. Therefore, unless forms of meiotic drive or gametic selection are known to be present, we shall assume that selection operates on survival and descendant births, between zygote formation and death or such advanced age that actual death becomes irrelevant to the genetic composition of future forests.

In studies of genetic differences among forest trees related to distinct selective histories, there is abundant evidence that climatic and physiographic effects have differentiated populations of trees with respect to length of growing season, growth rate, etc. (Wright 1962; Stern 1964; Hamrick and Libby 1972). Within populations it can be expected that generally unfavorable alleles will be continually eliminated while others of some phenotypic effect but of equivocal fitness value might manage to coexist.

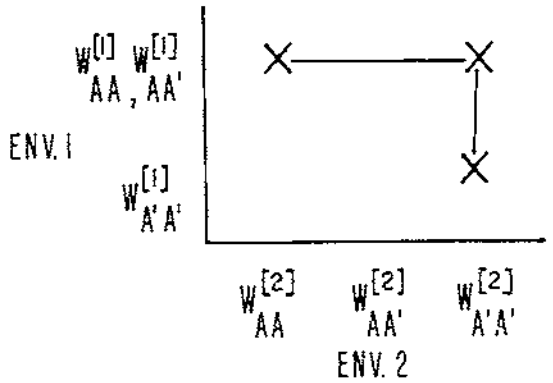
Among the many traits measured on forest trees, genetic differences may be caused by cryptic selective factors. Selection at one stage in the life of trees may affect zygotic frequencies of genes which cause phenotypic differences at some other life stage. Among trees within Douglas-fir stands, some genetic variance in height growth exists when trees are young, but there is little genetic variation for height growth within stands for older trees (Namkoong and others 1972). Such age-related selection can have additional unexpected results (Anderson and King 1970) on the effect of selection on the stability or instability of gene frequencies. While different genes operate on the same reactive processes at different stages of life, some genes can be expected to operate in a similar way in several life stages. However, the environment will have changed and the effect and selective advantage of those alleles can be quite different. Thus, some genes may have average effects over the lifetime of trees which cannot be predicted by their effects at any single stage of life. Hence, foresters in particular may have difficulties in defining effects for genes which act or affect fitness differently at different times. The environment can clearly change for trees, especially for pioneer species which are often established on open sites but which may have to regenerate under a closed stand in refuges in the next generation. In generalizing the effects of selection on evolution, MacArthur (1962) has emphasized that reproductive success can be far different under colonizing, low-density conditions than under closed community, high-density conditions (Pianka 1972). For most forest

tree species, a mixture of conditions will be present, sometimes requiring reproduction in a relatively stable environment where the site is already colonized and intraspecific competition is heavy. Here, the ability to survive crowding maximizes probabilities of success. For less stable environments which may be the result of newly created sites on the decay of established sites, greater success in colonizing sites may be required. The k and r types of selection as defined for logistic growth models in chapter 6 thus contrast in the kinds of behavior favored by selection. Such variations in effect imply that intraspecies competition may be important while selection for α types of interspecific competition may also be important. Therefore, selection on the basis of competitive behavior is complicated by intraspecific and interspecific density dependence. Important selection effects may then depend on the frequency with which alternate genotypes exist in the forest and hence will change according to the effects of previous selections. Thus, frequency-dependent selection can be a significant form of variation in selection pressures in forests. Thus, various forms of selection have been operating in our forests and require considerable analysis of gene actions and environmental variations to define how selection actually operates. While we generally consider simpler models of selection effects and generally have to assume that alleles have simple, average selective differences in some average environments at some average time of effect, the simplified models are not reasonable. The following theories then are useful only as first approximations, and even then require caution in applying to any of the complex phenomena of forest growth and development.

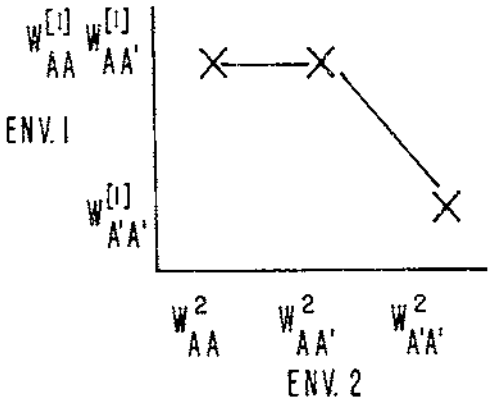
In the simple case in which one kind of environment almost exclusively exists with respect to genetically affected fitness, selection may be essentially unidirectional, and as long as the proper genotypes exist, a maximally adapted genotype would be expected to be fixed. If several kinds of environments exist, then the effect of selection may be either to still favor fitness in the more frequent environment or to favor an intermediate frequency as Levene's (1953) analysis would indicate. The outcome clearly depends on the fitnesses of each genotype in the various environments and on the frequency and sequences in which the environments exist. Levins (1968) describes the two factors in terms of a fitness of the genotypes in the environments and of a value function of the environments accounting for the frequency and sequence of the various environments. In the simplest case of three genotypes $AA{:}AA'{:}A'A'$ in two environments, the relative fitness values of each genotype (for example, W_{AA}) scored for each environment ($W_{AA}^{[1]}$, $W_{AA}^{[2]}$), and the pair of values is located in the two environmental dimensions. The same is true for $W_{AA'}^{[1]}$, $W_{AA'}^{[2]}$, and for $W_{A'A'}^{[1]}$, $W_{A'A'}^{[2]}$. A few cases are illustrated below in which the A allele is favored in environment 1 and A' is favored in environment 2.



Additivity in both environments.



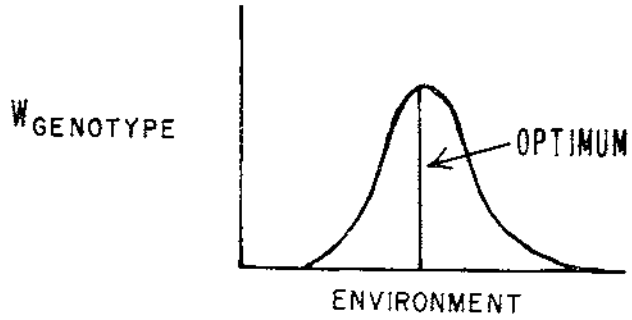
Dominance in both environments.



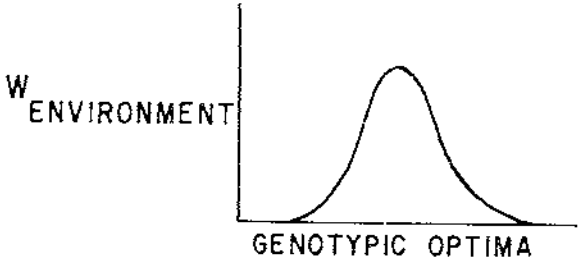
Dominance in environment 1
Additivity in environment 2

Alternate definitions of fitness sets by Levins (1968) create continuous functions connecting the points in a set of genotypes.

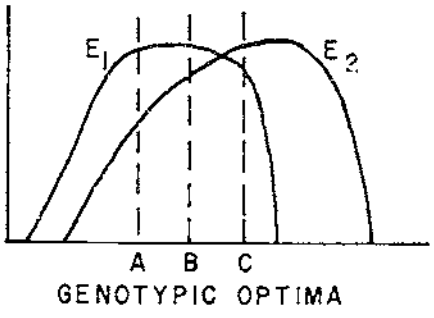
Consider that a genotype attains a maximum fitness at a particular state of an environmental variable and that its fitness declines as a quadratic or exponential function of the departure of the environment from that point. Thus, on an environmental scale, a genotype's fitness can be determined as a function of the environmental deviation from the optimum as in:



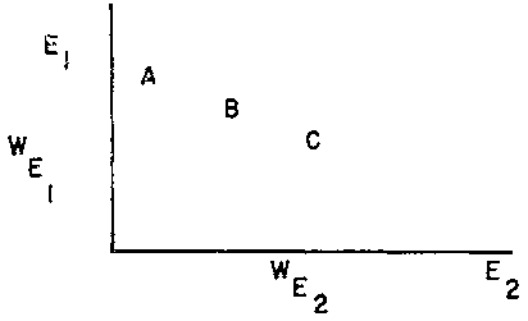
If several genotypes had differently located but similarly shaped curves, then at any particular level of the environmental variable, they would have an array of fitness values. Conversely then, any environment can be said to be a function of the genotypic fitnesses. Reducing characterization of the genotypes to their optimum locations, an environmental curve would attain a maximum for the genotypes whose optima coincide and would decline for genotypes with optima elsewhere. Thus, the environmental value would be a function of the distribution of optimum genotypes and may be similar in shape to the curve of genotypes on environments.



Then, two such environments would generate two curves for each genotypic optimum:

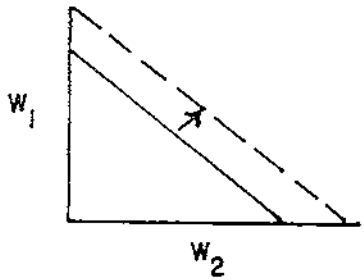


The two values of each genotypic point may alternatively be found in a two-dimensional mapping of environments 1 and 2 and the set of genotypes located in that space:



The closer the environments are, the more convex the set of points will be. If the environments are identical, the set degenerates to a line from the origin bisecting the angle of the axes. The further the environments are apart with respect to the variance of their optimum distribution, the closer the values around position "B" will be to the origin and hence the more concave the set will become in this central region.

Given any set of points on a particular straight or curved line describing the way the genotypes react to the environments, a second function can be drawn on the same graph indicating how the environment affects fitness. If a fine-grained environmental mixture exists, then the genotypic value simply depends on the average fitness as determined by the relative frequency of the environments. If p is frequency of environment 1 and $(1-p)$ the frequency of environment 2, then the fitness value would simply be $pW^{[1]}+(1-p)W^{[2]}$. If p is very high, then trees with high $W^{[1]}$ points have higher value than trees with low $W^{[1]}$ even if $W^{[2]}$ may be somewhat better. We can then draw lines of equal value for any given p to indicate the relative increases required in W_2 if W_1 decreases, such that fitness value remains constant:



$$A=pW_1+(1-p)W_2$$

$$W_1=\frac{A}{p}-\frac{(1-p)}{p}W_2$$

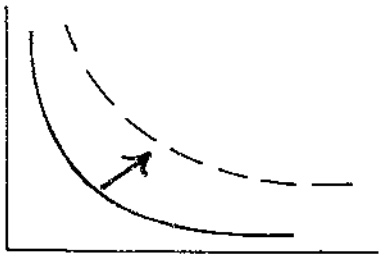
or $Y=a-bx$,

where $Y=W_1$, $a=\frac{A}{p}$, $b=\frac{1-p}{p}$, and $x=W_2$.

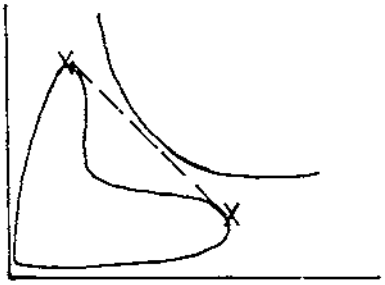
For a larger fitness value (A) to exist, the line of value must move to the upper right as indicated.

It is now clear that for a given set of genotypes, the type favored by selection will be the one which has the largest value function as determined by environment frequency, p. In all cases except for double dominance, it is possible that one or the other homozygote may be favored if the p or $1-p$ is very high.

A quite different result emerges, however, if a coarse-grained environment exists and the value is determined more by the sequential fitnesses that occur. Then, fitness is a multiplicative function of the environments and can be described by a function like $(W^{[1]})^p \cdot (W^{[2]})^{(1-p)}$ just as derived by Levene (1953) and by Li (1967) for the case when complete random mating occurs over all environments. In this case, the value function is proportional to a hyperbolic function, $p \log (W^{[1]}) + (1-p) \log (W^{[2]})$, as indicated below with the indicated direction of increases in value:



For coarse-grained environments then, it is far more difficult to favor a homozygote and selection can be expected to favor intermediate gene frequencies. Since trees do not move once germinated, they are susceptible to the accidents of seed fall and exist within neighborhoods on quite different soils and moisture regimes. Any major classification of environments would therefore be expected to include coarse-grained environmental variations. Hence, we might expect this type of environmental selection to predominate and determine the intermediacy of gene frequencies affected by soil factors. If the fitness set of genotypes has dominance relations such that the set is convex, then either of the fine- or coarse-grained environmental-value functions would most often yield optimal intermediate gene frequencies. For more nearly linear fitness sets, however, the hyperbolic-value function of the coarse-grained environments would more often yield intermediate gene frequency optima. This is Levene's (1953) case. For fitness sets which are concave, as drawn below, only the hyperbolic function can yield an intermediate optimum but one which can exist only if the extreme homozygotic types exist as mixtures.



Thus, intermediate gene frequencies can be favored by simple kinds of variable selection pressures without other factors like mutation or migration or even direct selection for intermediate forms. Genes can be maintained in intermediate equilibria by still other forms of selection, such as frequency-dependent selection as induced by competition or predator-prey, pathogen-host, or other frequency-dependent preferences. The possibility of frequency-dependent selection having strong, widespread effects in maintaining polymorphisms has been explored by Kojima (1971) and Kojima and Huang (1972) in *Drosophila* populations and for species with strong overlapping generations (Charlesworth and Giesel 1972). The effect of competitive interactions on maintaining selection for otherwise unexpected intermediate optima has been advanced by Mather (1969) and explored in trees by Huhn (1970c), who also concludes that even moderate levels of genotypic competition can lead to polymorphisms in trees.

A further extension of the simple model to include two loci also results in major differences in projected effects of selection that cannot be foreseen from the single-locus model. The simplest extension is to consider the loci to be independent in effect as well as in frequency. Linear, independent models can thereby be constructed for genetic distributions, and selection effects at one locus can be derived essentially independently of other loci. However, since genes do not act entirely independently and some exist in tight linkage groups, these simple, average-effect concepts may not be adequate, especially if gene frequencies change much and cause allelic combinations to interact nonlinearly.

MULTIPLE-LOCUS SELECTION*

In the single-locus case with simple environments, a fitness curve could be drawn as a function of gene frequency, and a $\frac{dp}{dt}$ curve could also be drawn as a function of gene frequency. In this case, the curves would show that the maximum fitness occurred at a value of p which was also at a point where $\frac{dp}{dt}=0$. Thus, any selectively induced changes in gene frequency would change

*Graduate-level statistical training required for thorough understanding.

the direction of maximizing fitness. However, in the multilocus case, the maximization of fitness at the same gene frequency as where $\frac{dp}{dt}=0$ is not a general result and the fitness surface itself is a function of zygotic frequencies, instead of being independent as in the single-locus case. These results have the same implications for breeding theory as for the genetics of natural populations.

If we first consider mating frequencies relating gametic frequencies for two generations with selection, we can derive the changes in gametic proportions and, hence, gene frequencies.

Let P_{AB} = frequency of AB gametes,

P_{Ab} = frequency of Ab gametes,

P_{aB} = frequency of aB gametes,

P_{ab} = frequency of ab gametes.

Assuming random mating in large populations without selection, the frequency of $AB\times AB$ unions will be P_{AB}^2, and these unions will yield all AB gametes. The frequency of $AB\times Ab$ unions will be $2P_{AB}P_{Ab}$, and these unions will yield AB and Ab gametes with equal frequency of $P_{AB}P_{Ab}$ each. Similarly, $AB\times aB$ unions occur with frequency $2P_{AB}P_{aB}$ and yield AB and aB gametes with $P_{AB}P_{aB}$ frequency each. The $AB\times ab$ unions occur with $2P_{AB}P_{ab}$ frequency and, with recombination at frequency r, yield Ab and aB gametes with frequency $rP_{AB}P_{ab}$ each. Without recombination, they yield AB and ab gametes with frequency $(1-r)P_{AB}P_{ab}$ each. The others are not affected by recombination. The only other kind of union that could give us new AB gametes would be recombinations of $Ab\times aB$ which occur with frequency $2P_{Ab}P_{aB}$. A recombination frequency r gives us AB half the time or with overall frequency $rP_{AB}P_{aB}$. Hence, the next generation's

$$P_{AB}^{[1]}=P_{AB}^{[0]}\left(P_{AB}^{[0]}+P_{Ab}^{[0]}+P_{aB}^{[0]}+P_{ab}^{[0]}\right)-r\left(P_{AB}P_{ab}-P_{Ab}P_{aB}\right)^{[0]}$$

$$=P_{AB}^{[0]}-r\begin{vmatrix}P_{AB}\,P_{AB}\\P_{aB}\,P_{ab}\end{vmatrix}^{[0]}=P_{AB}^{[0]}-rD^{[0]},$$

where $$D^{[0]}=\begin{vmatrix}P_{Ab}\,P_{Ab}\\P_{aB}\,P_{ab}\end{vmatrix}^{[0]}$$

All of the other frequencies can be similarly traced, and the collective result is:

$$\begin{pmatrix}P_{AB}^{[1]}\\P_{Ab}^{[1]}\\P_{aB}^{[1]}\\P_{ab}^{[1]}\end{pmatrix}=\begin{pmatrix}P_{AB}^{[0]}-rD^{[0]}\\P_{Ab}^{[0]}+rD^{[0]}\\P_{aB}^{[0]}+rD^{[0]}\\P_{ab}^{[0]}-rD^{[0]}\end{pmatrix}$$

The process is nonlinear since $D^{[0]}$ is a nonlinear function of the gamete frequencies:

$$\begin{aligned} D^{[1]} &= (P_{AB}P_{ab} - P_{Ab}P_{aB})^{[1]} \\ &= (P_{AB} - rD)^{[0]}(P_{ab} - rD)^{[0]} \\ &\quad - (P_{Ab} + rD)^{[0]}(P_{aB} + rD)^{[0]} \\ &= (P_{AB}P_{ab} - P_{Ab}P_{aB})^{[0]} \\ &\quad - (P_{AB} + P_{ab} + P_{Ab} + P_{ab})^{[0]} rD^{[0]} \\ &= D^{[0]} - rD^{[0]} \\ &= (1-r)D^{[0]} \end{aligned}$$

and $D^n = (1-r)^n D^{[0]}$.

Two features of the progress in D are notable, once a D value exists in a population, the disequilibrium D of gametic frequencies measured by the difference in coupling-repulsion-phase associations persists regardless of linkage. Even if unlinked loci recombine freely and $r = \frac{1}{2}$, D will decay at a rate of $\frac{1}{2}$ its former size per generation. The persistence of D is caused by the nonrandom association of alleles which cannot be immediately dissipated by any commonly known means, because increases in r decrease the yield of AB gametes from $AB \times ab$ unions but increase the yield of AB from $Ab \times aB$ unions. Hence, D persists. However, we can also see that each locus can immediately achieve Hardy-Weinberg equilibrium with random mating regardless of r because:

(1) AA homozygotes derive from $AB \times AB$ at P_{AB}^2 frequency, $AB \times Ab$ at $2P_{AB}P_{Ab}$ frequency, and $Ab \times Ab$ P_{Ab}^2 frequency.

The sum $= (P_{AB} + P_{Ab})^2 = P_A^2$, as required.

(2) Aa heterozygotes derive from $AB \times aB$ at $2P_{AB}P_{aB}$ frequency, $AB \times ab$ at $2P_{AB}P_{ab}$ frequency, $Ab \times aB$ at $2P_{Ab}P_{aB}$ frequency, and $Ab \times ab$ at $2P_{Ab}P_{ab}$ frequency.

$$\begin{aligned} \text{The sum} &= 2P_{AB}(P_{aB} + P_{ab}) + 2P_{Ab}(P_{aB} + P_{ab}) \\ &= 2P_{AB}(P_a) + 2P_{Ab}(P_a) \\ &= 2P_a(P_{AB} + P_{Ab}) = 2P_aP_A, \text{ also as required.} \end{aligned}$$

(3) The aa homozygotes derive from $aB \times aB$ at P_{aB}^2 frequency, $aB \times ab$ at $2P_{aB}P_{ab}$ frequency, and $ab \times ab$ at P_{ab}^2 frequency.

The sum $= (P_{aB} + P_{ab})^2 = P_a^2$, as finally required.

The same Hardy-Weinberg equilibrium would be reached by the locus B also. Hence, with random mating in large populations, all

individual loci immediately are in equilibrium even if the multiple-locus combinations are not.

We can now see what happens when epistatic types of selection are effective by imposing the following selection coefficients on the zygotic combinations:

	AB	Ab	aB	ab
AB	$W\left(\frac{AB}{AB}\right)$	$W\left(\frac{Ab}{AB}\right)$	$W\left(\frac{aB}{AB}\right)$	$W\left(\frac{ab}{AB}\right)$
Ab	$W\left(\frac{AB}{Ab}\right)$	$W\left(\frac{Ab}{Ab}\right)$	$W\left(\frac{aB}{Ab}\right)$	$W\left(\frac{ab}{Ab}\right)$
aB	$W\left(\frac{AB}{aB}\right)$	$W\left(\frac{Ab}{aB}\right)$	$W\left(\frac{aB}{aB}\right)$	$W\left(\frac{ab}{aB}\right)$
ab	$W\left(\frac{AB}{ab}\right)$	$W\left(\frac{Ab}{ab}\right)$	$W\left(\frac{aB}{ab}\right)$	$W\left(\frac{ab}{ab}\right)$

Marginal fitnesses of gametes can be defined by the fitnesses and frequencies of the zygotic combinations. For example:

$$W_{AB}=W\left(\frac{AB}{AB}\right)\times P\left(\frac{AB}{AB}\right)+W\left(\frac{AB}{Ab}\right)\times P\left(\frac{AB}{Ab}\right)$$
$$+W\left(\frac{AB}{aB}\right)\times P\left(\frac{AB}{aB}\right)+W\left(\frac{AB}{ab}\right)\times P\left(\frac{AB}{ab}\right).$$

Also, $W=W_{AB}\times P_{AB}+W_{Ab}\times P_{Ab}+W_{aB}\times P_{aB}+W_{ab}\times P_{ab}$.

If the table is simplified slightly by assuming that position effects are negligible, then $W\left(\frac{AB}{ab}\right)=W\left(\frac{Ab}{aB}\right)$, then the nine selection coefficients can be written more compactly as:

	BB	Bb	bb
AA	W_{22}	W_{21}	W_{20}
Aa	W_{12}	W_{11}	W_{10}
aa	W_{02}	W_{01}	W_{00}

Then the production of gametes for the next generation can be determined with the same assumption as before as:

$$P_{AB}^{[1]}=[W_{22}P_{AB}^{2}+W_{21}P_{Ab}P_{AB}+W_{12}P_{AB}P_{aB}$$
$$+W_{11}P_{AB}P_{ab}\ (1-r)+W_{11}P_{Ab}P_{aB}\ (r)]^{[0]}/\overline{W}^{[0]}$$

where $\overline{W}$ is the weighted mean fitness given above. Then,

$$P_{AB}^{[1]} = [P_{AB}^{[0]} W_{AB}^{[0]} - rW_{11}D^{[0]}]/\overline{W}^{[0]}$$

$$P_{Ab}^{[1]} = [P_{Ab}^{[0]} W_{Ab}^{[0]} + rW_{10}D^{[0]}]/\overline{W}^{[0]}$$

$$P_{aB}^{[1]} = [P_{aB}^{[0]} W_{aB}^{[0]} + rW_{01}D^{[0]}]/\overline{W}^{[0]}$$

$$P_{ab}^{[1]} = [P_{ab}^{[0]} W_{ab}^{[0]} - rW_{00}D^{[0]}]/\overline{W}^{[0]}.$$

In terms of changes in gametic frequencies:

$$\Delta P_{AB} = [P_{AB}^{[0]} \ [W_{AB} - \overline{W}]^{[0]} - rW_{11}D^{[0]}]/\overline{W}^{[0]}$$

$$\Delta P_{Ab} = [P_{Ab}^{[0]} \ [W_{Ab} - \overline{W}]^{[0]} + rW_{10}D^{[0]}]/\overline{W}^{[0]}$$

$$\Delta P_{aB} = [P_{aB}^{[0]} \ [W_{aB} - \overline{W}]^{[0]} + rW_{01}D^{[0]}]/\overline{W}^{[0]}$$

$$\Delta P_{aa} = [P_{ab}^{[0]} \ [W_{ab} - \overline{W}]^{[0]} - rW_{00}D^{[0]}]/\overline{W}^{[0]}.$$

For all changes in gamete frequencies to be zero, D must be zero and all W_{ij} must equal $\overline{W}$. If $D \neq 0$, then some $W_{ij} \neq \overline{W}$ at equilibrium frequencies, and hence some W_{ij} values can exist at equilibrium and would be associated with a nonzero disequilibrium D value. Several investigations of equilibria for different patterns of W_{ij} variations have shown that the number of stationary disequilibrium points for a given set of W_{ij} values may be as high as 7 (Karlin and Feldman 1969) while most studies also indicate that fitness is increased by having $D \neq 0$ (Kimura 1957; Lewontin and Kojima 1960; Bodmer and Parsons 1962; Wright 1967), and Moran (1964) has also shown that $\overline{W}$ is not generally maximized at equilibrium frequencies.

Thus, multilocus genetics presents qualitatively different problems and results than what might be otherwise expected from analysis of single-locus behavior. These differences must be considered if any large changes in gene frequencies or effects occur. In the short run, with large populations, the first approximation of largely independent genes in equilibrium may be reasonably accurate, but when long-term trends require large changes in allelic combinations or the introduction of allelic combinations with large effect on fitness, we cannot ignore the persistence of epistasis and disequilibria in gametic frequencies.

We should also recognize that dominance and overdominance relations may not translate directly from a phenotypic scale to a fitness scale. Wright (1935a) considered, for example, that a phenotypic optimum may exist such that phenotypes are maximally fit at some optimum environmental value and decline in fitness according to a quadratic function of the departures from that optimum. Thus, dominance and overdominance on the fitness scale are a quadratic function of the phenotypic scale. On a single-locus basis, overdominance on the fitness scale presents no new genetic features, but on a multiple-locus basis some dominance levels can

produce intermediate equilibria. Wright (1935a) first observed, if intermediate optima existed, that populations would evolve into a mixture of homozygotes at each locus such that the optimum phenotype was fixed and that at most one locus with overdominance on the fitness scale could remain segregating. However, Kojima (1959b) further showed that incomplete dominance at several loci could exist in an equilibrium, and Singh and Lewontin (1966) later showed that more equilibria could be expected if linkage disequilibria existed. Several possible patterns were explored by Wright (1969) for a two-locus system.

Investigations by Lewontin (1964) on five locus models indicated that large amounts of linkage disequilibrium can be generated from simple optimum phenotype models with epistasis on the fitness scale. Further study on even larger systems supported the long-term importance of epistasis on fitness and strong linkage disequilibria and led Franklin and Lewontin (1970) to conclude that, in the long run, the individual gene may be less appropriate than the whole chromosome for the study of evolution.

SELECTION-INDUCED POLYMORPHISM

Thus, selection models of reasonably simple form can yield populations with large amounts of genetic variance without recourse to mutation, migration, or other effects. The large reservoirs of genetic variance in many organisms can thus be explained by a variety of selection effects as well as other mechanisms. Even in commonly self-fertilizing organisms like oats (Allard 1965), there commonly exists sufficient outcrossing to generate reasonably large genetic variances. For tree species, there is a great deal of genetic variation in almost all traits studied. Due to the multiple mechanisms by which polymorphisms may be maintained, however, distinction of causes and the chances of the polymorphisms being more or less stable or subject to destruction are not known. Simple epistasis as well as variations in selective forces of several types can generate polymorphisms with some degree of stability. However, the study of systematic variations among populations may indicate the existence of some forms of selection.

We can conclude that many kinds of selection may well lead to polymorphisms among forest tree species, but the effects of one or two genes are difficult to detect and the existence of polymorphisms due to selection effects are difficult to establish. The major-gene phenomena associated with inbreeding depression of chlorophyll deficiencies have been easily observed, but they have not been associated in heterozygotes with any obvious heterotic effects. In fact, no simple cases of single-gene heterosis have been found in forest trees. The frequency of single-gene effects, such as chlorophyll mutants, is apparently more a function of mutation rates than selection for heterozygous effects (Franklin 1970b). Other single-gene effects at high frequency, such as oleoresin composition of slash pine (Squillace and Fisher 1966) and western

white pine (Hanover 1966) and flower color in Scots pine (Carlisle and Teich 1970), have not been associated with any observable selection pressures. However, for introduced pathogens, some disease resistances which would have current selective values may involve simple inheritance of one or two loci (Kinloch and others 1970), and their study over the next generation might be beneficial to our understanding of tree population dynamics.

If we consider only simple models of selection, we previously indicated that simple interactions of selection, migration, and mutation can also lead to equilibria. Another factor which we have not considered can lead to fixations—the operations of any of the above factors in small populations without free intermating. Many investigations have indicated that populations of trees may be of limited size (Sarvas 1963). Sakai (1971) concludes from examinations of isozymes in *Cryptomeria* that his population may exist with a great degree of isolation among adjacent stands. A more realistic model of forests would have to include the effects of the accidents of sampling as they affect small populations in some degrees of isolation. S. Wright's (1970) results indicate that variations in response to selection caused by partial isolations can be very potent in allowing populations to drift into gene combinations which may be useful for further evaluation in response to selection of types not possible in large random-mating populations. Interpopulational variances can thus be important to species evolution.

STOCHASTIC VARIATION

The general tendencies for average, population-wide effects caused by selection or other forces are not often exactly translated into events for individual trees. Average selective values for different genotypes may indicate the probabilities of survival on which trees will survive or reproduce "on the average." However, for a specific tree in a struggle for existence, life and death are qualitatively different events, and probabilities do not reflect the physiological nature of individual survivals. An individual lives or dies regardless of any probabilities. Similarly, for two individuals, they both may live or die, or one live and one die. Hence, the two may become 0, 1, or 2 individuals when counted again, even though we might expect an average of say 1.2 individuals. For larger populations, these accidents of sampling on an individual basis may lead to resulting numbers quite different from any average expectation. Thus 10 trees, each with a survival probability of 0.5, may all die, or some portion or all may live. While we may expect an average of 5 trees from very many 10-tree samples, any one such sample includes 0 to 10 survivors. If the process is sequentially repeated for the survivors, then again the outside limits remain 0 to 10 for any sample, though intuitively one may feel that somehow they should cluster around the average. Therefore, when variation in the occurrence of an event occurs,

the exact processes are no longer determined by simple equations giving some constant survival fractions. Variations in the occurrence of events lead to an interest not only in the average outcome of repeated samples but also in the variations that may occur among samples. Then, it may be desirable to estimate the probabilities of the population going to one extreme or another, or the probability of staying above some critically important level. In addition, if variations are very wide, they may be so important in affecting the outcome of population processes that a factor which might otherwise cause a stable equilibrium might become an unstabilizing force. We cannot review the entire theory of stochastic processes as applied to genetics problems, but rather we wish to review the basic concepts and tools useful in studying directed selection and other effects.

ANALYSIS OF STOCHASTIC PROCESSES*

To illustrate the kinds of analyses and effects of stochastic processes in population development, we recall the deterministic birth process of chapter 6. We originally considered that a growth rate or propensity to increase on an individual basis could be given a constant coefficient λ which we can consider to be a birth rate for each of n_t individuals alive at time t. Then the increase in n is

$\frac{dn}{dt} = \lambda n_t$, which when integrated yields

$$\log \frac{n_t}{n_0} = \lambda t,$$

or $$n_t = n_0 e^{\lambda t},$$

for the numbers (n_t) at time t as a function of the numbers (n_0) at some original time ($t=0$) and λ. The results were considered to be absolutely predictable and ratios of numbers of different types of individuals with specific λ or (r_a, r_A) rates were considered to provide relative selective values.

Now consider that λ actually expresses a tendency which is not exactly expressed by each tree, and hence in a small population we lack exact predictability. To estimate an average expectation and the variance which might be expected, we can analyze how variations may be generated and how they affect the probabilities of the possible numbers or gene frequencies and then compute the mean, variance, etc., of the population. The following example illustrates one approach to solving the problem for exponential growth by Bailey (1964), Feller (1957), and Pielou (1969). To simplify analyses, assume that λ is constant for some period and for some part of the population and that the probabilities of events are independent among all individuals. The probability of birth in a Δt time period then depends only on λ and the length of the

*Graduate-level statistical training required for thorough understanding.

time period and may be reasonably stated as:

$$Pr(\text{birth}) = \lambda \cdot \Delta t.$$

Then, starting at time t, for n trees we may model the birth probability in a Δt interval to be:

$$Pr(\text{births}) = \lambda \Delta t.$$

For there to be n trees after the time interval Δt, there would have had to have been n trees before the Δt interval and also no births; or $n-1$ trees before Δt and also 1 birth; or $n-2$ trees before Δt and also 2 births; etc., without death. If Δt is made sufficiently small and only 0 or 1 birth is possible in such a small Δt, then the probability of having n trees after the Δt interval is:

$$Pr(n;\ t+\Delta t) = Pr(n;\ t) \cdot Pr(0 \text{ births}) + Pr(n-1;\ t) \cdot$$
$$Pr(1 \text{ birth}) + Pr(n-x;\ t) \quad \cdot Pr(x \text{ births}) + \ldots.$$

Since $Pr(x \text{ births}) = 0$, for $x > 1$

$$Pr(n;\ t+\Delta t) = Pr(n;\ t) \cdot Pr(0 \text{ births})$$
$$+ Pr(n-1;\ t) \cdot Pr(1 \text{ birth}).$$

Taking $Pr(0 \text{ births}) = 1 - \lambda n \Delta t$

$$Pr(n;\ t+\Delta t) = Pr(n;\ t) \cdot (1 - \lambda n \Delta t) + Pr(n-1;\ t) \cdot n\Delta t$$

and hence,

$$\frac{Pr(n;\ t+\Delta t) - Pr(n;\ t)}{\Delta t} = -\lambda n \cdot Pr(n;\ t) + \lambda n \cdot Pr(n-1;\ t).$$

Allowing Δt to become infinitely small, the left-hand side becomes $\frac{d[Pr(n,t)]}{dt}$. Also, considering the initial state of the population, n may be zero after the initial Δt interval if n was zero before the Δt interval. Hence, for any initial size a, $Pr(a-1;\ t) = 0$. Therefore,

$$\frac{dPr(a;\ t)}{dt} = -\lambda a Pr(a;\ t),$$

and by integrating and using the boundary condition that $Pr(0;\ 0) = 1$, we derive $\log Pr(a;\ t) = -\lambda a t$ or $Pr(a;\ t) = e^{-\lambda a t}$. Then from the initial size, we can set up the probabilities of larger sizes at time t. The probability of being size $a+1$ at time $t+\Delta t$ is the sum of probabilities:

$$Pr(a+1;\ t+\Delta t) = Pr(a;\ t) \cdot Pr(1 \text{ birth})$$
$$+ Pr(a+1;\ t) \cdot Pr(0 \text{ births})$$

or
$$Pr(a+1;\ t+\Delta t) = Pr(a;\ t)\ (\lambda a \Delta t)$$
$$+ Pr(a+1;\ t)\ (1 - \lambda\ (a+1)\ \Delta t).$$

Using the same assumptions as before, we derive

$$\frac{d\,[Pr(a+1;t)]}{dt}+\lambda\ (a+1)\ Pr(a+1;\ t)=ae^{-\lambda at}$$

where $Pr(a;\ t)=e^{-\lambda at}$ on the right-hand side from the above derivation.

Integrating the left element by parts and using the initial condition that $Pr(a+1;\ t)=0$ at $t=0$, we can derive that:

$$Pr(a+1;\ t)=ae^{-\lambda at}\ (1-e^{-\lambda t}).$$

Repeating the process for $a+2$ would give us:

$$Pr(a+2;\ t)=\frac{a(a+1)}{2}e^{-\lambda at}(1-e^{-\lambda t})^2,$$

and in general for any $n>a$;

$$Pr(n;\ t)=\binom{n-1}{a-1}e^{-\lambda at}(1-e^{-\lambda t})^{n-a}.$$

We have thus derived the probability function for any n at time t as a function of a (or n_0) and λ and we can therefore determine the mean, variance, and higher moments of the process at that time. The traditional definitions of mean as $\sum_n nPr(n;\ t)$ and of variance as $\sum_n n^2Pr(n;\ t)-\mu^2$ can be derived for our case as:

$$\text{Mean}=ae^{\lambda t}$$

$$\text{Variance}=ae^{\lambda t}\ (e^{\lambda t}-1).$$

It can be noted that the mean is the same as the deterministic projection but that the variance can become very large at large t and increases faster than the mean.

A much simpler method of deriving moments is by deriving the whole sequence of probabilities of $n=1,\ 2,\ 3\ .\ .\ .$, etc., in the form of simple linear function. This is one of the few cases in which the method is often easier done than said since we use a transforming function which allows us to write the sequence of probabilities both as a linear function and as an exponential function. Such an expression is called a probability generating function (PGF) and can be derived for each probability function. For example, take a simple Poisson process in which the probability that a variable, x, is of size n can be stated as:

$$Pr(x=n)=\frac{e^{-\lambda}n}{n!},$$

where λ is the parameter of this distribution. Then its PGF can be stated as

$$\sum_{n=0}^{\infty} Pr(x=n)\cdot s^n,$$

where s is an indicator variable. Substituting the exponential function into the PGF, and multiplying by $e^{-\lambda s+\lambda s}$ gives:

$$e^{-\lambda}\frac{\Sigma(\lambda s)^n}{n!}\cdot\frac{e^{-\lambda s}}{e^{-\lambda s}}=\frac{e^{-\lambda}}{e^{-\lambda s}}\frac{\Sigma(\lambda s)^n e^{-\lambda s}}{n!}$$

Since $\frac{\Sigma(\lambda s)^n e^{-\lambda s}}{n!}$ is the summation over an entire Poisson distribution with parameter λs, the summation equals 1, and our PGF$=e^{-\lambda(1-s)}$. Since the exponential form can be expanded into a linear function with terms being the probabilities of $x=0, 1, 2 \ldots$ multiplied by $s^0, s^1, s^2 \ldots$, etc., $e^{-\lambda(1-s)}$ can be identified exactly with

$$Pr(x=0)s^0+Pr(x=1)s^1+Pr(x=2)s^2 \ldots, \text{ etc.}$$

Therefore, using

$$\frac{\delta(\text{PGF})}{\delta s}=Pr(x=1)+2Pr(x=2)s^1$$
$$+3Pr(x=3)s^2+\ldots,$$

we can see that evaluating $\frac{\delta(\text{PGF})}{\delta s}$ at $s=1$ gives us $\sum_n nPr(x=n)$ which is μ_n. We can also see that:

$$\text{Var}(n)=\frac{\delta^2(\text{PGF})}{\delta s^2}\Big|_{s=1}$$
$$+\frac{\delta(\text{PGF})}{\delta s}\Big|_{s=1}-\mu_n^2$$

and that $Pr(x=0)=\text{PGF}\Big|_{s=0}$.

While the derivations and uses of these functions will not be detailed, only a few theorems are required to develop the uses of the PGF for deriving probabilities of extinction, expected duration of processes before extinction or population growth explosion, etc. More complicated processes in which several types of organisms may be involved, such as age-dependent processes, genotypic arrays, etc., require only slightly more advanced consideration but can be useful in following forest processes (Namkoong and Roberds 1974).

STOCHASTIC GENETIC PROCESSES*

By determining the propensities for population growth among competing species or genotypes, or relative selective values, etc., and by considering that accidents of sampling occur according to some reasonable probability functions, we can examine the ex-

*Graduate-level statistical training required for thorough understanding.

pected outcome of population behavior in terms of averages, variances, or probabilities of extinction and duration of processes. The interest here is to derive a probability distribution of states of gene frequencies in a population or of the frequency of a single gene in several sampled populations. In addition to means and variances, we may be interested in the existence of gene-frequency stabilities for selection parameters, the absence of such stabilities, and the rate of change and approach to end points.

Where we previously considered Δt to be small, but let λn be of some size, let us now consider Δt to be small and the change in gene frequency also to be small. In a general diffusion process of this sort, we can consider that many loci may be varying in gene frequency in a population and that only random variations cause the frequency at a locus to vary. We can intuitively expect that most loci will remain near their original frequencies, though there are some small probabilities of rather distant drift, at least in a limited period of time.

Consider that as an approximation to the gene frequency process, a particle lies on a line and moves in small steps to the left or right according to how it is independently and randomly struck or otherwise moved from the right or left. If we let $p=Pr$(one step right), and $q=1-p=Pr$(one step left), the probability of being located 0, 1, 2 . . . r steps to the left or right after n steps is:

$$Pr(r;\, n+1)=p\cdot Pr(r-1;\, n)+q\cdot Pr(r+1;\, n).$$

Since r is the net result of several presumably independent steps to the right, say j, and the remaining $n-j$ steps to the left, then the probability of being at r is distributed binomially:

$$Pr(r;\, n)=\binom{n}{j}p^j q^{(n-j)}.$$

Making the changes in small steps, and making Δt also very small but such that the step sizes (Δx) remain such that $(\Delta x)^2$ approximates Δt, while both diminish towards zero in the limit, the net motion is $p\Delta x-q\Delta x$, and the variance is

$$p(\Delta x)^2+q(-\Delta x)^2-\mu^2=4pq(\Delta x)^2.$$

Then after a time period, t, in which $\frac{t}{\Delta t}$ independent events occur, the mean motion is $\frac{t}{\Delta t}(p-q)\ \Delta x$ with variance $\frac{4pqt(\Delta x)^2}{\Delta t}$.

Now allowing Δx and Δt to simultaneously become small, but such that $\frac{(\Delta x)^2}{\Delta t}=D$, and reparameterizing the mean change, we can express p as a function of the mean change M;

$$p = \tfrac{1}{2} + (\tfrac{1}{2D})\ \frac{M}{t}\Delta x$$

or letting $C = \dfrac{M}{2t}$,

$$p = \tfrac{1}{2} + (1/D)\,C\Delta x$$

and $$q = \tfrac{1}{2} - (1/D)\,C\Delta x$$

Then the mean is $2Ct$, and the variance is $2Dt$, since the probability distribution approximates normality for these $\dfrac{t}{\Delta t}$ independent trials, and for small Δx, $p \approx q \approx \tfrac{1}{2}$. Now, the probability function can be rewritten in terms of x steps and time as:

$$Pr(x;\ t+\Delta t) = pPr(x-\Delta x;\ t) + qPr(x+\Delta x;\ t)$$

We can now expand the left side around Δt deviations and the right side around Δx deviations using the general Taylor's series expansion:

$$f(y) = f(y_0) + \frac{\delta[f(y_0)]}{\delta y}\Delta y + \frac{\delta^2[f(y_0)]}{\delta y^2}\frac{(\Delta y)^2}{2!} + \frac{\delta^3[f(y_0)]}{\delta y^3}\frac{(\Delta y)^3}{3!} + \cdots$$

Then; $$Pr(x;\ t) + \frac{\delta Pr(x;\ t)}{\delta t}\Delta t + \frac{\delta^2 Pr(x;\ t)}{\delta t^2}\frac{(\Delta t)^2}{2!} \cdots$$

$$= p\left[Pr(x;\ t) - \frac{\delta Pr(x,\ t)}{\delta x}\Delta x + \frac{\delta^2 Pr(x,\ t)}{\delta x^2}\frac{(\Delta x)^2}{2!} - \cdots\right]$$

$$+ q\left[Pr(x;\ t) + \frac{\delta Pr(x,\ t)}{\delta x}\Delta x + \frac{\delta^2 Pr(x,\ t)}{\delta x^2}\frac{(\Delta x)^2}{2!} + \cdots\right]$$

$$= (p+q)\,Pr(x,\ t) + (q-p)\frac{\delta\,(Pr(x,\ t))}{\delta x}\Delta x + \frac{\delta^2\,(Pr(x,\ t))}{\delta x^2}(\Delta x)^2 + \cdots$$

Therefore, $$\frac{\delta Pr(x,\ t)}{\delta t}\Delta t + \frac{\delta^2 Pr(x,\ t)}{\delta t^2}\frac{(\Delta t)^2}{2!} = \frac{-(p-q)\,\delta Pr(x,\ t)}{\delta x}\Delta x + \frac{\delta^2 Pr(x,\ t)}{\delta x^2}(\Delta x)^2 + \cdots$$

and dividing by Δt and letting both Δx and Δt get small we can see:

$$\frac{\delta Pr(x,\ t)}{\delta t} = -(p-q)\frac{\Delta x}{\Delta t}\frac{\delta Pr(x,\ t)}{\delta t} + \frac{D\delta^2 Pr(x,\ t)}{\delta x^2}$$

or $$\frac{\delta Pr(x,\ t)}{\delta t} = -\frac{2C\delta Pr(x,\ t)}{\delta x} + \frac{D\delta^2 Pr(x,\ t)}{\delta x^2}.$$

This is the general Fokker-Planck equation, or diffusion equation, which Kimura (1957) applied to gene frequency drift by noting that c is proportional to the mean gene frequency change, M, and D is proportional to the variance, V, and hence can be written as:

$$\frac{\delta Pr(x,t)}{\delta t} = \frac{\delta^2}{\delta x^2}\frac{[V \cdot Pr(x,t)]}{2} - \frac{\delta}{\delta x}\left[M \cdot Pr(x,t)\right].$$

Wright (1940) derived these relationships in a slightly different way, but also sought the solution for the functional form of $Pr(x,t)$ which satisfied the equation and at the same time represented a stable condition where $\frac{\delta Pr(x,t)}{\delta t} = 0$.

The solution for t approaching infinity, $Pr(x, \infty)$ is

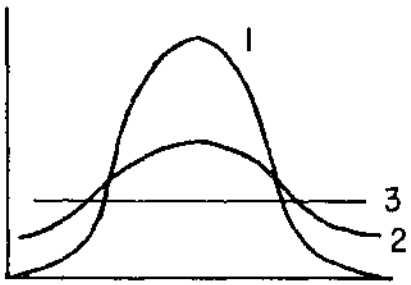
$$\frac{C}{V}\exp\left[-2\int\frac{M}{V}\,dx\right],$$

if the process can continue indefinitely. Thus, the expected distribution function for gene frequencies is dependent on average gene-frequency changes, M, and on the expected variance of such changes as may be induced by sampling variances.

We can derive the general behavior of the distribution function over time for certain types of conditions. For example, if the directional changes are small and the process starts at intermediate gene frequencies, then $M=0$ and $V=\frac{p(1-p)}{2N_e}$ and hence:

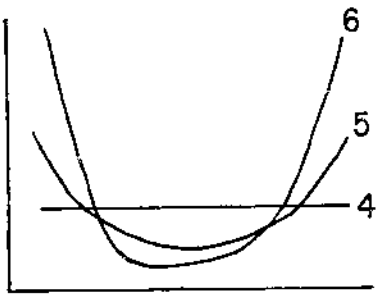
$$\frac{\delta Pr(p,t)}{\delta t} = \frac{1}{4N_e}\frac{\delta^2[p(1-p)Pr(p,t)]}{\delta p^2}$$

This expression yields a bell-shaped distribution for $Pr(p,t)$, which is dependent on t and which slowly changes to a rectangular distribution as the diffusion process makes all values of p equally likely for large values of N.



This solution, however, does not account for the fact that the end points of $p=0$ or 1 cannot ordinarily be escaped from. By adding these conditions, Kimura (1957) showed that as the intermediate

frequencies declined, the end points tended to absorb a high proportion of the distribution and the shape of distribution became more bowl-like:



Consider, for example, if $M=0$ and $V=\frac{p(1-p)}{2N_e}$, then the solution for

$$Pr(x, t) \approx \frac{2N_e}{p(1-p)}\, e^{-2/0dp} = \frac{2N_e}{p(1-p)}$$

which takes the inverse form of $p(1-p)$.

The solutions for the probability distribution function also depend on such factors as the initial starting point of the process and any directional effects in moving the average change towards an extreme or intermediate equilibrium and hence the process to some steady state. An initial low frequency might intuitively be expected to drift equally to the left and right but would tend to become fixed at zero more often than at one. Hence, a skewed distribution would be expected to develop for some period until all genes were fixed at zero more often than at one. Selection for the low-frequency allele would be expected to move the mean to the right and hence to develop a more symmetrical distribution for some period until the genes were more equally fixed at both ends. In this case, the tendency to drift rapidly to the left end points can be somewhat diminished by selection, though the relative effectiveness of drift versus selection will determine the ultimate success of selection in achieving a desired gene fixation.

To study the balance of forces between directional selection and diffusion drift, Kimura (1962) parameterized selection in a linear or additive model of gene effects as:

$$\left(1+\frac{s}{2}\right)AA : \left(1-\frac{s}{2}\right)A'A' : 1A'A.$$

Then for M, the change is the expected Δp value as before of

$$\frac{dp}{dt} = p(1-p)\,(r_A - r_{A'}),$$

and can be parameterized $p(1-p)s$. The variance in the change attributed to sampling variations in small populations is:

$$V(\Delta p) = \frac{p(1-p)}{2N_e}$$

Then to determine the probability that the favored gene A will ultimately be fixed, Kimura (1957) solved for the density functions and derived the relative probability of ultimately fixing A as:

$$\frac{\int_0^{p_0} \exp\left[-2\int \frac{sp(1-p)2N_e}{p(1-p)}\,dp\right]\,dp}{\int_0^1 \exp\left[-2\int \frac{sp(1-p)2N_e}{p(1-p)}\,dp\right]\,dp} = \text{ultimate probability of fixation (UPF)}$$

or $$\text{UPF} = \frac{1-\exp[-4N_e sp_0]}{1-\exp[-4N_e s]}$$

where p_0 is the initial-gene frequency.

Thus, the relationship between selection and the effective population size which determines drift is an intimate, multiplicative one in which large sizes of both N_e and s are required for successful selection. Note in this equation that if $Ns=0$, then $\text{UPF}=p_0$. However, if $Ns>0$, then

$$\text{UPF} = p_0 + 2Nsp(1-p) + 0\left[\frac{(2Ns)^2}{3}p(p-1)(2p-1)\right]$$

MUTATION, MIGRATION, SELECTION, AND STOCHASTIC VARIATIONS

Other effects can also be studied from this diffusion point of view such as mutation or migration having some influence on $M(\Delta p)$ in contrast to the deterministic models we previously developed. For example, if μ was the mutation frequency of $A \rightarrow A'$ and γ of $A' \rightarrow A$, then $M(\Delta p) = \gamma(1-p) - \mu p$ and with $V(\Delta p)$ as before,

$$\frac{M}{V} = 2N_e\left(\frac{\gamma}{p} - \frac{\mu}{1-p}\right)$$

$$\text{and } Pr(p,t) \approx \frac{2N_e}{p(1-p)} \exp\left[-4N_e \int \left(\frac{\gamma}{p} - \frac{\mu}{1-p}\right) dp\right]$$

$$= \frac{2N_e}{p(1-p)} \exp\left[-4N_e[\mu ln(1-p) + \gamma lnp]\right]$$

$$= \frac{2N_e}{p(1-p)} (1-p)^{4N_e\mu} p^{4N_e\gamma}$$

$$= 2N_e(1-p)^{(4N_e\mu-1)} p^{(4N_e\gamma-1)}.$$

One peculiarity of this form of the equation is that if $4N_e\mu = 1 = 4N_e\gamma$, or $\mu = \gamma = \frac{1}{4N_e}$, then $Pr(p, t)$ is proportional to $2N_e$ and is no longer a function of p and therefore $Pr(p, t)$ is uniform for all p. Thus, if mutation rates are on the order of $\frac{1}{4N_e}$ or if migration rates are on the order of one migrant per twice the N_e, then $m = \frac{1}{2N_e}$, and all gene frequencies can be equally likely. Therefore, such migration rates are sufficient to hold almost all allelic frequencies equally likely and therefore can maintain polymorphisms in spite of tendencies to drift to fixation.

On the other hand, if N or μ or m is large such that $4N_e\mu > 1$, then

$$Pr(p,t) \approx 2N_e(1-p)^{f_1(N_e,\mu)} p^{f_2(N_e,\gamma)}$$

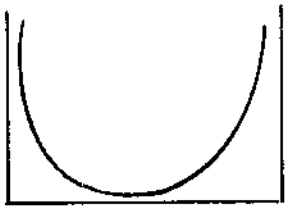
which is a function of $p(1-p)$ which has a peak in the intermediate values of p. In fact, at very high values of N the solution for $Pr(p, t)$ is proportional to:

$$Pr(p,t) \approx \mu[ln(4N_e\mu - 1)] p^{4N_e\gamma - 1} (1-p)^{4N_e\mu - 1}$$

$$+ \gamma[ln(4N_e\gamma - 1] p^{4N_e\gamma - 1} (1-p)^{4N_e\mu - 1}$$

which is close to zero everywhere except at $p = \frac{\mu}{\mu + \gamma}$, which was the deterministic solution we previously reached.

We can also see that if $N\mu$ is very small, that the solution for $Pr\ (p,t) \approx (1-p)^{-1}\ p^{-1}$ which is the reciprocal of the peaked quadratic function and has a deep concavity in the intermediate ranges of p:



Hence, if $N\mu$ or Nm is small, the random processes of drift fix the loci at one or the other allele. If N_e is low, drift occurs without much effect of otherwise effective mutation, or migration.

The effects that limited population sizes jointly exercise on selection and mutation-migration can also be examined in the diffusion model by hypothesizing that the mean change in gene frequency

$$M(\Delta p) = sp(1-p) - \mu p + v(1-p).$$

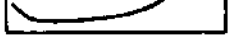
Using $V(\Delta p)$ as the $\frac{p(1-p)}{2N_e}$ drift function, we obtain

$$\frac{M}{V} = 2N_e\left(s - \frac{\mu}{1-p} + \frac{v}{p}\right),$$

and $$\int \frac{M}{V}dp = 2N_e s + 2N_e\mu \log\ (1-p) + 2N_e v \log\ (p),$$

so that $Pr(p, t) \approx 2N_e e^{4N_e sp} p^{4N_e v - 1} (1-p)^{4N_e \mu - 1}$.

This function now shows that the selection and mutation-migration independently have simple product-like effects on the probability distribution where a large s can push the distribution to the right-hand state as long as N is large enough and the effects of mutation or migration in increasing the alternate allele are not high. Consider, for example, that N_e is large but that both μ and v are small so that $4N_e\mu = 4N_e v \approx 1$. Then $Pr(p, t) \approx 2N_e e^{4N_e sp}$ which is an increasing function of p and hence tends to decrease the probability of having low gene frequencies and increases the probability of high frequencies:



It also indicates that, if s is on the order of $\frac{1}{4N_e}$, even at low initial p, the population can keep the favored allele. On the other hand, if $4N_e s$ is very low such that $e^{-4N_e sp} \approx 1$, then $Pr(p, t)$ can be largely determined by the balance between drift and the effects of mutation-migration. Then, introduction of new alleles through mutation-migration, even at a low frequency will maintain genetic variability. Thus, if migration may exist, very strong isolating mechanisms are required if populations are to diverge in the kinds of alleles carried. On the other hand, alleles can be lost under these conditions if breeding populations are developed with either low N_e or low s.

A question of great significance to our understanding of evolutionary mechanisms in tree populations is whether small popu-

lations have evolved and display divergent allelic frequencies or selection and migration among breeding units has been under uniform selection and homogenizing.

MIGRATION, INBREEDING, AND STOCHASTIC VARIATIONS

Since even small migration rates can affect the existence and stability of intermediate gene frequencies, we should consider the effects of migration in terms of its homogenizing effect on the population. It is conceivable, for example, that occasional pollen migrations may be enough to keep even relatively isolated yellow-poplar stands from diverging. Thus, if small interbreeding units exist, the divergencies in gene frequency which may otherwise exist may be nullified by even rare migrants. In subdivided populations with an overall average gene frequency of $\bar{p}$, the variance among such samples, $\mathrm{Var}(p)$, due to subdivision is $\mathrm{Var}(p) = F\bar{p}(1-\bar{p})$ according to Wahlund's principle as described earlier in this chapter. However, the variance among unit means due to limitations on random mating within units, for a unit with frequency p_i and the sampling variance of

$$\sigma_{p_i}^2 = \frac{p_i(1-p_i)}{2N_i}.$$

Averaged over all units (say k of them),

$$\sigma_{p_i}^2 = \frac{1}{k}\Sigma\frac{p_i(1-p)}{2N_i}.$$

If all N_i were equal,

$$\sigma_{p_i}^2 = \Sigma\frac{p_i(1-p_i)}{2Nk}.$$

Since the frequency of heterozygotes with the random-mating units is $2p_i(1-p_i)$, over the whole population the heterozygote frequency would be $\frac{2\Sigma p_i(1-p_i)}{k}$, which must also satisfy $2\bar{p}(1-\bar{p})(1-F)$. Therefore, we can write

$$\sigma_p^2 = \frac{\bar{p}(1-\bar{p})(1-F)}{2N}$$

We can now determine the relationship of F to migration rate, m, by defining F as the probability of identity by descent which increases by $\frac{1}{2N}$ each generation in a closed population. The remaining $\frac{2N-1}{2N}$ portion of the population presumably does not increase F, and hence

$$F = \frac{1}{2N} + \left(\frac{2N-1}{2N}\right)F'.$$

However, F can increase only among matings of nonmigrants, and at equilibrium

$$F=(1-m)^2\left[\frac{1}{2N}+\left(\frac{2N-1}{2N}\right)F\right]$$

or $$F=\frac{(1-m)^2}{2N-(2N-1)(1-m)^2}$$

Substituting this value into $\sigma_{p_T}^2$ yields:

$$\sigma_{p_T}^2=\frac{\bar{p}(1-\bar{p})(1-m)^2}{2N-(2N-1)(1-m)^2}\approx\bar{p}(1-p)/[4Nm+1],$$

for small values of m. Projecting the process from the diffusion equation approach gives the same approximate results for this model in which each subpopulation is considered to have an internal free, intermating system of size N, and a large external pool of the general species which feeds in migrants which carry the general species average of the gene frequency. The intimate relationship between population size and migration rates again indicates that small migration rates can have some effect in maintaining intermediate gene frequencies if N is large enough, but small N can permit fixations of any alleles though large sampling variations would exist in which alleles are fixed. A major problem exists in biology in general and in forestry in particular, however, in estimating both N and m, and hence in determining the effectiveness of migration in preventing genetic loci from becoming fixed. Pollen and seed dispersal studies, such as carried on by Sarvas (1963) and summarized by Wright (1962), are required but alone cannot satisfy the need for independent estimates of N and m. Furthermore, the effective flow rates on equilibria achieved are inevitably the resultant function of migration, balanced selections, population size, and mutation.

If the problem can be simplified to isolate just the N and m factors, however, we can begin to understand the effect of migrations on population evolution. Since the migration model used is clearly a very simple one, more realistically complicated models have also been developed to extend the projections of the relationship of migration to gene frequency distributions. An extension of the previously developed concept of population islands imbedded in a sea of the general average population is that each subpopulation is partially isolated but can share migrants with immediately adjacent neighbors to its left and right at one rate, m, and with the general population at another rate, m_∞. This model is called a steppingstone model by Kimura and Weiss (1964), who showed that the differentiation among populations expressed as σ_p^2 is approximately

$$\sigma_p^2=\bar{p}(1-\bar{p})/[1+4Nm_1(1-r)],$$

where r is a correlation factor of frequencies among adjacent

subpopulations and is a function of migrations such that it is approximately proportional to $\exp[-\sqrt{2m_\infty / m_1}]$. Increases in m_∞ at the expense of m_1 would decrease r and make σ_p^2 close to the island model solution. If m_1 is larger relative to m_∞, however, r increases and σ_p^2 increases. If the population is dispersed into subunits such that migratory exchange can occur in two dimensions, the r factor is proportional to $\exp[-k\sqrt{4m_\infty / m_1} / \sqrt{k}]$, where k is a function of step distance between subpopulations. In such a case, the r increases more rapidly for any increases in m_1 relative to m_∞, and hence, for the same amount of migration but split into more adjacent sources, the differentiation among population increases. For this two-dimensional dispersal case, Kimura and Maruyama (1971) have shown that only if Nm is less than one can population differentiation be expected. This is a slightly looser condition than previously suggested for the island model of populations. If $Nm>4$ the steppingstone model leads to a result very close to panmixia. More complicated cases and more complete analyses are derived by Weiss and Kimura (1965) which tend to show the same results.

NEIGHBORHOOD INBREEDING MODELS*

An alternate model of population dispersion and the effects of migration on maintaining genetic correlations among units is one in which the larger population is not actually physically discontinuous. Rather, the isolation may only be affected by higher or lower probabilities of neighbors being related than more distantly located trees. An effective isolation by distance may then exist, causing some tendency for more distantly dispersed trees to have drifted to different gene frequencies. The problem addressed by Wright (1940, 1943, 1949, 1951) was one of defining an effective population size, N_e, useful for computing an expected variance of gene frequencies, $V(p)$, among random neighborhoods of a larger continuous population. Since N_e is a function of the inbreeding coefficient, F, and the determination of F can be stated in terms of the probability of common parentage, we can eventually determine N_e as a function of the probability that a tree's parents are close enough to have been related. Assuming that the one-generation change in inbreeding occurs by the union of gametes from the same individual, and assuming that parents disperse offspring according to a normal distribution, the probability that an individual is inbred can be computed. If a uniform density of d trees per unit area exists and offspring dispersal follows a bivariate normal distribution with each direction having variance σ^2, then N_e can be derived to be $N_e=4\pi\sigma^2 d$ as follows.

Let the distance of a parent to an offspring be distributed with a probability function of

*Graduate-level statistical training required for thorough understanding.

$$y_i=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-x_i^2}{2\sigma^2}\right),$$

and also let the potential number of parents in an area, n, be a function of the dispersal variance. In particular, let $n\approx 4\sigma^2$ or a square of 2σ on a side. Then with n parents, in a density of $d=\frac{n}{4\sigma^2}$, each parent occupies an area of

$$\frac{1}{d}=\frac{4\sigma^2}{n}\int_{x_1-\frac{\sigma}{n}}^{x_1+\frac{\sigma}{n}}\int_{x_2-\frac{\sigma}{n}}^{x_2+\frac{\sigma}{n}} dx_1dx_2 \text{ in two dimensions.}$$

Then the probability that a gamete at a spot comes from a particular ith parent is $\frac{y_i}{d}$ and the probability of two such events is $\frac{y_i^2}{d^2}$. Then summing this for all parents gives us the probability of uniting gametes being identical by descent or

$$\frac{1}{N_e}=\sum_i \frac{4\sigma^2}{n}\int_{x_1-\frac{\sigma}{n}}^{x_1+\frac{\sigma}{n}}\int_{x_2-\frac{\sigma}{n}}^{x_2+\frac{\sigma}{n}} y_i^2dx_1dx_2=\frac{1}{n\pi}.$$

Since $n=4d\sigma^2$, $N_e=4\pi d\sigma^2$. Since a circle of radius $2\sqrt{d\sigma^2}$ has this area, N_e is equivalent to the number of trees within a circle radius $2\sqrt{d\sigma^2}$. If only one sex disperses its gametes, $N_e=2\pi d\sigma^2$, and N_e is equivalent to the number of trees within a circle of radius $\sqrt{2d\sigma^2}$. For trees distributed along one dimension as on a river bank, $N_e=\sqrt{4\pi\sigma d^2}$ (or $\sqrt{2\pi\sigma d^2}$ for single sex gametic dispersal) and hence is even smaller and generates a greater variance in gene frequency. Other formulations, such as given by Malécot (1969), give slightly different parameterizations but provide similar proportionate effects of density and dispersal distances.

This model is also clearly a crude approximation of actual dispersal patterns, as Wright (1962) has noted for pollen and seed dispersal patterns in many trees, but Wright (1969) indicates that relatively little divergencies would result from using the exponential functions of Bateman (1947). A principal result of use in investigating these models for forest trees, in which migration rates are determined by d and σ^2, is that if $d\sigma^2>16$ then little differentiation will occur in two dimensionally dispersed populations.

GEOGRAPHIC VARIATION IN FOREST TREES

In forest trees, various investigations indicate that while extremely long-distance pollen and seed migrations have been recorded and may be responsible for new colonizations, most indicate

that established forests do not disperse gametes very widely. Wright's (1962) summarization of dispersion studies generally indicates a very strong exponential decay for pollen dispersal regardless of pollen size, structure, vector, or wind velocity and hence that pollen flights are similar for trees as for even herbs and shrubs. Seed dispersal in pines is also restricted and strongly exponential in its decay rate (Pomeroy and Korstian 1949), (Boyer 1966). Wang and others (1960) also suggested that migration may be very low in established populations. Hence, migration may not be more effective than mutation in affecting gene-frequency variations.

However, populations of most subclimax species are not regularly distributed over time or space, and large variations occur in dispersal behavior. We can expect that neighbors do share gametes more often then distant trees or stands but that long-distance migrations are not rare (Sarvas 1967; Lanner 1966). There is some evidence to suggest that pollen flight characteristics are not the same at ground level as at upper-crown levels of established stands where the female flowers are often borne on wind-pollinated species. Air turbulences which lift the pollen into this area can also effectively carry pollen for many miles (Buell 1947; Boyer 1966), causing a more diffuse gametic dispersal than might be expected from ground-level studies. While some loss of viability should be expected (Sluder 1970), larger neighborhood sizes should still be expected.

The direct evidence on actual gene migrations for any tree species is meager. The possibilities of very restricted as well as widely dispersed panmixia exist for many tree species without clear data on the effective migration rates or population sizes. It is clear that some natural inbreeding can occur in pine stands (Squillace and Kraus 1963), and by tracing a mutant spruce allele, Langner (1953) indicated very restricted effective migration. Sluder's (1970) literature survey indicated that major migration effects were most often limited to seed dispersal—often by animal vectors but also by wind and water movements for those species so adapted. The evidence on stand-to-stand variations does not clearly support one hypothesis or the other. Most species display some stand-to-stand genetic differences, but most such differences only become clear over large distances and on ecologically distinct areas, and then the confounding effects of selection obscure the testability of migration hypotheses. Nevertheless, genetic differences within apparently contiguous stands have been shown to exist in isozyme differences by Conkle (1971), by Sakai (1971) and Sakai and others (1971), while population migration and differentiation paths have also been analyzed by tracing isozyme differences over widely separated stands of *Abies* (Matsuura and Sakai 1972). An exception may be *Populus deltoides* in the Mississippi Valley where stand differences over a wide geographic region are very small. This condition may indicate the strength of river migrations of seed.

A species in which isolation might be expected to have resulted in strongly divergent populations is yellow-poplar, which is often insect pollinated, possesses low stand densities among physically isolated stands, and displays no effective seed migration mechanisms. For the relatively lightly or currently unselected traits of leaf shape, Kellison (1970) found that the stand-to-stand variation within geographic regions, σ_s^2, was about the same as the family variance among trees within stands, σ_f^2. In contrast, the more selectively critical trait of early height growth had a much smaller σ_s^2 relative to its σ_f^2, indicating that selection may be effective in making stands genetically more uniform for some traits but that a relatively low N_e without selection can cause gene frequency divergences.

In more widely and uniformly dispersed loblolly pine, old-field stands tend to display larger N_e and lower inbreeding among open-pollinated families (Franklin 1968). While gametic unions often occur among related trees and even self-pollinations occur, the inbreds are highly susceptible to zygotic mortality before the seeds mature and through the first post-germination year. Thus, inbreeding is largely eliminated in the next generation. The effective gametic unions are then more likely to be from more distant migrants than what a pollen survey would indicate. There is, in fact, considerable evidence that inbreeding is genetically controllable itself and that outcrossing as well as isolation can have some selective advantages and can be selectively changed in populations (Levin and Kerster 1967, 1969, 1971).

Selection, therefore, considerably influences the expression of both the dispersive effects of inbreeding in small populations and the homogenizing effects of migration. This influence is perhaps most obvious with respect to migrations across species reproductive barriers or the sequences of development and destruction of reproductive barriers among partial isolates. Since fitness is only secondarily associated with reproductive barriers, population response to the relative advantages of isolation versus panmixia are often slow. Many of the hard pines of the Southeastern United States display a great ambiguity with respect to the status of species barriers. While the species concept is generally held to be valid, hybridization is so commonly observed that the barriers must be quite weak. However, introgression is rare among species and appears to be responsive to the vagaries of the selective advantages of hybrid phenotypes. In the Sonderegger pine hybrids, the parental species generally maintained themselves as distinct species; but during periods of great environmental disturbances in certain areas, hybrids and introgressants were very common. Upon restoration of a more normal environment, however, the newer generations were composed of no new hybrids and gave only weak evidence of any effective or long-lasting introgressive effects (Namkoong 1966c).

As a disruptive force, selection, on the other hand, might over-

come any homogenizing effects of migration. Haldane (1948) and Fisher (1950) have suggested that one effect may be that if there is a sharp reversal in selective advantage of alternate alleles, migration would tend to produce a gradual clinal shift over the segment of the population which surrounds the area of selective differences. A similar theoretical result was obtained by Hanson (1966) who investigated an island population within which the selective advantage was opposite to that in the general population surrounding it. In such islands, a minimal population radius of 6σ to 7σ is required to avoid being completely swamped by migrants from the general population.

Clines can be generated by selection in which an environmental series creates a series of populations with gradual changes in the optimum gene frequency. In such cases, migration may tend to homogenize the population. Alternatively, as Endler (1973) suggests, migration could be irrelevant to the attainment of a stable equilibrium series of the sequence of optimum selection gene frequencies even at high migration rates among adjacent populations. It is therefore far from clear in any specific case of clinal variations, what balancing of mechanisms between selection, migration, and drift these resultant clines may represent (Stern 1964). The existence of clines among serial tree populations can hardly be doubted as a general phenomenon (Langlet 1963; Sarvas 1970; Fryer and Ledig 1972), but its causes remain obscure in forestry. Measurement of presumed selectively different traits among adjacent populations along and across environmental gradients is a most useful approach to the problem of the relative effectiveness of selection versus migration (Hamrick and Libby 1972). However, we may not be observing stable population configurations in forest trees, and hence must consider that forest population strategies may require a slower, or less than immediate, response to selection pressures in unstable environments. It, therefore, remains an open question as to whether local populations possess even currently optimum gene frequencies (Namkoong 1969).

Thus, variation patterns among populations are the response of those populations to mutation and drift as confounded by variations in selection pressures over time and space and by migration patterns, all of which are further confounded by genetic changes in capacity to respond to mutagens, selection pressure, migration rates, and fluctuations in those effects. Therefore, the very simple concepts of static forest tree populations, in which some form of a balancing, optimal selection maintains optimal populations, are often likely to be excessively naive. While steady states may exist such that large changes in gene frequencies may not occur, such steady states are more likely to result from at least migration and drift forces in addition to selection, rather than from balancing selection alone. Selection for varying environments may itself have led to stable equilibria in gene frequencies, but migration and drift can significantly inhibit responses of perennial organisms to

local selection pressures (Antonovics 1968b).

We have investigated the models which have been developed for projecting the effects of the genetic factors influencing population evolution first as simple, independent effects. More extensive models of some effects such as multiple-locus selection and various migration patterns in themselves indicate that qualitatively different results may apply to population projections than would be indicated from simple effects alone. Thus, the existence of multiple-loci and varying environments lead to projected equilibria unexpected from single-locus or single-environment projections. The joint effects of selection, mutation, and migration were the simple resultant of their independent effects if it was assumed that no feedback mechanisms exist to modify mutation or migration rates as secondary selection effects. The dispersive effects of drift in populations of small size may be quite strong in some tree species and greatly complicate the effects of selection on gene frequency patterns. The general result of our considerations was that the combined effect of N_e and s, μ, or m had to be such that the product exceeded one for the deterministic effects to be significant. However, this result also required the assumption that the effects were constant and no feedback mechanisms existed such that the effects might change in response to indirect selection or to N_e itself. Thus, the independent effects were themselves quite complicated phenomena to which our first approximate models barely do justice. For joint effects, we often use very simple models which are known to be excessively naive and to require unrealistic assumptions on the independence of effects. Nevertheless, we have been able to develop models and test some hypotheses about them which rationalize our biological concepts and provide testable hypotheses.

Understanding the evolution of tree populations requires study of means, variances, and entire distributions of traits and genes as well as the distributions of the forces affecting them. The variations that have evolved and the genetic control of responses to variations determine the capacity of populations to respond to future variations in the genetic and external environments. The patterns of variation now present have been determined by the past factors which are rarely separable for convenient testing to determine the relative strengths of say selection versus migration in molding clinal variations. Nevertheless, the relative strengths of the independent forces acting on populations is an important first step to determine the causes of any steady or variable states, and hence the possible stabilities of those states with respect to variations in selection, migration rates, etc. Eventually, a more complete systems analysis will have to be made to account for secondary selective effects on mutation and migration rates as well as modifications of gene action to affect selection response itself. In addition, there is still considerable debate not only on the patterns of variation that may exist but on the presence of large amounts of genetic variation in all populations currently being studied for their isozyme gene frequencies. Regardless of what mixture of

factors may be responsible for the presence of those genetic variations, there seems little doubt that a considerable portion of the genes in a genome has a meaningful frequency of variants. Much of this variation may be among alleles which offer no selective differences. A few loci of detrimental mutants may be held in balance between selection and mutation rate as suggested by Kimura and Ohta (1971) or among alleles which at some time may have been advantageous to have been held in stable polymorphic systems, but which now exist without elimination (Robertson 1970). Which of these systems might explain the high rates of lethal equivalents in Douglas-fir (Sorenson 1969)? In addition, how many alleles are maintained in intermediate frequency by varying environments (Levins 1968), or by a combination of a few overdominant loci and high epistacy (Franklin and Lewontin 1970), or by frequency dependent selection (Kojima and Tobari 1969)? The stability and utility of those variations remains a critical question for those concerned with utilizing the system as evolved to respond to future variations of the environment and of possible breeding systems.

CHAPTER 10
THE VIEW AHEAD FOR FOREST GENETICS

At this time, foresters have the unique opportunity of initiating scientific breeding of forest tree species with relatively unmanaged gene pools. This does not mean that the gene pools have not been evolving to meet new evolutionary demands. They have been and will continue to do so. The quantification of natural and human-directed effects on the genetic composition and dynamics of the forests thus requires both description and model development. An understanding of natural evolutionary systems provides the basic data and model of a functioning genetic system. Such understanding is required for developing directed breeding systems and also helps provide a respect for the beauty and complexity with which the natural world operates. While the economic value of scientific breeding cannot be strongly doubted, the greater task is to build a more complete understanding of forest tree genetics so that better models and breeding methods can be devised.

Evolutionary genetics has provided the breeder with guidelines on how his natural populational source of materials may have evolved into partially segregated subpopulations and hence on how the structure of his breeding populations may be modified to maintain or to hybridize among any existing stand differences. A study of the evolution of his natural populations may also indicate how the genetic control of traits may change over the life cycle of individual trees or among environments. Thus, the bioengineering of the breeding populations may be a feasible alternative.

In addition, by understanding the evolution of forest systems, the forest geneticist can contribute to our general understanding of the variety of ways our general living systems have evolved. Indeed, the forest geneticist is obliged to expand our awareness of and appreciation for the integrated living systems within which we move. In particular, the evolution of nonequilibrium communities with pioneer species, including overlapping and unstable age classes, which grow in semi-isolated pockets of variable size and duration, characterizes common ecosystems of many forests and is a rich source of diverse life styles available for study.

In the past, as reviewed in this book, many difficult problems have been faced by a large number of scientists, and their first approximations have often been found to be most useful. Foresters

have primarily concerned themselves with ecological control and effects, and only in the last 20 years have they begun to extensively attack the underlying genetic system. Part of their efforts have been directed to understanding and parameterizing environmental variations and the interactions between sites and genes. The particularly difficult problems of understanding the nature of genotype × environment relationships and choosing the best genotypes for a given variety of environments have been at least partially solved by present methods. The main problem of determining the size of the interactions may not require sophisticated testing, but determining the form of the response to sampled sets of environmental variations requires a high degree of skill in design and analysis. The additional problem of determining whether all individuals in a breeding population should exhibit homeostasis or if a mixed populational homeostasis should be sought also requires a high degree of genetic sophistication.

The choice of a breeding system is dependent on the genetic knowledge of the breeder and the kinds of genetic variations which he can use. A vast array of opportunities to further increase gain is available to the breeder who may choose to use breeding methods capable of utilizing any dominance and epistatic types of gene action which may be present. Various kinds of pure breeding or hybrid breeding systems may be appropriate, and the separation of the breeding population from the seed-production orchard further expands the operating options of the breeder to create maximum long- and short-run gains. His knowledge of theoretical quantitative genetics also provides guidelines on how he may compromise between selection intensity and breed population size. It also suggests methods of subdividing populations to achieve greater flexibility in long-term breeding programs.

With all of these means by which the trained quantitative geneticist can affect breeding practices, he is very likely to achieve greater gains than a person who relies on uninformed intuition. However, recognition of the great value that genetic analyses and the subsequent synthesis of breeding programs may have does not imply that many problems do not remain before breeders can be fully effective. Some problems are difficult to solve because we lack experience with forest trees, and although a general theory may be available, forestry data may be lacking. Other problems are particularly difficult because the basic theory is inadequate to meet our needs. For example, optimal breeding programs must account for interactions between trees and other organisms, which themselves are affected by the breeding system. If competitive effects among genotypes are important, as described by Sakai's (1955) distinguished series of studies, then genetic systems, such as described in a series of studies by Huhn (1970c) or by Griffing (1967), will have to be extended and applied to tree breeding programs. If breeding affects insect or disease pathogen populations, then we will have to modify the breeding program to achieve gain while the pathogen evolves in some minimally harm-

ful directions. Other problems which will require theoretical developments include the nature of provenance differences and their use in breeding programs. Problems for which a theoretical foundation may be adequate, but where experience with trees is lacking, include long-term selection studies in which inbreeding can adversely affect selection response. Both theoretical and practical problems of great complexity remain for tree breeders much as they do for animal breeders (Barker 1967).

In spite of these major problems in achieving maximal gains, the remarkable success of modern breeders lies in the application of genetic principles to the studied and bred organisms. The very simple models of gene action used have thus led to substantial improvements through selection and breeding of genotypic compositions. More accurate and precise testing methods and wiser choices of traits and materials for breeding have vastly increased breeding efficiency. Perhaps the most significant contribution of quantitative genetic theory to tree breeding has been the adoption of simple gene models. By applying known principles of gene action, some predictive power has been achieved for a variety of breeding procedures. The advance from no genetic model to simple models has thus fostered considerable economic gains. However, as old revolutionaries tend to become the new conservatives, new models tend to generate their own orthodoxies and impose their own limits on concepts of how breeders may control future populations. In somewhat oversimplied terms, it may be argued that the principal effect of quantitative genetics has been to apply a linear model of gene effects to many genetic loci. As a result, breeding theory is largely the adaptation of linear statistical models to crossing and breeding experiments. The fact that the simple models are not truly adequate may be well recognized, but our thinking remains limited to approaches available with linear models.

The simple models were never intended to include such complicating effects as nonindependence among alleles and among loci. However, since linkage groups do exist and inbreeding does occur, the genetic models are not inclusive of these possibly significant effects and may not be an adequate basis for predicting selection effects. The smaller the effective population size and the fewer genes of large effect there are, the greater will be the discrepancies. While some theoretical studies have been conducted (for example, Gill 1965b; Latter 1966), our thinking is largely restricted to how well the simple models behave under nonindependence conditions, and little has been successfully done to develop more adequate models for selection. The major effects of epistasis are similarly inadequately modeled, and except for computer simulation studies on multiplicative gene-action models (Franklin and Lewontin 1970), the combined effects of epistasis, linkage, and small effective population sizes under selection have not been adequately studied. Thus, any realistic mixture of effects with varying gene frequencies are poorly modeled. Studies of

changes in selection intensity, especially of the kind that switches from positive to negative over periods of time greater than one generation, and under competition or other frequency-dependent effects, have only been studied under the simplest genetic models but can obviously be of major importance to breeding theory. Also, as gene frequencies change under the influences of selection, correlated changes occur in the genetic variances (Rawlings 1970). They can also be expected to occur in the effects of the genes themselves, because the entire genetic background of the individuals in the breeding populations is changing.

It thus seems clear that we should at least determine the adequacy of the presently used models rather than to assume their restrictive definitions. For example, instead of defining dominance effects as deviations from the linear, additive effects, we might construct models where dominance and epistatic second-order effects are defined first. Then, fitting alternate models may indicate the adequacy of one or the other model, and means and variances defined according to the most appropriate mode. Selection effects may then be more easily modeled if second-order effects are important. Thus, we need not rely entirely on linear models if others fit better, and we could begin to develop models of quantitative gene action which are less restricted than our simpler ones.

The use of linear economic models in forestry is obviously a serious limitation. Yet, breeding theories on several traits require not only linear economic models but also independence of value among traits. Clearly, nonlinear and dependent models of value must be developed and used for truly adequate evaluation.

Similarly, the problems of predicting selection effects or modeling the evolution of gene systems need not be restricted to such assumptions as the constancy of gene effects or other parameters of selection, migration, etc. These parameters do vary, sometimes randomly and sometimes in correlated patterns. Thus, the diffusion theoretic basis for selection under small population sizes, as developed by Kimura (1964), may well be expanded to include nonrandom parameter variations of certain forms and could encompass the variations suggested by Levins (1968). The theoretic basis for selection and breeding theory is thus likely to expand to include variations in selection pressure, population size and inbreeding, and migration or gene-pool exchange rates. The control of such variations in multiple, small population replicates may be achievable.

One type of change which breeders may anticipate is in the form of the physical and economic environments within which the future commercial breeds must operate. Since variations of uncertain form and extent must be anticipated, the problem for breeders is to determine not only how their breed populations change if selection parameters change, but more importantly, what kind of breed populations should they construct to yield maximum value in an uncertain future. Thus, new concepts of optimum population forms are required which will include vari-

ance control and higher moment specifications, in addition to our present concern with mean values and maximization principles only. Newer techniques of mathematical programming will likely be used to define optimum selections under conditions of both economic and environmental uncertainty (Namkoong 1970b).

While tree breeding itself is explosively advancing, and many problems require solution, the present state of the theoretical art can be described as having reached a plateau of development. We have used linear models with tremendous success in advancing both our concepts of breed control and development as well as in vastly increasing operating efficiencies. There remain many problems in which linear models of independent gene actions can still be applied for guidance in optimizing breeding practices. However, forest geneticists cannot afford to assume that the present models adequately define all important kinds of genetic variations. Hence, geneticists should not be limited in their concepts to the restrictions and limitations of linear models. The biological questions of how breeds actually develop and of how genes actually interact to give responses to selection have not been solved by the application of linear statistical models to breeding theory. The models used have provided a basis for testing certain hypotheses on the existence of forms of genetic variances and the efficiency of breeding. The next step in the development of a better theory is to conduct experiments on the adequacy of the models and to propose more inclusive or more accurate models. Then the scientific process from forming a model to testing the model, to observation and proposal of better models can proceed.

At this time, the overriding need in quantitative genetics of forestry is for biologists and breeders to use and understand the simple models, to test their adequacy, and to propose models which more closely fit the biological facts. We require tests of how inbreeding affects response to selection when both inbreeding depression and small population sizes have some effect. For gene actions in populations with hybrid-crossing systems, we require tests on the nature of any heterotic responses. For long-term selection programs, we require experiments on the changes developed in the variances and in the gene effects themselves. The nature of environmental interactions and age changes still requires far better definition than now available, and the optimal use of environmental and economic variations by breed populations requires more imaginative solutions than those developed thus far. Therefore, the role of quantitative genetics in the future is to apply quantitative analyses to more inclusive and more accurate genetic models, as suggested by experimental testing and the fertile minds of foresters.

LITERATURE CITED

Anonymous.
1961. Status and methods of research in economic and agronomic aspects of fertilizer response and use. Natl. Acad. Sci., Natl. Res. Counc. Publ. 918, 89 p.

Aalders, L. E.
1966. A recurrent selection program for perennial crop species designed to minimize inbreeding. Can. J. Genet. Cytol. 8:293–295.

Adams, W. T., J. H. Roberds, and B. J. Zobel.
1973. Inter-genotypic interactions among families of loblolly pine. Theor. & Appl. Genet. 43:319–322.

Allard, R. W.
1960. Principles of plant breeding. 485 p. John Wiley & Sons, Inc., New York.

Allard, R. W.
1965. Genetic systems associated with colonizing ability in predominantly self-pollinated species, p. 49–76. *In* H. G. Baker and G. L. Stebbins [eds.], Genetics of colonizing species. Acad. Press, New York.

Allard, R. W., and J. Adams.
1969. The role of intergenotypic interactions in plant breeding. Proc. XII. Int. Congr. Genet. 3:349–370.

Allard, R. W., and A. D. Bradshaw.
1964. Implications of genotype-environmental interactions in applied plant breeding. Crop Sci. 4:503–508.

Allard, R. W., J. Harding, and C. Wehrhahn.
1966. The estimation and use of selective values in predicting population change. Heredity 21:547–563.

Anderson, R. L., and T. A. Bancroft.
1952. Statistical theory in research. 321 p. McGraw-Hill Book Co., New York.

Anderson, T. W.
1958. An introduction to multivariate statistical analysis. 374 p. John Wiley & Sons, Inc., New York.

Anderson, W. W.
1971. Genetic equilibrium and population growth under density-regulated selection. Am. Nat. 105:489–498.

Anderson, W. W., and C. E. King
1970. Age-specific selection. Natl. Acad. Sci. Proc. 66:780–786.

Antonovics, J.
1968a. Evolution in closely adjacent plant populations. V. Evolution of self-fertility. Heredity 23:219–238.

Antonovics, J.
1968b. Evolution in closely adjacent plant populations. VI. Manifold effects of gene flow. Heredity 23:507–524.

Armitage, F. B., and P. M. Burrows.
1966. Preliminary heritability estimates for *Pinus patula* in Rhodesia. Rhod. Zambia Malaya J. Agric. Res. 4:111–117.

Ayala, F. J.
1971. Competition between species: frequency dependence. Science 171:820–824.

Ayala, F. J.
1972. Competition between species. Am. Sci. 60:348–357.

Bailey, N. T. J.
1964. The elements of stochastic processes. 249 p. John Wiley & Sons, Inc., New York.
Baker, L. H., and R. N. Curnow.
1969. Choice of population size and use of variation between replicate populations in plant breeding selection programs. Crop Sci. 9:555–560.
Bal, B. S., C. A. Suneson, and R .T. Ramage.
1959. Genetic shift during generations of natural selection in barley. Agron. J. 51:555–557.
Bannister, M. H.
1965. Variation in the breeding system of *Pinus radiata*, p. 353–374. *In* H. G. Baker and G. L. Stebbins [eds.], Genetics of colonizing species. Acad. Press, New York.
Barber, J. C.
1961. Growth, crown form, and fusiform rust resistance in open-pollinated slash pine progenies. Sixth South. Conf. For. Tree Improv. Proc. 1961:97–104.
Barber, J. C.
1964. Inherent variation among slash pine progenies at the Ida Cason Callaway Foundation. U.S. For. Serv. Res. Pap. SE–10, 90 p. Southeast. For. Exp. Stn., Asheville, N.C.
Barker, J. S. F.
1967. Modern problems of population genetics in animal husbandry. Der Zuchter 37:309–323.
Bateman, A. J.
1947. Number of s-alleles in a population. Nature 160:337.
Becker, W. A.
1967. Manual of procedures in quantitative genetics. 2d. ed., 130 p. Wash. State Univ. Press, Pullman.
Beineke, W. F.
1967. Genetic variation in the ability to withstand transplanting shock in loblolly pine (*Pinus taeda* L.). Diss. Abstr. 27:4197–B.
Bey, C. F.
1970. Geographic variation for seed and seedling characters in black walnut. U.S. For. Serv. Res. Note NC–101, 4 p. North Cent. For. Exp. Stn., St. Paul, Minn.
Bhagwat, S. G.
1963. Comparison of first-year wood fibers among different poplar clones. Third Cent. States For. Tree Improv. Conf. Proc. 1962:5–14.
Billingsley, P.
1961. Statistical methods in Markov chains. Ann. Math. Stat. 32:12–40.
Bingham, R. T., R. J. Olson, W. A. Becker, and M. A. Marsden.
1969. Breeding blister rust resistant western white pine. V. Estimates of heritability, combining ability, and genetic advance based on tester mating. Silvae Genet. 18:28–38.
Blackith, R. E., and R. A. Reyment.
1971. Multivariate morphometrics. 412 p. Acad. Press, New York.
Bodmer, W. F., and P. A. Parsons.
1962. Linkage and recombination in evolution, p. 1–100. *In* E. W. Caspari and J. M. Thoday [eds.], Advances in genetics. Vol. 11. Acad. Press, New York.
Bogyo, T. P.
1964. Coefficients of variation of heritability estimates obtained from variance analyses. Biometrics 20:122–129.
Bohren, B. B., W. G. Hill, and A. Robertson.
1966. Some observations on asymmetrical correlated responses to selection. Genet. Res. 7:44–57.
Box, G. E. P., and H. L. Lucas.
1959. Design of experiments in non-linear situations. Biometrika 46:77–90.

Boyer, W. D.
1966. Longleaf pine pollen dispersal. For. Sci. 12:367–368.

Braaten, M. O.
1965. The union of partial diallel mating designs and incomplete block environmental designs. N.C. State Univ. Inst. Stat. Mimeogr. Ser. 432, 77 p. Raleigh.

Buell, M. F.
1947. Mass dissemination of pine pollen. (Abstr.) J. Elisha Mitchell Sci. Soc. 63:163–167.

Burdon, R. D., and C. J. A. Shelbourne.
1971. Breeding populations for recurrent selection: conflicts and possible solutions. New Zealand J. For. Sci. 1:174–193.

Burkill, H. M.
1959. Large scale variety trials of *Hevea brasiliensis* Muell-Arg. on Malayan estates 1934–53. Rubber Res. Inst. Malaya J. 16:1–37.

Burrows, P. M.
1967. Seed orchard systems for tree breeding. Rhod. Zambia Malaya J. Agric. Res. 5:273–280.

Burrows, P. M.
1970. Coancestry control in forest tree breeding plans, p. 27–36. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Butcher, A. C., J. Croft, and M. Grindle.
1972. Use of genotype-environmental interaction analysis in the study of natural populations of *Aspergillus nidulans*. Heredity 29:263–283.

Byrd, B. W., Jr., D. F. Matzinger, and T. J. Mann.
1965. Intergenotypic competition in flue-cured tobacco. Tob. Sci. 9:12–16.

Callaham, R. Z., and A. A. Hasel.
1961. *Pinus ponderosa*. Height growth of wind-pollinated progenies. Silvae Genet. 10:33–42.

Campbell, R. K.
1964. Recommended traits to be improved in a breeding program for Douglas fir. Weyerhaeuser Co. For. Res. Note 57, 19 p.

Carlisle, A., and A. H. Teich.
1970. The Hardy-Weinberg law used to study inheritance of male inflorescence color in a natural Scots pine population. Can. J. Bot. 48:997–998.

Charlesworth, B., and J. T. Giesel.
1972. Selection in populations with overlapping generations. IV. Fluctuations in gene frequency with density-dependent selection. Am. Nat. 106:402–411.

Clarke, B.
1972. Density-dependent selection. Am. Nat. 106:1–13.

Clausen, J., D. D. Keck, and W. M. Hiesey.
1948. Experimental studies on the nature of species. III. Environmental responses of climatic races of Achillea. Carnegie Inst. Publ. 581, 129 p. Washington, D. C.

Cockerham, C. C.
1954. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. Genetics 39:859–882.

Cockerham, C. C.
1956. Effects of linkage on the covariances between relatives. Genetics 41:138–141.

Cockerham, C. C.
1959. Partitions of hereditary variance for various genetic models. Genetics 44:1141–1148.

Cockerham, C. C.
1963. Estimation of genetic variances, p. 53–94. *In* W. D. Hanson and H. F. Robinson [eds.], Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc. Publ. 982. Washington, D.C.

Cockerham, C. C.
1967. Group inbreeding and coancestry. Genetics 56:89–104.

Cockerham, C. C.
1970. Avoidance and rate of inbreeding, p. 104–127. *In* K. Kojima [ed.], Mathematical topics in population genetics. Springer-Verlag, New York.

Cockerham, C. C.
1971. Higher order probability functions of identity of alleles by descent. Genetics 69:235–246.

Cockerham, C. C.
1973. Analyses of gene frequencies. Genetics 74:679–700.

Cockerham, C. C., and D. F. Matzinger.
1966. Simultaneous selfing and partial diallel test crossing. III. Optimum selection procedures. Aust. J. Biol. Sci. 19:795–805.

Comstock, R. E., and R. H. Moll.
1963. Genotype-environment interactions, p. 164–196. *In* W. D. Hanson and H. F. Robinson [eds.], Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc., Washington, D.C.

Comstock, R. E., and H. F. Robinson.
1948. The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. Biometrics 4:254–266.

Conkle, M. T.
1970. Hybridization—application to pine populations, p. 131–133. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Conkle, M. T.
1971. Inheritance of alcohol dehydrogenase and leucine aminapeptidase isozymes in knobcone pine. For. Sci. 17:190–194.

Conkle, M. T.
1973. Growth data for 29 years from the California elevational transect study of ponderosa pine. For. Sci. 19:31–39.

Cress, C. E.
1966a. A comparison of recurrent selection systems. Genetics 54:1371–1379.

Cress, C. E.
1966b. Heterosis of the hybrid related to gene frequency differences between two populations. Genetics 53:269–274.

Cress, C. E.
1967. Reciprocal recurrent selection and modifications in simulated populations. Crop Sci. 7:561–567.

Crow, J. F.
1968. Some analyses of hidden variability in Drosophila populations, p. 71–86. *In* R. C. Lewontin [ed.], Population biology and evolution. Syracuse Univ. Press, New York.

Crow, J. F., and M. Kimura.
1970. An introduction to population genetics theory. 591 p. Harper & Row Publ., Inc., New York.

Cruden, D.
1949. The computation of inbreeding coefficients for closed populations. J. Hered. 40:248–251.

Crumpacker, D. W., and R. W. Allard.
1962. A diallel cross analysis of heading date in wheat. Hilgardia 32:275–318.

Curnow, R. N., and L. H. Baker.
1968. The effect of repeated cycles of selection and regeneration in populations of finite size. Genet. Res. 11:105–112.
Dawson, P. S.
1969. A conflict between Darwinian fitness and population fitness in Tribolium "competition" experiments. Genetics 62:413–419.
Dawson, P. S.
1972. Evolution in mixed populations of Tribolium. Evolution 26:357–365.
Demetrius, L.
1969. The sensitivity of population growth rate to perturbations in the life cycle components. Math. Biosci. 4:129–136.
DeWitt, C. T.
1960. On competition. Verslagen van landbouwkundige onder zoekingen 66:82.
Dickinson, A. G., and J. L. Jinks.
1956. A generalized analysis of diallel crosses. Genetics 41:65–78.
Dietrichson, J.
1961. Breeding for frost resistance. Silvae Genet. 10:172–179.
Eberhart, S. A., and W. A. Russell.
1966. Stability parameters for comparing varieties. Crop Sci. 6:36-40.
Ehrenberg, C. E.
1961. Increment and branch development of pine progenies. Skogen 48(1):6–8. For. Abstr. 22(3):366.
Ehrenberg, C. E.
1966. Parent-progeny relationship in Scots pine (*Pinus silvestris* L.). Results from three progeny tests with plus and minus tree progenies in Southern Sweden. Stud. Forestalia Suec. 40. 54 p.
Eisen, E. J.
1967. Mating designs for estimating direct and maternal genetic variances and direct-maternal genetic covariances. Can. J. Genet. Cytol. 9:13–22.
Eldridge, K. G.
1966. Genetic improvement of *Eucalyptus regnans* by selection of parent trees. Appita 19:133–138.
Empig, L. T., C. O. Gardner, and W. A. Compton.
1971. Theoretical gains for different population improvement procedures. Nebr. Agric. Exp. Stn. Agric. Misc. Publ. 26, 22 p. Lincoln.
Endler, J. A.
1973. Gene flow and population differentiation. Science 179:243–250.
Eriksson, G., I. Ekberg, and A. Jonsson.
1972. Meiotic and pollen investigations as a guide for localization of forest tree seed orchards in Sweden. Proc. Joint Symp. For. Tree Breed. Genet. Subj. Group, IUFRO, (Sect. 5), For. Trees, SABRAO. Part B-4(1), 28 p. Gov. For. Exp. Stn., Tokyo.
Ewens, W. J.
1963. Numerical results and diffusion approximations in a genetic process. Biometrika 50:241–249.
Falconer, D. S.
1955. Patterns of response in selection experiments with mice. Cold Spring Harbor Symp. Quant. Biol. 20:178–196.
Falconer, D. S.
1960. Introduction to quantitative genetics. 356 p. Ronald Press Co., New York.
Feller, W.
1951. Diffusion processes in genetics, p. 227–246. *In* J. Negman [ed.], Second Berkeley Symposium on Math. Stat. Probl. Univ. Calif. Press, Berkeley.
Feller, W.
1957. An introduction to probability theory and its applications. Vol. 1. 461 p. John Wiley & Sons, Inc., New York.

Finlay, K. W.
1963. Adaptation—its measurements and significance in barley breeding. First Int. Barley Genet. Symp. Proc., p. 351–359.

Finlay, K. W., and G. N. Wilkinson.
1963. The analysis of adaptation in a plant-breeding programme. Aust. J. Agric. Res. 14:742–754.

Finney, D. J.
1956. The consequences of selection for a variate subject to errors of measurements. Rev. Inst. Int. Stat. 24:1–10.

Fisher, R. A.
1950. Gene frequencies in a cline determined by selection and diffusion. Biometrics 6:353–361.

Fisher, R. A.
1958. Genetical theory of natural selection. 2d rev. ed. 291 p. Dover Publ., New York.

Fisher, R. A.
1965. The theory of inbreeding. 2d ed. 150 p. Acad. Press, New York.

Food and Agriculture Organization of the United Nations.
1970. Second world consultation on forest tree breeding. 2 vols. 1587 p. Food & Agric. Organ., U. N., Rome.

Fowler, D. P., and D. T. Lester.
1970. The genetics of red pine. USDA For. Serv. Res. Pap. WO–8, 13 p Washington, D.C.

Frankham, R., L. P. Jones, and J. S. F. Barker.
1968a. The effects of population size and selection intensity in selection for a quantitative character in Drosophila. I. Short-term response to selection. Genet. Res. 12:237–248.

Frankham, R., L. P. Jones, and J. S. F. Barker.
1968b. The effects of population size and selection intensity in selection for a quantitative character in Drosophila. II. Long-term response to selection. Genet. Res. 12:249–266.

Frankham, R., L. P. Jones, and J. S. F. Barker.
1968c. The effects of population size and selection intensity in selection for a quantitative character in Drosophila. III. Analyses of the lines. Genet. Res. 12:267–283.

Franklin, E. C.
1968. Artificial self-pollination and natural inbreeding in *Pinus taeda* L. Diss. Abstr. Sect. B, 29(4):1225.

Franklin, E. C.
1970a. Inbreeding depression effects and their influences on selection programs, p. 16. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Franklin, E. C.
1970b. Survey of mutant forms and inbreeding depression in species of the family Pinaceae. USDA For. Serv. Res. Pap. SE–61, 21 p. Southeast. For. Exp. Stn., Asheville, N.C.

Franklin, I., and Lewontin, R. C.
1970. Is the gene the unit of selection? Genetics 65:707–734.

Freeman, G. H., and J. M. Perkins.
1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. Heredity 27:15–23.

Frey, K. J., and T. Horner.
1957. Heritability in standard units. Agron. J. 49:59–62.

Fryer, J. H., and F. T. Ledig.
1972. Microevolution of the photosynthetic temperature optimum in relation to the elevational complex gradient. Can. J. Bot. 50:1231–1235.

Funk, D. T.
1970. Genetics of black walnut. U.S. For. Serv. Res. Pap. WO–10, 13 p. Washington, D.C.
Gantmacher, F. R.
1964. The theory of matrices. 276 p. Vol. 2. Chelsea Publ., New York.
Gardner, C. O.
1963. Estimates of genetic parameters in cross-fertilizing plants and their implications in plant breeding, p. 225–252. *In* W. D. Hanson and H. F. Robinson [eds.], Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc., Washington, D.C.
Gardner, C. O., and J. H. Lonnquist.
1959. Linkage and degree of dominance of genes controlling quantitative characters in maize. Agron. J. 51:524–528.
Gaylor, D. W., and R. L. Anderson.
1960. The construction and evaluation of some designs for the estimation of parameters in random models. N.C. State Univ. Inst. Stat. Mimeogr. Ser. 256, 88 p. Raleigh.
Giesel, J. T.
1971. The relations between population structure and rate of inbreeding. Evolution 25:491–496.
Gill, J. L.
1965a. Effects of finite size on selection advance in simulated genetic populations. Aust. J. Biol. Sci. 18:599–617.
Gill, J. L.
1965b. A Monte Carlo evaluation of predicted selection response. Aust. J. Biol. Sci. 18:999–1007.
Gill, J. L.
1965c. Selection and linkage in simulated geometric population. Aust. J. Biol. Sci. 18:1171–1187.
Gilpin, M. E.
1972. Enriched predator-prey systems: theoretical stability. Science 177: 902–904.
Gilpin, M. E., and K. E. Justice.
1972. Reinterpretation of the invalidation of the principle of competitive exclusion. Nature 236:273–274, 299–301.
Goddard, R. E., and C. L. Brown.
1961. An examination of seed orchard concepts. J. For. 59:252–256.
Goggans, J. F.
1961. The interplay of environment and heredity as factors controlling wood properties in conifers, with special emphasis on their effects on specific gravity. N.C. State Univ. Sch. For., Tree Improv. Program Tech. Rep. 11, 56 p. Raleigh.
Goldsmith, C. H., and D. W. Gaylor.
1970. Three stage nested designs for estimating variance components. Technometrics 12:487–498.
Graybill, F. A.
1961. An introduction to linear statistical models. 463 p. McGraw-Hill Book Co., New York.
Greenwood, J. J. D.
1969. Apostatic selection and population density. Heredity 24:157–161.
Griffing, B.
1956. Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9:463–493.
Griffing, B.
1960. Theoretical consequences of truncation selection based on the individual phenotype. Aust. J. Biol. Sci. 13:307–343.
Griffing, B.
1967. Selection in reference to biological groups. I. Individual and group selection applied to populations of unordered groups. Aust. J. Biol. Sci. 20:127–139.

Grisjuk, N. M.
1959. The inheritability of thorn formation in *Gleditsia triacanthos*. Bjulleten Moskovskogo Obscestva Ispytatelej Prirody. (Otd. Biol.) 64(2):117–122.

Hadley, G.
1964. Nonlinear and dynamic programming. 484 p. Addison-Wesley Publ. Co., Reading, Mass.

Haldane, J. B. S.
1948. The theory of a cline. J. Genet. 48:277–284.

Haldane, J. B. S., and S. D. Jayakar.
1963. Polymorphism due to selection of varying direction. J. Genet. 58:237–242.

Hamrick, J. L., and W. J. Libby.
1972. Variation and selection in western U.S. montane species. I. White fir. Silvae Genet. 21:29–35.

Hanover, J. W.
1962. Clonal variation in western white pine. I. Graftability. Intermt. For. & Range Exp. Stn. Res. Note 101, 4 p. Ogden, Utah.

Hanover, J. W.
1966. Gene control of monoterpene levels in *Pinus monticola* Dougl. Heredity 21:73–84.

Hanover, J. W., and B. V. Barnes.
1962. Heritability of height growth in year-old western white pine. For. Genet. Workshop Proc. 1962, South. For. Tree Improv. Comm. Sponsored Publ. 22:71–76. Macon, Ga.

Hanover, J. W., and B. V. Barnes.
1969. Heritability of height growth in western white pine seedlings. Silvae Genet. 18:80–82.

Hanson, W. D.
1963. Heritability, p. 125–140. *In* W. D. Hanson and H. F. Robinson [eds.], Statistical genetics and plant breeding. Natl. Acad. Sci., Res. Counc. Publ. 982, Washington, D.C.

Hanson, W. D.
1964. Genotype-environment interaction concepts for field experimentation. Biometrics 20:540–552.

Hanson, W. D.
1966. Effects of partial isolation (distance), migration, and different fitness requirements among environmental pockets upon steady state gene frequencies. Biometrics 22:453–468.

Hanson, W. D.
1970. Genotypic stability. Theor. Appl. Genet. 40:226–231.

Harlan, H. V., and M. L. Martini.
1938. The effect of natural selection in a mixture of barley varieties. J. Agric. Res. 57:189–199.

Harris, D. L.
1964. Genotypic covariances between inbred relatives. Genetics 50:1319–1348.

Hartley, H. O.
1964. Exact confidence regions for the parameters in non-linear regression laws. Biometrika 51:347–353.

Hartley, H. O.
1969. Some recent developments in nonlinear least squares estimation. Biometrics 25:793.

Hattemer, H. H.
1966. Die Eignung einiger blatt-und verzweigungsmerkmale fur die unterscheidung von schwarzpappel-hybridklonen. Der Zuchter 36: 317–327.

Hayman, B. I.
1958. The separation of epistatic from additive and dominance variation in generation means. Heredity 12:371–390.

Hayman, B. I.
1960a. Maximum likelihood estimation of genetic components of variation. Biometrics 16:369–381.
Hayman, B. I.
1960b. The theory and analysis of diallel crosses. Genetics 45:155–172.
Hayman, B. I., and K. Mather.
1953. The progress of inbreeding when homozygotes are at a disadvantage. Heredity 7:165–183.
Henderson, C. R.
1963. Selection index and expected genetic advance, p. 141–163. *In* W. D. Hanson and H. F. Robinson [eds.], Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc., Washington, D.C.
Hill, W. G.
1969. On the theory of artificial selection in finite populations. Gen. Res. 13:143–163.
Hill, W. G.
1971. Design and efficiency of selection experiments for estimating genetic parameters. Biometrics 27:293–311.
Hill, W. G., and A. Robertson.
1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.
Huhn, V. M.
1969. Untersuchungen zur Konkurrenz zwischen vershiedenen Genotypen in Pflanzenbestanden. I. Modifikation der Methode von Sakai zur Schatzung der genetischen-, Umwelt- und Konkurrenzvarianz einer Population. Silvae Genet. 18:186–192.
Huhn, V. M.
1970a. The competitive environment and its genetic reaction variations, p. 62–86. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South For. Exp. Stn., New Orleans, La.
Huhn, V. M.
1970b. Unterschungen zur Konkurrenz zwischen verschiedenen Genotypen in Pflanzenbestanden. III. Das Korrelationsmuster eines Bestandes. Silvae Genet. 19:77–89.
Huhn, V. M.
1970c. Untersuchungen zur Konkurrenz zwischen verschiedenen Genotypen in Pflanzenbestanden. IV. Probleme der optimalen Parzellengrossein felduersuchen. Silvae Genet. 19:151–163.
Hyun, S. K.
1971. Developing advanced generation breeding population for a hybrid breeding program. Paper presented to the Working Group on Quantitative Genetics, Section 22, IUFRO, 1971. Gainesville, Fla. 18 p.
Karlin, S., and M. W. Feldman.
1969. Linkage and selection: new equilibrium properties of the two-locus symmetric viability model. Natl. Acad. Sci. Proc. 62:70–74.
Kearsey, M. J.
1965. Biometrical analysis of a random mating population: a comparison of five experimental designs. Heredity 20:205–235.
Kellison, R. C.
1970. Phenotypic and genotypic variation of yellow-poplar (*Liriodendron tulipifera* L.). Ph.D. Thesis, N.C. State Univ., Raleigh, 112 p.
Kempthorne, O.
1957. An introduction to genetic statistics. 545 p. John Wiley & Sons, Inc., New York.
Kempthorne, O., and A. W. Nordskog.
1959. Restricted selection indices. Biometrics 15:10–19.
Kendall, M. G.
1961. A course in multivariate analysis. 185 p. Hafner Publ. Co., New York.

Kendall, M. G., and A. Stuart.
1963. The advanced theory of statistics. Vol. 1. 433 p. Hafner Publ. Co., New York.
Kendall, M. G., and A. Stuart.
1966. The advanced theory of statistics. Vol. III. 552 p. Hafner Publ. Co., New York.
Keyfitz, N.
1968. Introduction to the mathematics of population. 450 p. Addison-Wesley Publ. Co., Redding, Mass.
Kimura, M.
1957. Some problems of stochastic processes in genetics. Ann. Math. Stat. 28:882–901.
Kimura, M.
1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.
Kimura, M.
1964. Diffusion models in population genetics. J. Appl. Probab. 1:177–232.
Kimura, M.
1969. The rate of molecular evolution considered from the standpoint of population genetics. Natl. Acad. Sci. Proc. 63:1181–1188.
Kimura, M., and J. F. Crow.
1963. On the maximum avoidance of inbreeding. Genet. Res. 4:399–415.
Kimura, M., and T. Maruyama.
1971. Pattern of neutral polymorphism in a geographically structured population. Gen. Res. 18:125–131.
Kimura, M., and T. Ohta.
1971. Theoretical aspects of population genetics. Monogr. Popul. Biol. 4, 219 p. Princeton Univ. Press, Princeton, N.J.
Kimura, M., and G. H. Weiss.
1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.
King, J. P.
1965. Seed source x environment interactions in Scotch pine. I. Height growth. Silvae Genet. 14:105–115.
King, J. P., and H. Nienstaedt
1965. Variation in needle cast susceptibility among 29 jack pine seed sources. Silvae Genet. 14:194–198.
Kinloch, B. B., Jr., G. K. Parks, and C. W. Fowler.
1970. White pine blister rust: simply inherited resistance in sugar pine. Science 167:193–195.
Kojima, K.
1959a. Role of epistasis and overdominance in stability of equilibria with selection. Natl. Acad. Sci. Proc. 45:984–989.
Kojima, K.
1959b. Stable equilibria for the optimum model. Natl. Acad. Sci. Proc. 45:989–993.
Kojima, K.
1961. Effects of dominance and size of population on response to mass selection. Genet. Res. 2:177–188.
Kojima, K.
1971. Is there a constant fitness value for a given genotype? No! Evolution 25:281–285.
Kojima, K., and S. L. Huang.
1972. Effects of population density on the frequency-dependent selection in the esterase-6 locus of *Drosophila melanogaster*. Evolution 26: 313–321.
Kojima, K., and T. M. Kelleher.
1963a. A comparison of purebred and crossbred selection schemes with two populations of *Drosophila pseudoobscura*. Genetics 48:57–72.

Kojima, K., and T. M. Kelleher.
1963b. Selection studies of quantitative traits with laboratory animals. p. 395–422. *In* Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc. Publ. 982. Washington, D.C.

Kojima, K., and Y. N. Tobari.
1969. Selective modes associated with karyotypes in *Drosophila ananassae*. II. Heterosis and frequency-dependent selection. Genetics 63: 639–651.

Kraus, J. F.
1967. Heritability of some seed characteristics in slash pine (*Pinus elliottii* Engelm.). Diss. Abstr. 27:4199–B.

Kriebel, H. B.
1965. Parental and provenance effects on growth of red oak seedlings. Fourth Cent. States For. Tree Improv. Conf. Proc. 1964:19–25.

Kriebel, H. B., and W. J. Gabriel.
1969. Genetics of sugar maple. USDA For. Serv. Res. Pap. WO–7, 17 p. Washington, D.C.

Kriebel, H. B., G. Namkoong, and R. A. Usanis.
1972. Analysis of genetic variation in 1-, 2-, and 3-year old eastern white pine in incomplete diallel cross experiments. Silvae Genet. 21:44–48.

Lacaze, J. F., and M. Arbez.
1971. Variabilité infraspecifique de l'épicéa. Héritabilité et corrélations génétiques de quelque caractéres au stade juvénile. Ann. Sci. For. 28:141–183.

Langlet, O.
1936. Studier över tallens fysiologiska variábilitet och dess samband med klimatet. Medd. Stat. Skogsforskningsinst. 29:219–470.

Langlet, O.
1963. Patterns and terms of intra-specific ecological variability. Nature 200:347–348.

Langner, W.
1953. Eine mendelspaltung bei Aurea-Formen von *Picea abies* (L.) Karst. als Mittel zur Klärung der Befruchtungsverhältnisse im Walde. Z. f. Forstgenet. u. Forstpflanzenzücht. 2:49–51.

Lanner, R. M.
1966. Needed: a new approach to the study of pollen dispersion. Silvae Genet. 15:50–52.

Latter, B. D. H.
1965. The response to artificial selection due to autosomal genes of large effect. I. Changes in gene frequency at an additive locus. Aust. J. Biol. Sci. 18:585–598.

Latter, B. D. H.
1966. The interaction between effective population size and linkage intensity under artificial selection. Genet. Res. 7:313–323.

Latter, B. D. H., and A. Robertson.
1962. The effects of inbreeding and artificial selection on reproductive fitness. Genet. Res. 3:110–138.

Laude, H. H., and A. F. Swanson.
1942. Natural selection in varietal mixtures of winter wheat. J. Am. Soc. Agron. 34:270–274.

Ledig, F. T.
1970. Genotype x environment interaction in controlled environments: the physiological basis for differential response, p. 90–99. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics., Section 22, IUFRO 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Ledig, F. T., and T. O. Perry.
1967. Variation in photosynthesis and respiration among loblolly pine progenies. Ninth South. Conf. For. Tree Improv. Proc. 1967:120–128.

Lerner, I. M.
1954. Genetic homeostasis. 134 p. Oliver & Boyd, London.
Levene, H.
1953. Genetic equilibrium when more than one ecological niche is available. Am. Nat. 87:331–333.
Levin, B. R.
1971. The operation of selection in situations of interspecific competition. Evolution 25:249–264.
Levin, D. A., and H. W. Kerster.
1967. Natural selection for reproductive isolation in phlox. Evolution 21:679–687.
Levin, D. A., and H. Kerster.
1969. Density-dependent gene dispersal in Liatris. Am. Nat. 103:61–74.
Levin, D. A., and H. W. Kerster.
1971. Neighborhood structure in plants under diverse reproductive methods. Am. Nat. 105:345–354.
Levins, R.
1968. Evolution in changing environments; some theoretical explorations. Monographs in population biology. 120 p. Princeton Univ. Press, Princeton, N.J.
Lewontin, R. C.
1964. The interaction of selection and linkage. II. Optimum models. Genetics 50:757–782.
Lewontin, R. C.
1965. Selection for colonizing ability, p. 77–94. *In* H. G. Baker and G. L. Stebbins [eds.], Genetics of colonizing species. Acad. Press, New York.
Lewontin, R. C., and K. Kojima.
1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472.
Li, C. C.
1955. Population genetics. 366 p. Univ. Chicago Press, Chicago.
Li, C. C.
1967. Genetic equilibrium under selection. Biometrics 23:397–484.
Libby, W. J.
1964. Clonal selection, and an alternative seed orchard scheme. Silvae Genet. 13:32–40.
Libby, W. J.
1969. Seedling vs. vegetative orchards, p. 306–316. *In* FAO-N.C. State For. Tree Improv. Cent. Lecture Notes, N.C. State Univ. 1969, 334 p.
MacArthur, R. H.
1962. Some generalized theories on natural selection. Natl. Acad. Sci. Proc. 48:1893–1897. Washington, D.C.
Madalena, F. E., and W. G. Hill.
1972. Population structure in artificial selection programmes: simulation studies. Genet. Res. 20:75–99.
Malécot, Gustave.
1969. The mathematics of heredity. Revised, edited, and translated by D. M. Yermanos. 88 p. W. H. Freeman, San Francisco.
Marcuse, S.
1949. Optimum allocation and variance components in nested sampling with an application to chemical analysis. Biometrics 5:189–206.
Marquardt, D. W.
1963. Least squares estimation of non-linear parameters. 30 p. IBM SHARE SDA 2094.01 (NLIN).
Mather, K.
1969. Selection through competition. Heredity 24:529–540.
Mather, K., and B. I. Hayman.
1952. The progress of inbreeding where heterozygotes are at an advantage. (Abstr. 183) Biometrics 8:176.

Matsuura, T., and K. Sakai.
1972. Geographical variation on an isozyme level in Abies sacalinensis. Proc. Joint IUFRO-SARAO Symp., Tokyo, A(9), 12 p.
Mergen, F.
1963. Evaluation of spontaneous, chemical, and radiation-induced mutations in Pinaceae. World Consult. For. Genet. & Tree Improv. Proc. 23–30 August 1963, Stockh. FAO/Forgen–63, Vol. I, 1/:1–15.
Misra, R. K.
1966. Vectorial analysis for genetic clines in body dimensions in populations of *Drosophila subobscura* Coll. and a comparison with those of *D. robusta* Sturt. Biometrics 22:469–487.
Moffett, A. A., and K. M. Nixon.
1963. One parent progeny testing with black wattle (*Aracia mearnsii* de wild). World Consult. For. Genet. & Tree Improv., Stockh. FAO/Forgen–63, Vol. I, 2a/5:1–12.
Moll, R. H., J. H. Lonnquist, J. V. Fortuno, and E. C. Johnson.
1965. The relationship of heterosis and genetic divergence in maize. Genetics 52:139–144.
Moll, R. H. and H. F. Robinson.
1967. Quantitative genetic investigations of yield of maize. Der Zuchter 37:192–199.
Moll, R. H., W. S. Salhuana, and H. F. Robinson.
1962. Heterosis and genetic diversity in variety crosses of maize. Crop Sci. 2(3):197–198.
Moll, R. H., and C. W. Stuber.
1971. Comparisons of response to alternative selection procedures initiated with two populations of maize (*Zea mays* L.). Crop. Sci. 11: 706–711.
Moran, P. A. P.
1964. On the nonexistence of adaptive topographies. Ann. Hum. Genet., 27(Pt. 2):383–393. London.
Morgenstern, E. K., and A. H. Teich.
1969. Phenotypic stability of height growth of jack pine provenances. Can. J. Genet. Cytol. 11:110–117.
Mostafa, M. G.
1967. Designs for the simultaneous estimation of functions of variance components from two-way crossed classification. Biometrika 54: 127–131.
Mukai, T.
1969. Maintenance of polygenic and isoallelic variation in populations. Proc. XII Int. Congr. Genet. 3:293–308.
Muzik, T. J., and Cruzado, H. J.
1958. Transmission of juvenile rooting ability from seedlings to adults of *Hevea brasilienseis*. Nature 181:1288.
Namkoong, G.
1965. Inbreeding effects on estimation of additive genetic variance. For. Sci. 12:8–13.
Namkoong, G.
1966a. Application of Nelder's designs in tree improvement research. Eighth South. Conf. For. Tree Improv. Proc. 1965:24–37.
Namkoong, G.
1966b. Family indices for seed-orchard selection, p. 7–12. *In* Joint proceedings second genetics workshop of the Society of American Foresters and the seventh Lake States forest tree improvement conference 1965. U.S. For. Serv. Res. Pap. NC–6. North Cent. For. Exp. Stn., St. Paul, Minn.
Namkoong, G.
1966c. Statistical analysis of introgression. Biometrics 22:488–502.
Namkoong, G.
1967. Multivariate analysis of multiple regression in provenances analysis. Proc. 14th IUFRO Congr. Sect. 22, p. 308–318.

Namkoong, G.
1969. Nonoptimality of local races. Tenth South. Conf. For. Tree Improv. Proc. 1969:149–153.

Namkoong, G.
1970a. Optimum allocation of selection intensity in two stages of truncation selection. Biometrics 26:465–476.

Namkoong, G.
1970b. Problems of multiple-trait breeding. FAO/IUFRO. Second World Consult. For. Tree Bred. Vol. 1 (1969):775-781.

Namkoong, G.
1971. Criteria for choosing mating designs for tree breeding. Paper presented to the Working Group on Quantitative Genetics Section 22, IUFRO. Gainesville, Fla., 10 p.

Namkoong, G.
1972. Persistance of variances for stochastic, discrete-time, population growth models. Theor. Popul. Biol. 3:507–518.

Namkoong, G., A. C. Barefoot, and R. G. Hitchings.
1969. Evaluating control of wood quality through breeding. Tappi 52:1935–1938.

Namkoong, G., R. C. Biesterfeldt, and J. C. Barber.
1971. Tree breeding and management decisions. J. For. 69:138–142.

Namkoong, G., and D. L. Miller.
1968. Estimation of non-linear parameters for a non-asymptotic function. Biometrics 24:439–440.

Namkoong, G., and J. H. Roberds.
1974. Extinction probabilities and the changing age structure of redwood forests. Am. Nat. 108:355–368.

Namkoong, G., E. B. Snyder, and R. W. Stonecypher.
1966. Heritability and gain concepts for evaluating breeding systems such as seedling seed orchards. Silvae Genet. 15:76–84.

Namkoong, G., and E. B. Snyder.
1969. Accurate values for selection intensities. Silvae Genet. 18:172–173.

Namkoong, G., and A. E. Squillace.
1970. Problems in estimating genetic variance by Shrikhande's method. Silvae Genet. 19:74–77.

Namkoong, G., R. A. Usanis, and R. R. Silen.
1972. Age related variation in genetic control of height growth in Douglas-fir. Theor. Appl. Genet. 42:151–159.

Nanson, A.
1967. Modèle théorique pour l'étude des tests précoces. Biometrie-Praximetrie 8:84-107.

Nanson, A.
1970. Juvenile and correlated trait selection and its effect on selection programs. p. 17–26. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Nasoetion, A. H.
1965. An evaluation of two procedures to estimate genetic and environmental parameters in a simultaneous selfing and partial diallel test crossing design. N.C. State Univ. Inst. Stat. Mimeogr. Ser. 425, 69 p. Raleigh.

Nei, M.
1963. Effect of selection on the components of genetic variances, p. 501–515. *In* Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc. Publ. 982, Washington, D.C.

Nicholls, J. W. P., H. E. Dadswell, and J. M. Fielding.
1964. The heritability of wood characteristics of *Pinus radiata*. Silvae Genet. 13:68–71.

Nilsson, B., and E. Andersson.
1970. Spruce and pine racial hybrid variations in northern Europe, p. 118–128. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Oakes, M. W.
1967. The analysis of a diallel cross of heterozygous or multiple allelic lines. Heredity 28:83–95.

Orr-Ewing, A. L.
1965. Inbreeding and single crossing in Douglas-fir. For. Sci. 11:279–290.

Osborne, R., and W. S. B. Paterson.
1952. On the sampling variance of heritability estimates derived from variance analyses. R. Soc. Edinb. B. Proc. 64:456–461.

Owen, G.
1968. Game theory. 228 p. W. B. Saunders Co., Philadelphia, Pa.

Patel, R. M., C. C. Cockerham, and J. O. Rawlings.
1962. Selection among factorially classified variables. N. C. State Univ. Inst. Stat. Mimeogr. Ser. 317, 43 p. Raleigh.

Patel, R. M., C. C. Cockerham, and J. O. Rawlings.
1969. Selection among diallel classified variables. Biometrics 25:49–61.

Penny, L. H., W. A. Russell, and G. F. Sprague.
1962. Types of gene action in yield heterosis in maize. Crop Sci. 2:341–344.

Penny, L. H., W. A. Russell, G. F. Sprague, and A. R. Hallauer.
1966. Recurrent selection, p. 352–367. *In* Statistical genetics and plant breeding. Natl. Acad. Sci., Natl. Res. Counc. Publ. 982, Washington, D.C.

Perkins, J. M., and J. L. Jinks.
1968. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. Heredity 23:339–356.

Perry, T. O., and Chi-Wu Wang.
1958. The value of genetically superior seed. J. For. 56:843–845.

Pianka, E. R.
1972. r and K selection or b and d selection? Am. Nat. 106:581–588.

Pielou, E. C.
1969. An introduction to mathematical ecology. 286 p. Wiley-Interscience, New York.

Pike, D. J.
1969. A comparison of two methods for predicting changes in the distribution of gene frequency when selection is applied repeatedly to a finite population. Genet. Res. 13:117–126.

Pimentel, D., and A. B. Soans.
1970. Animal populations regulated to carrying capacity of plant host by genetic feedback. Adv. Study Inst. Dyn. Numbers in Popul. Proc.: 313–326.

Pollard, J. H.
1966. On the use of the direct matrix product in analysing certain stochastic population models. Biometrika 53:397–415.

Pomeroy, K. B., and C. F. Korstian.
1949. Further results on loblolly pine seed production and dispersal. J. For. 47:968–970.

Prout, T.
1968. Sufficient conditions for multiple niche polymorphism. Am. Nat. 102:493–496.

Rao, C. R.
1971. Minimum variance quadratic unbiased estimation of variance components. J. Multivariate Anal. 1:445–456.

Rawlings, J. O.
1970. Present status of research on long and short-term recurrent selection in finite populations—choice of population size, p. 1–15. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Righter, F. I.
1960. Forest tree improvement through inbreeding and intraspecific and interspecific hybridization. Fifth World For. Congr. Vol. 2:783–787.

Robertson, A.
1960. A theory of limits in artificial selection. London R. Soc. B. Proc. 153:234–249.

Robertson, A.
1970. The reduction in fitness from genetic drift at heterotic loci in small populations. Genet. Res. 15:257–259.

Robertson, F. W., and E. C. R. Reeve.
1952. Studies in quantitative inheritance. I. The effects of selection of wing and thorax length in *Drosophila melanogaster*. J. Genet. 50:414–448.

Robinson, H. F., and R. H. Moll.
1959. Implications of environmental effects on genotypes in relation to breeding. Am. Seed Trade Assoc. Hybrid Corn Div. Rep. Hybrid Corn Ind.-Res. Conf. 14:24–31.

Rojas, B. A., and G. F. Sprague.
1952. A comparison of variance components in corn yield trials: III. General and specific combining ability and their interaction with locations and years. Agron. J. 44:462–466.

Rouvier, R.
1966. L'analyse en composantes principales: Son utilisation en genetique et ses rapports avec l'analyse discriminatoire. Biometrics 22:343–357.

Roy, S. N.
1957. Some aspects of multivariate analysis. 214 p. John Wiley & Sons, Inc., New York.

Rudolph, T. D., and H. Nienstaedt.
1962. Polygenic inheritance of resistance to winter injury in jack pine-lodgepole pine hybrids. J. For. 60:138–139.

Russell, W. A., G. F. Sprague, and L. H. Penny.
1963. Mutations affecting quantitative characters in long-time inbred lines of maize. Crop Sci. 3:175–178.

Saeterstal, L. S.
1963. The rate of drying of young excised plants of various provenances of Norway Spruce and Douglas fir. Meddeleser Vestlanders Forstlige Forsksstasjon 12(1) No. 38, 88p.

Sakai, K.
1955. Competition in plants and its relation to selection. Cold Spring Harbor Symp. Quant. Biol. 20:137–157.

Sakai, K.
1961. Competitive ability in plants: its inheritance and some related problems. Symp. Soc. Exp. Biol. No. 15, p. 245–263.

Sakai, K.
1965. Contributions to the problem of species colonization from the viewpoint of competition and migration, p. 215–241. *In* H. G. Baker and G. L. Stebbins [eds.], Genetics of colonizing species. Academic Press, New York.

Sakai, K.
1971. Genetic differentiation in a natural population of forest tree species. Sabrao Newsl. 3:71–72.

Sakai, K., and S. Hatakeyama.
1963. Estimation of genetic parameters in forest trees without raising progeny. Silvae Genet. 12:152-157.

Sakai, K., Y. Miyazaki, and T. Matsuura.
1971. Genetical studies in natural populations of forest trees. Silvae Genet. 20:168–173.

Sakai, K., and H. Mukaide.
1967. Estimation of genetic, environmental, and competitional variances in standing forests. Silvae Genet. 16:149–152.

Sakai, K., H. Mukaide, and K. Tomita.
1968. Intraspecific competition in forest trees. Silvae Genet. 17:1–5.

Sarvas, R.
1963. Pollen dispersal and its significance in silviculture and genetics. Finland For. Res. Inst. Annu. Rep. 2, 2 p. (Summary).

Sarvas, R.
1967. Pollen dispersal within and between subpopulations; role of isolation and migration in microevolution of forest tree species. XIV. IUFRO-Congr. Proc. Vol. III Munich 1967:332–345.

Sarvas, R.
1970. Genetical adaptation of forest trees to the heat factor of the climate. FAO/IUFRO Second World Consult. For. Tree Breed. Vol. 1(1969):187–202.

Schreiner, E. J.
1966. Maximum genetic improvement of forest trees through synthetic multiclonal hybrid varieties. Northeast. For. Tree Improv. Conf. Proc. 13:7–13.

Schreiner, E. J.
1968. Forest tree breeding. Unasylva 22(3):3–9. No. 90.

Schutz, W. M., and C. A. Brim.
1971. Inter-genotypic competition in soybeans. III. An evaluation of stability in multiline mixtures. Crop Sci. 11:684–689.

Schutz, W. M., and C. C. Cockerham.
1962. The effect of field blocking on gain from selection. N. C. State Univ. Inst. Stat. Mimeogr. Ser. 328, 86 p. Raleigh.

Schutz, W. M., and S. A. Usanis.
1969. Inter-genotypic competition in plant populations. II. Maintenance of allelic polymorphisms with frequency-dependent selection and mixed selfing and random mating. Genetics 61:875–891.

Searle, S. R.
1971. Linear models. 532 p. John Wiley & Sons, Inc., New York.

Shelbourne, C. J. A.
1973. Planning breeding programs for tropical conifers grown as exotics, p. 155–179. *In* J. Burley and D. G. Nikles [eds.], Selection and breeding to improve some tropical conifers. Commonw. For. Inst., Oxford.

Shrikhande, V. J.
1957. Some considerations in designing experiments on coconut trees. J. Indian Soc. Agric. Stat. 9:82–99.

Singh, M., and R. C. Lewontin.
1966. Stable equilibria under optimizing selection. Natl. Acad. Sci. Proc. 56:1345–1348.

Slobodkin, L. B.
1961. Growth and regulation of animal populations. 184 p. Holt, Rinehart and Winston, New York.

Sluder, E. R.
1970. Gene flow patterns in forest tree species and implications for tree breeding. FAO-IUFRO Second World Consult. For. Tree Breed. Vol. 2(1969):1139–1150.

Smith, D. C.
1966. Plant breeding—development and success, p. 3–54. *In* Kenneth J. Frey [ed.], Plant breeding: a symposium held at Iowa State University. Iowa State Univ. Press, Ames, Iowa.

Smith, W. J.
1967. The heritability of fibre characteristics and its application to wood quality improvement in forest trees. Silvae Genet. 16:41–50.

Snyder, E. B.
1961. Measuring branch characters of longleaf pines. USDA For. Serv. Occas. Pap. 184, 4 p. South. For. Exp. Stn., New Orleans, La.

Snyder, E. B.
1966. Lattice and compact family block designs in forest genetics, p. 12–17. *In* Joint proceedings second genetics workshop of the Society of American Foresters and the seventh improvement conference 1965. U.S. For. Serv. Res. Pap. NC–6. North Central For. Exp. Stn., St. Paul, Minn.

Snyder, E. B.
1969. Parental selection versus half-sib family selection of longleaf pine. Tenth South. Conf. For. Tree Improv. Proc. 1969:84–88.

Synder, E. B., and R. M. Allen.
1971. Competitive ability of slash pine analyzed by genotype x environment stability method. Eleventh Conf. South. For. Tree Improv. Proc. 1971:142–147.

Sorensen, Frank.
1969. Embryonic genetic load in coastal Douglas-fir, *Pseudotsuga menziesii* var. *menziesii*. Am. Nat. 103:389–398.

Sprague, G. F.
1966. Quantitative genetics in plant improvement, p. 315–354. *In* Kenneth J. Frey [ed.], Plant breeding: a symposium held at Iowa State University. Iowa State Univ. Press, Ames, Iowa.

Sprague, G. F.
1967. Plant breeding. Annu. Rev. Genet. 1:269–294.

Squillace, A. E.
1966a. Combining superior growth and timber quality with high gum yield in slash pine. Eighth South. Conf. For. Tree Improv. Proc. 1965:73–76.

Squillace, A. E.
1966b. Geographic variation in slash pine. Soc. Ann. For., For. Sci. Monogr. 10, 56 p.

Squillace, A. E.
1966c. Racial variation in slash pine as affected by climatic factors. U.S. For. Serv. Res. Pap. SE–21, 10 p. Southeast. For. Exp. Stn., Asheville, N.C.

Squillace, A. E.
1970. Genotype-environment interactions in forest trees, p. 49–61. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

Squillace, A. E.
1973. Comparison of some alternative second-generation breeding plans for slash pine. Twelfth South. For. Tree Improv. Conf. Proc. 1973:2–13.

Squillace, A. E., R. T. Bingham, G. Namkoong, and H. F. Robinson.
1967. Heritability of juvenile growth rate and expected gain from selection in western white pine. Silvae Genet. 16:1–6.

Squillace, A. E., and G. S. Fisher.
1966. Evidences of the inheritance of turpentine composition in slash pine, p. 53–60. *In* Joint proceedings second genetics workshop of the Society of American Foresters and the seventh Lake States forest tree improvement conference 1965. U.S. For. Serv. Res. Pap. NC–6. North Cent. For. Exp. Stn., St. Paul, Minn.

Squillace, A. E., and J. F. Kraus.
1959. Early results of a seed source study of slash pine in Georgia and Florida. Fifth South. Conf. For. Tree Improv. Proc. 1959:21–34.

Squillace, A. E., and J. F. Kraus.
1963. The degree of natural selfing in slash pine as estimated from albino frequencies. Silvae Genet. 12:46–50.
Steele, R. G. D., and J. H. Torrie.
1960. Principles and procedures of statistics. 481 p. McGraw-Hill Book Co., New York.
Stephens, S. G.
1961. Species differentiation in relation to crop improvement. Crop Sci. 1:1–5.
Stern, K.
1962. Über die relative Bedeutung von Erbgut und umwelt für die variation einiger Merkmale innerhalb von Waldbaumpopulationen. Forstliche Mitteilungen 15:131–134.
Stern, K.
1963. Über die abhangigkeit des blühens der Sandbirke von Erbgut und umwelt. [The dependence of the flowering of *Betula verrucosa* on inheritance and environment.] Silvae Genet. 12:26–31. [In German, English Summary]
Stern, K.
1964 [The intensity of natural selection along an altitudinal cline], p. 139–146. *In* H. Schmidt-Vogt [ed.], Forstsamengewinnung und Pflanzenanzucht für das Hochgebirge. BLV Verslagsgesellschaft. Munich.
Stern, K.
1972. Breeding population and productive population in forest tree breeding. *In* Proceedings joint symposium for forest tree breeding of IUFRO and SABRAO. Gov. For. Exp. Stn. Symp. A. 9 p. Tokyo, Japan.
Stern, K., and H. H. Hattemer.
1964. Problems involved in some models of selection in forest tree breeding. Silvae Genet. 13:27–32.
Stettler, R. F., K. S. Bawa, and G. K. Livingston.
1970. Haploidy: An approach to the development of high-yielding varieties. FAO/IUFRO Second World Consult. For. Tree Breed. Vol. 2(1969):1031–1039.
Stonecypher, R. W.
1966. The loblolly pine heritability study. Int. Pap. Co., Southlands Exp. For. Tech. Bull. 5, 128 p. Bainbridge, Ga.
Strickland, R. K., and R. E. Goddard.
1966. Inheritance of branching and crown characteristics in slash pine. Eighth South. Conf. For. Tree Improv. Proc. 1965:57–63.
Stuber, C. W.
1970. Theory and use of hybrid population statistics, p. 100–112. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.
Stuber, C. W., and C. C. Cockerham.
1966. Gene effects and variances in hybrid populations. Genetics 54:1279–1286.
Sykes, Z. M.
1969. Some stochastic versions of the matrix model for population dynamics. J. Am. Stat. Assoc. 64:111–130.
Tai, G. C. C.
1971. Genotypic stability analysis and its application to potato regional trials. Crop Sci. 11:184–190.
Tallis, G. M.
1962. A selection index for optimum genotype. Biometrics 18:120–122.
Texas Forest Service.
1957. Forest tree improvement program of the Texas Forest Service. Fifth Prog. Rep. For. Serv. Circ. 58, 14 p.

Tigerstedt, Von P. M. A.
1966. [Development of the genetic variances of growth in height in a field experiment with *Betula verrucosa*.] Silvae Genet. 15:135–137. [In German, Title in German, English Summary].

Toda, R.
1956. On the crown slenderness in clones and seedlings. Z. fur Forstgenet. und Forstpflanzenzucht. 5:1–5.

Toda, R.
1961. Studies on the genetic variance in Cryptomeria. Japan: For. Exp. Stn. Bull. 132:1–46.

van Buijtenen, J. P.
1970. Applications of interspecific hybridization in forest tree breeding, p. 113–117. *In* Papers presented at the second meeting of the Working Group on Quantitative Genetics, Section 22, IUFRO, 1969. USDA For. Serv. South. For. Exp. Stn., New Orleans, La.

van Buijtenen, J. P., and W. W. Saitta.
1972. Linear programming applied to the economic analysis of forest tree improvement. J. For. 70:164–167.

Vandermeer, J. H.
1972. On the covariance of the community matrix. Ecology 53:187–189.

Varnell, Ray J., A. E. Squillace, and G. W. Bengtson.
1967. Variation and heritability of fruitfulness in slash pines. Silvae Genet. 16:125–128.

Wakeley, P. C.
1954. Planting the southern pines. USDA For. Serv. Agric. Monogr. 18, 233 p. Washington, D.C.

Walker, L. C., and R. D. Hatcher.
1965. Variation in the ability of slash pine progeny groups to absorb nutrients. Soil Sci. Soc. Am. Proc. 29:616–621.

Wallace, B.
1968. Polymorphism, population size, and genetic load, p. 87–108. *In* R. C. Lewontin [ed.], Population biology and evolution. Syracuse Univ. Press, New York.

Wang, C. W., T. O. Perry, and A. G. Johnson.
1960. Pollen dispersion of slash pine (*Pinus elliottii* Engelm.) with special references to seed orchard management. Silvae Genet. 9:78–86.

Webb, C. D.
1970. Early results of performance trials of American sycamore. (Abstr.) First N. Am. For. Biol. Workshop Aug. 5–7, 1970. Mich. State Univ., East Lansing.

Weiss, G. H., and M. Kimura.
1965. A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probab. 2:129–149.

Weissner, E. W.
1971. Multitype branching processes in random environments. J. Appl. Probab. 8:17–31.

Wells, O. O., and P. C. Wakeley.
1966. Geographic variation in survival, growth, and fusiform-rust infection of planted loblolly pine. Soc. Am. For., For. Sci. Monogr. 11, 40 p.

Wilcox, J. R.
1970. Inherent variation in south Mississippi sweetgum. Silvae Genet. 19:91–94.

Wilcox, J. R., and R. E. Farmer, Jr.
1967. Variation and inheritance of juvenile characters of eastern cottonwood. Silvae Genet. 16:162–165.

Wilcox, J. R., and R. E. Farmer, Jr.
1968. Heritability and *C* effects in early root growth of eastern cottonwood cuttings. Heredity 23:239–245.

Williams, J. S.
1962. The evaluation of a selection index. Biometrics 18:375–393.
Wright, J. W.
1962. Genetics of forest tree improvement. FAO For. & For. Prod. Stud. 16, 399 p., Rome.
Wright, J. W.
1963. Genetic variation among 140 half-sib Scotch pine families derived from 9 stands. Silvae Genet. 12:83–89.
Wright, J. W.
1964a. Flowering age of clonal and seedling trees as a factor in choice of breeding system. Silvae Genet. 13:21–27.
Wright, J. W.
1964b. Hybridization between species and races. Unasylva 18:30–39.
Wright, J. W.
1970. Genetics of eastern white pine. U.S. For. Serv. Pap. WO–9, 16 p. Washington, D.C.
Wright, J. W.
1971. Second generation seed orchards of white and red pines in Michigan. Paper presented at the Working Group on Quantitative Genetics, Section 22, IUFRO, 1971. Gainesville, Fla.
Wright, J. W., L. F. Wilson, and W. K. Randall.
1967. Differences among Scotch pine varieties in susceptibility to European pine sawfly. For. Sci. 13:175–181.
Wright, S.
1922. Coefficients of inbreeding and relationship. Am. Nat. 56:330–338.
Wright, S.
1935a. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. J. Genet. 30:243–256.
Wright, S.
1935b. Evolution in populations in approximate equilibrium. J. Genet. 30:257–266.
Wright, S.
1940. Breeding structure of populations in relation to speciation. Am. Nat. 74:232–248.
Wright, S.
1943. Isolation by distance. Genetics 28:114–138.
Wright, S.
1949. Population structure in evolution. Am. Phil. Soc. Proc. 93:471–477.
Wright, S.
1951. The genetical structure of populations. Ann. Eugen. 15:323–354.
Wright, S.
1967. "Surfaces" of selective value. Natl. Acad. Sci. Proc. 58:165–172.
Wright, S.
1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. 511 p. Univ. Chicago Press, Chicago.
Wright, S.
1970. Random drift and the shifting balance theory of evolution, p. 1–31. *In* K. Kojima [ed.], Mathematical topics in population genetics. Springer-Verlag, New York.
Young, S. S. Y.
1961. A further examination of the relative efficiency of three methods of selection for genetic gains under less-restricted conditions. Genet. Res. 2:106–121.
Young, S. S. Y.
1964. Multi-stage selection for genetic gain. Heredity 19:130–145.
Zobel, B.
1961. Inheritance of wood properties in conifers. Silvae Genet. 10:65–70.
Zobel, B., D. Cole, and R. Stonecypher.
1962. Wood properties of clones of slash pine. For. Genet. Workshop Proc. 1962, South. For. Tree Improv. Comm. Sponsored Publ. 22: 32–39. Macon, Ga.

Zobel, B., and R. L. McElwee.
1964. Seed orchards for the production of genetically improved seed. Silvae Genet. 13:4–11.
Zobel, B., R. L. McElwee, and C. Browne.
1961. Interrelationship of wood properties of loblolly pine. Sixth South. Conf. For. Tree Improv. Proc. 1961:142–163.

☆ U.S. GOVERNMENT PRINTING OFFICE 1979 O—262-855

END