EFFECT OF CORRELATED GENE DISTRIBUTION DUE TO SAMPLING ON THE DIALLEL ANALYSIS

RAJA F. NASSAR¹

Department of Genetics, University of California, Davis

Received December 7, 1964

THE theory and analysis of diallel crosses as presented by JINKS and HAYMAN (1953) and JINKS (1954), and extended by HAYMAN (1954), has one of its features the V_r , W_r distribution which is used to investigate simultaneously the genetical properties of a group of homozygous parents. The assumptions underlying this theory are: (i) diploid segregation in the parents; (ii) no difference between reciprocal crosses; (iii) homozygous parents; (iv) two alleles per locus; (v) no nonallelic interaction; (vi) noncorrelated gene distribution between the parents. If all of the above assumptions are met, then the line of regression of W_r , the covariance between the offspring of the r th parent and their nonrecurrent parents, on V_r , the variance of all these offspring, would have a slope of one. The line of unit slope through the origin is that of complete dominance. Movement of the line of unit slope, upward or downward, relative to the line of complete dominance would reveal decreasing or increasing dominance respectively. The intercept of the line on the ordinate is then a measure of the average degree of dominance. A test for the validity of the above model is the constancy of W_r , V_r (HAYMAN 1954). It is obvious that the characteristics of the V_r , W_r distribution are disturbed if any of the listed assumptions fails. HAYMAN (1954b, 1958) discusses what form the distribution may take under failure of these various assumptions. Of particular relevance is the effect of correlated gene distribution. He states that the V_r , W_r curve is convex upward with association and is convex downward with dispersion. The effect is slight and hardly alters the slope of the graph.

It should be emphasized that the correlation of gene distribution is inevitable in any small sample of parental genotypes, and may be a frequent source of deviations from the expectations provided by the model in the absence of any other failure in the list of assumptions.

The purpose of this study is to investigate by computer simulation the magnitude of effects of gene correlation arising from random sampling of parental genotypes on the slope and intercept of the (W_r , V_r) regression line under complete and unidirectional dominance. General and specific combining ability variances were also estimated as outlined by GRIFFING (1956a) to check on bias, variances and distributions of estimates.

¹ Present address: Department of Animal Husbandry, Institute of Agriculture, St. Paul, Minnesota.
This research supported in part by Public Health Service Grant 11778-01.

MATERIALS AND METHODS

A program was written for the IBM 7040 computer to independently make each parental locus 1/1 or 0/0 with a probability of 0.5, according to whether random number was greater than or equal to 0.5, or less than 0.5, each random number being $0 \leq \text{r.n.} < 1$.

The genotype values employed per locus, and genotype frequencies in the base population implied by the method of constructing parental genotypes, are as follows:

Genotype	Genotypic value (G_i)	Frequency in base population
1/1	$2u = 1$	$q^2 = 1/4$
1/0	$u + au = 1$	$2q(1 - q) = 1/2$
0/0	0	$(1 - q)^2 = 1/4$

The total genotypic value for an individual (parent or F_1) was obtained by summing the appropriate values for the ten loci, $G = \sum_{i=1}^{10} G_i$.

To summarize, the system of simulation provided for: (i) homozygosity of parents; (ii) no reciprocal differences or maternal effects; (iii) no epistasis; (iv) diploid segregation; (v) two alleles per locus; (vi) no correlation between loci in gene distribution in the base population from which homozygous parents were derived; (vii) random sample of parents from base population; (viii) no forces to change the gene frequency in the homozygous parents from that of the base population. Correlation between loci in gene distribution in homozygous parents was then the only factor that could account for deviation of the V_r , W_r distribution from expectation.

Environmental effects were not simulated (that is genotypic values were used as phenotypic values) since to have done so would only have increased the random variance of results.

A total of 180 diallels were simulated, as follows:

No. of diallels	No. of parents per diallel
60	6
50	10
50	20
20	40

For each diallel cross the genotypic means of the F_1 progenies were calculated and then analyzed by the regression of W_r on V_r (JINKS 1954; HAYMAN 1954) and by GRIFFING's (1956a) method 3.

RESULTS

V_r , W_r distribution: It is clear from Figure 1 that the frequency distribution of the regression coefficients for the dialleles with varied number of parents showed a consistent downward deviation from 1. As expected, the intensity of the deviation was greatest for the six- and ten-parent diallels and decreased with increase of number of parents. In association with the deviation from a unit slope the regression line almost always intersected the W_r -axis at a point below the origin, thus giving rise to a pseudo-overdominance effect (Figure 2). The intensity of this effect was greatest in the diallel of six parents and decreased with increase in the number of parents. The means of the distributions of regressions and intercepts of Figures 1 and 2 are tabulated below. They clearly show the intensity of deviation from a unit slope of 1 as well as from a zero intercept at the origin.

	Number of parents per diallel			
	6	10	20	40
Regression	$0.8873 \pm .0173$	$0.8742 \pm .0200$	$0.9098 \pm .0089$	$0.9397 \pm .0130$
Intercept	$-0.6838 \pm .0525$	$-0.4142 \pm .0316$	$-0.1096 \pm .0236$	$-0.0655 \pm .0197$

FIGURE 1.—Frequency distributions of regression coefficients of W_r on V_r for diallel size 6, 10, 20 and 40.

General (gca) and specific combining ability (sca) variances: *Bias:* In the conduct of this work, the assumptions required by GRIFFING (1956b) were met. Accordingly, the following should hold in the absence of nonallelic gene interaction.

$$2\sigma_{gca}^2 = \sigma_g^2 \text{ (additive genetic variance)}$$

$$\sigma_{sca}^2 = \sigma_d^2 \text{ (dominance variance)}$$

σ_g^2 and σ_d^2 can be written

$$\sum_{i=1}^n 2q_i(1-q_i) [1+(1-2q_i)a_i]^2 u_i^2$$

FIGURE 2.—Frequency distributions of the intercepts (W_r values when $V_r = 0$) for diallel size 6, 10, 20 and 40.

TABLE 1

Observed means of general and specific combining ability variance estimates for each set of parents

	Number of parents per diallel			
	6	10	20	40
Mean gca	.643 ± .068	.561 ± .044	.651 ± .031	.606 ± .028
Mean sca	.573 ± .033	.605 ± .018	.619 ± .007	.615 ± .007

$$\sum_{i=1}^n 4 q_i^2 (1-q_i)^2 a_i^2 u_i^2$$

where the summation is over the number of loci (COMSTOCK and ROBINSON 1948). Substituting the appropriate numerical values, $u = 1/2$, $a = 1$, $q = 1 - q = 1/2$ and $n = 10$, σ_g^2 and σ_d^2 for the base population are found to be 1.25 and 0.625. The expected value of gca and sca variances is then 0.625.

The observed mean values (Table 1) of the distributions in Figures 3 and 4 as compared to the expected 0.625 reveals that there is no bias in the estimates.

Variance of estimates: When the assumption of normality of effects is met and sampling is random, the variance of any mean square, M is

$$\sigma_M^2 = \frac{2[E(M)]^2}{f} \quad (1)$$

where $E(M)$ is the expectation of the mean square and f , the degrees of freedom on which the mean square is based. When an estimated variance component is derived from two independent mean squares as $(M_1 - M_2)/N$, its variance is

$$\left[\frac{2[E(M_1)]^2}{f_1} + \frac{2[E(M_2)]^2}{f_2} \right] / N^2 \quad (2)$$

It is of importance to know whether the observed standard deviation of the variance component estimates is in accord with formula 2. In calculating the standard deviation, the mean of all mean square estimates per diallel size was used for the expected mean square in formula 2.

TABLE 2

Means of the general and specific combining ability estimates, and the standard deviations around them

	Number of parents per diallel			
	6	10	20	40
gca variance estimates				
empirical	.643 ± .528	.561 ± .313	.651 ± .225	.606 ± .125
formula	.643 ± .505	.561 ± .300	.651 ± .223	.606 ± .141
sca variance estimates				
empirical	.573 ± .262	.605 ± .126	.619 ± .051	.615 ± .032
formula	.573 ± .278	.605 ± .144	.619 ± .067	.615 ± .032

FIGURE 3—Frequency distributions of general combining ability variance estimates for 6, 10, 20 or 40 parents per diallel.

The means of the gca and sca variance estimates and the standard deviations around them, calculated from formula 2 and empirically, are presented in Table 2. There is good agreement between the standard deviations calculated by the two methods.

Distribution of estimates: The distributions of gca and sca variance estimates are shown in Figures 3 and 4. As anticipated the variation among estimates varies inversely with number of parents, and is less for sca than for gca estimates

FIGURE 4—Frequency distributions of specific combining ability estimates for 6, 10, 20 or 40 parents per diallel.

within the same set of parents. This is also revealed from the respective standard deviations as tabulated under variance of estimates.

Fiducial limits: The normal approximation limits referred to by BROSS (1950) were computed for each estimate within a set of parents. These were set at two standard deviations on either side of the estimate. Standard deviation for each estimate was computed from formula (2) by substituting the mean square value for that estimate for the expected mean square, $E(M_i)$.

Table 3 clearly reveals that the distributions of gca and sca variance estimates are skewed to the right when the number of degrees of freedom for estimates is small. It is then that these distributions are similar to the chi square distributions of the mean squares from which the estimates were derived. Greater symmetry in distribution is obtained with more degrees of freedom until the fit is normal. This agrees with the results of COMSTOCK *et al.* (1951) in corn.

Fiducial limits can be computed, as suggested by BROSS (1950) for variance component estimates where the mean squares are based on few degrees of freedom. Such limits were computed for the gca variance estimates when the number of parents was six. Results in terms of the number of times the limits failed to include the mean of the estimates are:

No. of parents per diallel	No. of estimates	Lower limit too high	Upper limit too low
6	60	2	2

Whereas about 19% of the estimates failed to include the mean (Table 3) under the normal approximation limit, only 5% did so under 90% Bross fiducial limits. The agreement in the latter case is reasonable. If, in computing the normal approximation limits, the standard deviation is used (as computed from formula 2 by substituting the average of all mean squares for the expected mean square), then the gca and sca variance estimates for diallels 6 and 10 of Table 3 agree with the limits.

TABLE 3

*Failures of normal approximation limits about individual estimates
to include the means of the estimates*

No. of parents per diallel	No. of estimates	Normal approximation limit	
		lower limit too high	upper limit too low
6	60 gca	0	11
	sca	0	7
10	50 gca	0	8
	sca	0	3
20	50 gca	0	1
	sca	0	0
40	20 gca	0	0
	sca	0	0

TABLE 4

Genotype of parent of the rth array plus the genotypes and frequencies of all parents in that array along with their progenies and expected genetic values

Genotype of parent of rth array	Genotypes (and their frequencies) of all parents			
	<i>AABB</i> (<i>p</i>)	<i>AAbb</i> (<i>r</i>)	<i>aaBB</i> (<i>s</i>)	<i>aabb</i> (<i>t</i>)
Coded genetic values of parental genotypes	$u_1 + u_2$	$u_1 - u_2$	$-u_1 + u_2$	$-u_1 - u_2$
Coded genetic values of progenies				
<i>AABB</i>	$u_1 + u_2$	$u_1 + a_2 u_2$	$a_1 u_1 + u_2$	$a_1 u_1 + a_2 u_2$
<i>AAbb</i>	$u_1 + a_2 u_2$	$u_1 - u_2$	$a_1 u_1 + a_2 u_2$	$a_1 u_1 - u_2$
<i>aaBB</i>	$a_1 u_1 + u_2$	$a_1 u_1 + a_2 u_2$	$-u_1 + u_2$	$-u_1 + a_2 u_2$
<i>aabb</i>	$a_1 u_1 + a_2 u_2$	$a_1 u_1 - u_2$	$-u_1 + a_2 u_2$	$-u_1 - u_2$

Contribution of gene correlation to V_r and W_r : For two loci, and following the method of COMSTOCK *et al.* (1952), table 4 gives the genotype of the parent of the *r*th array and the genotypes and frequencies of all parents in that array, along with their progenies and expected genetic values. Values for V_r and W_r for arrays associated with parents of each of the four genotypes can be derived from Table 4. For example, the contribution of these two loci to V_r for the array associated with an *AAbb* parent is

$$\begin{aligned}
 & [p+r - (p+r)^2] u_1^2 + [(s+t - (s+t)^2) a_1^2 u_1^2 + [p+s - (p+s)^2] a_2^2 u_2^2 \\
 & + [(r+t) - (r+t)^2] u_2^2 + 2 [p - (p+r)(p+s)] u_1 a_2 u_2 - 2 [r - (p+r)(r+t)] \\
 & \quad u_1 u_2 + 2 [s - (p+s)(s+t)] a_1 u_1 a_2 u_2 - 2 [t - (s+t)(r+t)] a_1 u_1 u_2 \\
 & \quad - 2 [(p+r)(s+t) a_1 u_1^2 + 2(p+s)(r+t) a_2 u_2^2].
 \end{aligned}$$

Note that $p + r + s + t = 1$, $p + r = q_1$, $p + r = q_2$, $s + t = 1 - q_1$, and $s + t = 1 - q_2$.

Substituting in the above variance-expression and simplifying we get

$$q_1(1-q_1)(u_1 - a_1 u_1)^2 + q_2(1-q_2)(u_2 + a_2 u_2)^2 + 2(pt-rs)(u_1 - a_1 u_1)(u_2 + a_2 u_2)$$

For the present genetic model, $u_1 = u_2 = a_1 u_1 = a_2 u_2 = 1/2$.

Therefore $V_r(AAbb) = q_2(1-q_2)$.

Following the same procedure for all four parental genotypes one obtains the results of Table 5.

From the above results it is evident that disregarding the correlated gene distri-

TABLE 5

Contribution of correlated gene distribution to V_r and W_r under complete dominance

Genotype	Frequency	W_r	V_r
<i>AABB</i>	<i>p</i>	0	0
<i>AAbb</i>	<i>r</i>	$q_2(1-q_2) + (pt-rs)$	$q_2(1-q_2) + 0$
<i>aaBB</i>	<i>s</i>	$q_1(1-q_1) + (pt-rs)$	$q_1(1-q_1) + 0$
<i>aabb</i>	<i>t</i>	$q_1(1-q_1) + q_2(1-q_2) + 2(pt-rs)$	$q_1(1-q_1) + q_2(1-q_2) + 2(pt-rs)$

bution effect in the term $(pt-rs)$, the contributions to W_r and V_r by each parental genotype are equal according to expectation.

DISCUSSION

Bias in the regression of W_r on V_r : Table 5 shows that for a pair of loci the contribution of correlated gene distribution to W_r is $(r+s+2t)(pt-rs)$ and to V_r is $2t(pt-rs)$. It can be shown that the expectations of $r(pt-rs)$ and $s(pt-rs)$ are negative, and that of $t(pt-rs)$ and $p(pt-rs)$ are positive. The proof for $r(pt-rs)$ follows. (The author is indebted to Dr. R. E. Comstock for his help in this proof.)

The probability distribution of $AABB, AAbb, aaBB, aabb$ (k_1, k_2, k_3, k_4) under random sampling from a population follows a multinomial distribution

$$P(k_1, k_2, k_3, k_4) = \frac{N!}{k_1! k_2! k_3! k_4!} P_1^{k_1} P_2^{k_2} P_3^{k_3} P_4^{k_4}$$

where $k_1 + k_2 + k_3 + k_4 = N =$ number of parents, and $P_1, P_2, P_3, P_4 =$ probabilities corresponding to $AABB, AAbb, aaBB, aabb$.

$$E[r(pt-rs)] = E(ptr) - E(r^2s)$$

$$\text{Let } p = k_1/N, r = k_2/N, s = k_3/N, t = k_4/N.$$

Let $p = k_1/N, r = k_2/N, s = k_3/N, t = k_4/N$.

$$\text{Then } E(ptr) = \sum P(k_1, k_2, k_3, k_4) \frac{k_1 k_2 k_4}{N^3}$$

$$\text{and } E(r^2s) = \sum p(k_1, k_2, k_3, k_4) \frac{k_2 k_3}{N^3} + \sum P(k_1, k_2, k_3, k_4) \frac{k_2(k_2-1)k_3}{N^3}$$

From that it can be shown that

$$E(ptr) = P_1 P_2 P_4 \frac{(N-1)(N-2)}{N^2}$$

$$E(r^2s) = P_2 P_3 \frac{N-1}{N^2} + P_2^2 P_3 \frac{(N-1)(N-2)}{N^2}$$

$$\text{Therefore, } E(ptr) - E(r^2s) = - \left[P_2 P_3 \frac{N-1}{N^2} \right].$$

Following the same procedure it can be shown that

$$E[s(pt-rs)] = - \left[P_2 P_3 \frac{N-1}{N^2} \right]$$

$$E[t(pt-rs)] = E[p(pt-rs)] = \left[P_1 P_4 \frac{N-1}{N^2} \right]$$

$$\text{Therefore, } E[V-W] = 2 P_2 P_3 \frac{(N-1)}{(N^2)}$$

$$\text{For } M \text{ loci, } E [V - W] = 2 P_2 P_3 \frac{(N-1)}{N^2} \frac{m(m-1)}{2}.$$

The consequence of this, with reference to Table 5, is the V_r is almost always larger than W_r . Another thing to consider is that the difference between V_r and W_r is more for large than for small values. This is due to the fact that larger V_r , W_r values are sustained by more recessive genes which contribute to the ($pt-rs$) effect that causes this difference.

If V_r is larger than W_r in almost all cases and the difference between them is proportional to their values, then a regression of W_r on V_r is in most cases below a unit slope. The results have shown that this is so. It is suggested that the bias downward in deviation from unit slope and from a zero intercept is even more intense with overdominance and less so with partial dominance. This is revealed in Table 6 and obtained by substituting $a=1/2$, 2 in the general V_r and W_r expressions. This bias is also less when dominance is not all in one direction.

In the case of no dominance ($a=0$), W_r is always twice V_r and all the (V_r , W_r) points have the same value as is expected under noncorrelated gene distribution. This was, also, borne out by actual data which were not presented in RESULTS.

Judging from the simulation results, correlated gene distribution seems to cause serious disturbances in the V_r , W_r regression line. These are of frequent occurrence especially in diallels of the size often used by plant breeders. Similar deviations in the slope of the regression line have been reported to be due to non-allelic interactions (JINKS 1954, 1955; HAYMAN 1957; JOHNSON 1963). The evidence, however, can never be definitive. One can never know whether deviation from expectations in the (W_r , V_r) graph is due to correlated gene distribution or to another deviation from assumptions (like epistasis) unless number of parents is much larger than practical. It is not uncommon, also, to get negative W_r values (Figure 5), which strongly suggests to the plant breeder that an interaction exists in some of the crosses. The situation becomes even worse when one considers that the disturbances due to gene correlations are exaggerated with increase in the number of loci controlling the character.

TABLE 6

Contribution of correlated gene distribution to V_r and W_r under different degrees of dominance

		<i>AABB</i>	<i>AAbb</i>	Genotype <i>aaBB</i>	<i>aabb</i>
Partial dominance					
$(a = 1/2)$	W_r	$1/2 (pt-rs)$	$(pt-rs)$	$(pt-rs)$	$3/2 (pt-rs)$
	V_r	$1/8 (pt-rs)$	$3/8 (pt-rs)$	$3/8 (pt-rs)$	$9/8 (pt-rs)$
No dominance					
$(a = 0)$	W_r	$(pt-rs)$	$(pt-rs)$	$(pt-rs)$	$(pt-rs)$
	V_r	$1/2 (pt-rs)$	$1/2 (pt-rs)$	$1/2 (pt-rs)$	$1/2 (pt-rs)$
Overdominance					
$(a = 2)$	W_r	$— (pt-rs)$	$(pt-rs)$	$(pt-rs)$	$3 (pt-rs)$
	V_r	$1/2 (pt-rs)$	$-3/2 (pt-rs)$	$-3/2 (pt-rs)$	$9/2 (pt-rs)$

The position of the point (V_r, W_r) on the regression line denotes the relative proportion of dominant and recessive genes in the r th parent. Parents which are recessive correspond to points at the upper end of the line and those that are dominant correspond to points at the lower end of the line. In cases of deviations from a unit slope the ability to estimate the relative proportions of dominant and recessive genes in the parents is reduced owing to the overlap produced in the positions of the V_r, W_r points on the regression line (Figure 5). The intercept (W_r value when $V_r = 0$) is, also, predominantly negative indicating true overdominance when in fact it is pseudo-overdominance.

The results of the gca and sca variance estimates show that one can with fairly low accuracy make inferences about the base population from a diallel of small size. It must be noted that the standard deviations around the means of the gca and sca variance estimates are minimized, since there was no inclusion of environmental variation in the model. Hence, the picture with regard to the accu-

FIGURE 5—Regressions of W_r on V_r for six diallels of size 10. The numbering in the graph of each diallel denotes the number of recessive genes in the r th parent of the V_r, W_r point.

acy of the gca and sca variance estimates is even less favorable than the results suggest.

While this paper was being prepared for publication, HILL (1964) published a paper on the effects of correlated gene distributions in the analysis of diallel crosses. His data were limited to one *Drosophila* experiment in which he adjusted the real diallel table by adding and subtracting whole arrays in the hope of simulating coupling and repulsion effects. HILL concluded that coupling produced an upward curvature on the W_r/V_r graph while dispersion had no effect. His results are not in agreement with my results. However, adding to or subtracting from data in a diallel table is by no means an adequate representation of the true association and dispersion effects, hence any results arrived at cannot be taken to be representative of the situation at large.

The author is indebted to DR. A. S. FRASER for providing the facilities and encouragement in research, DR. R. E. COMSTOCK for his helpful discussions and suggestions in evaluating the results and for helpful criticism of the manuscript, and DR. P. HANSCHKE for the use of his simulation program with modifications to perform the crosses and compute genotypic mean values of the F_1 progenies and for helpful discussions in programming.

SUMMARY

Computer simulation of diallel crosses was carried out to investigate (1) the effect of correlated gene distribution due to small samples on the slope of the W_r/V_r graph, and (2) the variances of the general and specific combining ability parameter estimates arising from differences among samples of parents. The results showed that correlated gene distribution under complete dominance caused serious deviations in the slope of the W_r, V_r regression line and the intercept (W_r value when $V_r = 0$). As a result, the intercept was predominantly below the origin (simulating overdominance) and the regression line almost consistently had a slope less than one. These deviations are sufficiently frequent in occurrence to throw doubt on any inference from diallel cross data.

General and specific combining ability variance estimates were unbiased, but were subjected to large sampling errors when they were computed from mean squares based on few degrees of freedom. This caused a nonsymmetry in the form of the normal distribution. With more degrees of freedom, however, greater symmetry in distribution, as well as reduced sampling errors, was obtained. The observed standard deviations of the estimates agreed well with values obtained from mean values of pertinent mean squares.

LITERATURE CITED

- BROSS, I., 1950 Fiducial intervals for variance components. *Biometrics* **6**: 136-144.
- COMSTOCK, R. E., and H. F. ROBINSON, 1948 The components of genetic variance in populations. *Biometrics* **4**: 254-266. ——— 1951 Consistency of estimates of variance components. *Biometrics* **7**: 75-82. ——— 1952 Estimation of average dominance of genes. pp. 494-516. *Heterosis*. Edited by J. W. GOWEN. Iowa State College Press, Ames.
- GRIFFING, B., 1956a A Concept of general and specific combining ability in relation to diallel

- crossing systems. *Australian J. Biol. Sci.* **9**: 463-493. — 1956b A generalized treatment of the use of diallel crosses in quantitative inheritance. *Heredity* **10**: 31-50.
- HAYMAN, B. I., 1954 The theory and analysis of diallel crosses. *Genetics* **39**: 789-809. — 1957 Interaction, heterosis and diallel crosses. *Genetics* **42**: 336-355 — 1958 The theory and analysis of diallel crosses. II. *Genetics* **43**: 63-85.
- HILL, J., 1964 Effects of correlated gene distributions in the analysis of diallel crosses. *Heredity* **19**: 27-46.
- JINKS, J. L., 1954 The analysis of heritable variation in a diallel cross of *Nicotiana rustica* varieties. *Genetics* **39**: 767-788. — 1955 A survey of the genetical basis of heterosis in a variety of diallel crosses. *Heredity* **9**: 223-238.
- JINKS, J. L., and B. I. HAYMAN, 1953 The analysis of diallel crosses. *Maize Genet. Coop. News Letter* **27**: 48-54.
- JOHNSON, L. P. V., 1963 Applications of the diallel-cross techniques to plant breeding. *Statistical Genetics and Plant Breeding*. Natl. Acad. Sci.-Natl. Res. Council Publ. **982**: 561-570.