THE USE OF INCOMPLETE FAMILY DATA IN THE STUDY OF SELECTION AND POPULATION STRUCTURE IN MARSUPIALS AND DOMESTIC ANIMALS¹

D. W. COOPER

*Department of Animal Breeding, Agricultural College of Sweden, Uppsala 7, Sweden and
the Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.*

Received September 28, 1967

DURING the last decade numerous biochemical polymorphisms in the enzymes and other proteins of many species have been discovered. In these polymorphisms it is usual to find that each genotype has a unique phenotype. Such systems are said to be codominant. The purpose of this paper is to suggest how the genetics of protein polymorphisms (and any other type of codominant system) may be studied using incomplete family data (I.F.D.), i.e., data upon the phenotypes of one parent and its offspring. In marsupials these data are readily available in the field as mother and pouch young, in placental mammals as mother and foetal young, and in domestic animals usually but not invariably as sire groups. The method of analysis is an extension of a simple idea for analyzing sire group data (COOPER 1966).

The use of the Hardy-Weinberg expectations: The discussion of the use of I.F.D. will be clarified if the more usual method of treating field data is first mentioned. This is to compare genotypic numbers observed in a putatively random mating population with those predicted by the Hardy-Weinberg expectations. There is a fundamental weakness in the use of Hardy-Weinberg expectations: they are sensitive to both selective differences between genotypes and to the effects of population structure. The problems are best exemplified by reference to domestic animal populations.

For example, among cattle and sheep, there are two ways in which the genotypic numbers may depart from the Hardy-Weinberg expectations without there being any selection. Both stem from the restricted numbers of male parents in any one generation in the typical breeding groups. This restriction allows considerable gene frequency drift to occur between breeding groups separated from the common parental population by only a few generations. If Hardy-Weinberg expectations are calculated on the overall gene frequency for a population subdivided into different subpopulations with different gene frequencies, the proportion of heterozygotes will be overestimated. This was first demonstrated by WAHLUND (1928) and so has been referred to as Wahlund's effect.

Within each subpopulation the restricted number of male parents allows

¹ Paper No. 1229, Laboratory of Genetics, University of Wisconsin. Supported in part by grant AI-03204 from the National Institutes of Health.

TABLE 1
Expected ratios for I.F.D. where the data are derived from different subpopulations with different gene frequencies

Parental genotype	Parental expected frequency	Progeny genotypes and their expected frequencies		
		A_1A_1	A_1A_2	A_2A_2
A_1A_1	$p^2 + V$	$p^3 + 3pV + M$	$p^2(1-p) + (1-3p)V - M$	
A_1A_2	$2p(1-p) - 2V$	$O_{1,1}$ $p^2(1-p) + (1-3p)V - M$	$O_{1,12}$ $p(1-p) - V$	$p(1-p)^2 - (2-3p)V + M$
A_2A_2	$(1-p)^2 + V$	$O_{12,1}$	$O_{12,12}$ $p(1-p)^2 - (2-3p)V + M$ $O_{2,12}$	$O_{12,2}$ $(1-p)^3 + 3(1-p)V - M$ $O_{2,2}$

p = gene frequency

V = variance in gene frequency between subpopulations

M = third moment of gene frequency

chance differences in gene frequency to arise between the sexes. When the gene frequency among the progeny within such a subpopulation is used to calculate the Hardy-Weinberg expectations, the proportion of heterozygotes will tend to be underestimated. ROBERTSON (1965) has drawn attention to this possibility in cattle populations and so it will be referred to as the Robertson effect.

The extent to which these problems might be encountered in wild populations will usually be entirely a matter of conjecture. In any case it will be demonstrated that the use of I.F.D. allows the effects of population structure to be distinguished from those of selection.

The method of analysis: The treatment will be confined to the two allele case, which is sufficient to illustrate the main issues involved. Consider that there are two codominant alleles A_1 and A_2 with frequencies p and $1-p$. Consider that I.F.D. pertaining to this locus exist.

Then, on the hypothesis that the data come from a large random mating population, the expectations for the ratios of the genotypes in the progeny from the three possible classes of parental genotypes are as in Table 1 when $V = M = 0$. The symbol O indicates progeny and the subscripts indicate genotype of parent and offspring, e.g., $O_{1,12}$ is the number of A_1A_2 offspring from A_1A_1 male parents. It is obvious that

$$\left. \begin{array}{l} (1) O_{12,12} : [O_{12,1} + O_{12,2}] = 1:1 \\ (2) O_{1,1} : O_{1,12} \\ O_{12,1} : O_{12,2} \\ O_{2,12} : O_{2,2} \end{array} \right\} = p:1-p$$

(3) the ratio of the number of A_1 to A_2 genes in the parents of known genotype is also $p:1-p$

The first ratio will be referred to as the *1:1 ratio*, the next three as the *gametic gene ratios*, and the last one as the *parental gene ratio*. The sum of the gametic gene ratios will be called the *total gametic gene ratio*.

In Table 2 is set out a series of χ^2 tests which can be performed using the ratios in Table 1 (when $V = M = 0$) as expected values. Several factors may cause a body of I.F.D. to depart from such expectations and these can be conveniently discussed under the headings below.

1. *Selection:* The detection of the action of selection on the progeny using the 1:1 ratio has been discussed before (COOPER 1966). That this ratio is affected only by selection on the progeny and not by population structure constitutes one of the chief advantages of I.F.D. over population data. The gametic gene ratios will also be affected by selective differences among the progeny. If there is a significant departure from the 1:1 ratio, it is likely that there will be significant heterogeneity among the gametic gene ratios.

2. *Multiparity:* When carrying out the heterogeneity test upon the gametic gene ratios (Test (b)) upon data from organisms which are multiparous, it may be necessary to modify the usual χ^2 test. The word *multiparous* can be used in two senses: (a) to describe a female which has had more than one offspring and (b) to describe a female which produces more than one offspring as a result of a

TABLE 2

χ^2 tests which can be carried out on incomplete family data (I.F.D.) and null hypotheses which are tested

Test	Null hypotheses tested
(a) χ^2_1 on an expected 1:1 for ($O_{12.1} + O_{12.2}$) : $O_{12.12}$.	No selective differences between A_1A_2 and A_1A_1 and/or A_2A_2 .
(b) χ^2_2 for heterogeneity of the individual gametic ratios using total gametic gene ratio as the expected, i.e., on $O_{1.1} : O_{1.12} ; O_{12.1} : O_{12.2} ;$ $O_{2.12} : O_{2.2}$.	May test null hypothesis in a), <i>and</i> that there exists no subdivision of the population such that p is not constant <i>and</i> that the animal is not multiparous and does not tend to be monogamous.
(c) χ^2_1 for agreement with HARDY-WEINBERG expectations of parental genotypic ratios.	All above null hypotheses except inflation of χ^2 due to multiparity.

See text for explanation of O symbols.

single mating. χ^2 must always be modified for data from a species with the latter kind of multiparity and will be necessary for the former kind when the species is monogamous, or tends to be so. The modification is necessary because it is an assumption of the heterogeneity test on the gametic gene ratios that in producing their offspring, the parents whose phenotypes are known should have as their mates individuals whose array of genotypes is a random sample of that of the parental population. This assumption is violated when the production of one offspring by any pair of animals makes it more likely, on a chance basis, that the same pair will produce further offspring. For example, data from cattle will usually not have this problem but those from pigs, man, the opossum, and all Dasyuridae (marsupial mice, etc.) do. In obtaining I.F.D. from these species we have first divided the parental population into three parts (according to whether their mates are A_1A_1 , A_1A_2 , or A_2A_2), a process which gives rise to chance differences in gene frequency, and then inflated the statistical significance of these differences by continued sampling with replacement as gametes are produced.

Alternatively, the unmodified use of χ^2 on such data may be looked upon as violating the principle that for χ^2 to be applicable each of the observations must be statistically independent. As the designation *gametic gene ratios* implies, we are in effect examining the ratio of the alleles in three samples of the gametes of the unexamined parent. In a strictly monogamous population or in a single litter the maximum number of statistically independent "observations" which can be made for each unexamined parent is 2, i.e., each parent has two alleles at any one locus. Of these two genes in the unexamined parents $2(1-(1/2)^s)$ will be represented on the average at least once among s sibs. Or, in other words, though s sibs contain s genes from their unexamined parent, only $2(1-(1/2)^s)$ of these are non-identical genes. Notice that it does not matter whether the unexamined parent is heterozygous or homozygous—in either case this parent has two non-

identical, statistically independent genes. If we make use of this simple weight, an expression for the χ^2 for heterogeneity of gametic gene ratios can be obtained which allows for the lack of independence among the observations.

Let t_{11} , t_{12} , and t_{22} be the number of A_1A_1 , A_1A_2 , and A_2A_2 parents, f_{11x} , f_{12x} , and f_{22x} be the proportion of A_1A_1 , A_1A_2 , and A_2A_2 parents with x offspring, and p_{11} , p_{12} , p_{22} be the proportion of the A_1 allele among the unexamined mates of A_1A_1 , A_1A_2 , and A_2A_2 parents ($p_{11} = O_{1.1}/(O_{1.1} + O_{1.12})$, $p_{12} = O_{12.1}/(O_{12.1} + O_{12.2})$, $p_{22} = O_{2.12}/(O_{2.12} + O_{2.2})$). Then we may calculate three quantities T_{11} , T_{12} , and T_{22} as follows; for example

$$T_{11} = 2t_{11} \sum f_{11x} (1 - (1/2)^x)$$

χ^2 for heterogeneity of gametic gene ratios is then correctly calculated as

$$\frac{1}{p(1-p)} [T_{11} (p_{11}-p)^2 + T_{12} (p_{12}-p)^2 + T_{22} (p_{22}-p)^2]$$

where p is, as before, the proportion of A_1 among all unexamined parents.

The necessity of allowing for this underestimation of the variance means that collecting I.F.D. becomes progressively less efficient as the family size increases. Efficiency is measured by the fraction of the two non-identical genes added by each successive sib (last two columns of Table 3). This fraction will depend upon the genotype of the examined parent. For each family of a homozygous examined parent, the average number of non-identical genes of the unexamined parent contributed by s sibs is

$$2(1 - (1/2)^s).$$

But for a heterozygous examined parent, s sibs contribute only

$$2(1 - (3/4)^s),$$

because the expectation for heterozygous progeny of heterozygous parents is not dependent upon gene frequency and, therefore, contributes no information about the genes of the unexamined parent. If H = the proportion of heterozygotes in the population, then the average fraction of the two non-identical genes of the unexamined parents represented by a random family of s sibs is

$$Y_{(s)} = 2(1-H)(1 - (1/2)^s) + 2H(1 - (3/4)^s).$$

TABLE 3

Efficiency of typing related to size of sibship (s) and frequency of heterozygotes (H) when the species is multiparous*

No. of offspring (s)	Cumulative total of non-identical genes $Y(s)$		Number of non-identical genes added by each sib ($Y(s) - Y(s-1)$)	
	$H = 0$	$H = 1/2$	$H = 0$	$H = 1/2$
1	1	0.75	1	0.75
2	1.5	1.19	0.5	0.44
3	1.75	1.45	0.25	0.27
4	1.88	1.62	0.13	0.17
5	1.94	1.73	0.06	0.10
6	1.97	1.81	0.03	0.08

* Measured as the no. of non-identical genes of the unexamined parent added by each successive sib, $Y(s) - Y(s-1)$. See text for further explanation of $Y(s)$.

This expression has its least value at $H = 1/2$ and its greatest at $H = 0$, ($0 \geq p \geq 1$, and if one assumes that H cannot be greater than $1/2$). Table 3 gives values for $Y_{(s)}$ for sibship sizes from 1 to 6 and $H = 0$ and $1/2$. As far as the heterogeneity χ^2 is concerned, there is little point in typing more than five or six sibs. Very little sensitivity is added to the test by doing so.

It should be emphasized that it is not the number of offspring per parent nor the identicalness of the genes sampled which are *per se* the point here. If every parent had many offspring, it would not be necessary to go through the procedure outlined above if, after the production of one offspring, mates were exchanged at random throughout the population.

It should also be emphasized that there is no limit to number of offspring which can be usefully typed for a heterozygous examined parent, because the 1:1 ratio is unaffected by multiparity.

3. *Subdivision or stratification of the population—the Wahlund effect:* If I.F.D. from a large population composed of subpopulations with differing gene frequencies are considered as a unit, the gametic gene ratios will show heterogeneity in Test (b). The expectations for such a population are to be found in Table 1 when V , the variance of the gene frequency in the population, assumes a positive value and M the third moment of the gene frequency about its mean, may be positive, negative or zero. Let p_i be the frequency of the gene A_1 in the i^{th} subpopulation, and k_i be the relative size of the i^{th} subpopulation, $\sum k_i = 1$. Then the expectations in Table 1 are obtained quite simply. For example, the proportion of A_1A_1 offspring from A_1A_1 parents in the i^{th} population is p_i^2 , and for the whole population is $\sum k_i p_i^2$. Now

$$\sum k_i p_i^3 = \sum k_i (p + (p_i - p))^3 = p^3 + 3pV + M$$

where

$$V = \sum (k_i p_i^2) - p^2$$

$$M = \sum k_i (p_i - p)^3$$

The other six elements of Table 1 are derived in the same way. In a subdivided population, the ratio among the progeny of A_1A_1 parents will tend to indicate a higher value of the frequency of A_1 while the ratio among the progeny of A_2A_2 will tend to indicate a lower value. The value indicated by the ratio among the progeny of A_1A_2 will be intermediate, tending to depart from the real value for the whole population, p , in a direction which will depend upon the value of p . There will also, of course, be a tendency for a significant deficiency of heterozygotes to occur when the parental genotypic numbers are examined for agreement with the Hardy-Weinberg expectations.

In addition to V , the ratios in Table 3 include M , the third moment of p_i about its mean p . The third moment is a measure of the departure of the distribution of p_i from symmetry. It will be 0 if the distribution is symmetrical and will take positive or negative values according to the direction of skewness. Generally speaking, we can expect that in the neighborhood of $p = 1/2$, there will be no skewness but that for extreme values there may be. KIMURA (1955) has shown that the frequency distribution of gene frequencies under random

genetic drift is skewed when the initial gene frequency departs from 50%, although doubtless selection for an equilibrium value will tend to eliminate extreme values and so some of the skewness. In any case, M will be small in comparison to V for most situations, and can be safely ignored when estimating V .

If both selection and the Wahlund effect occur, the direction of the departures observed in each test will differ from those where only one of the two occur. For example, suppose a selection for heterozygotes and the Wahlund effect exist. The former will be apparent from the 1:1 test. In the gametic gene ratio heterogeneity test the two will tend to reinforce each other while in the Hardy-Weinberg test they will operate in opposite directions. Selection alone will give significance in all three tests while the Wahlund effect will give it only in the latter two, with a deficiency of heterozygotes in the Hardy-Weinberg test.

4. *Sex differences in gene frequency—the Robertson effect:* These can occur in animal breeding when male and female parents come from different parts of the same breed or from different breeds altogether (for an example, see RENDEL 1958). It is theoretically possible that sex differences in selection coefficients could give rise to sex differences in gene frequency (OWEN 1953), but it is questionable whether such situations will be detected. It is also possible that in a population with a very small effective population size, random differences in gene frequency between the sexes could result in significant differences in the total gametic gene ratios for each sex of parent (ROBERTSON 1965).

In general, the use of I.F.D. circumvents the Robertson effect. However, it is important, when dealing with domestic animals, that data for male and female parents be analyzed separately. The 1:1 ratio will not be affected by differences in parental gene frequency but the gametic gene ratios will be, as can be seen from examples like the following extreme one: If all female parents were A_1A_1 and all male parents A_2A_2 the ratio of $A_1A_1:A_1A_2$ among the progeny of A_1A_1 would be 0:1 and among the progeny of A_2A_2 the ratio of $A_1A_2:A_2A_2$ would be 1:0. The ratio of $p:1-p$ in both sexes of parents combined and among the whole of the progeny is 1:1, given that both sexes of parents reproduce equally and are equally represented in parents and progeny.

Other considerations: A possible source of error in dealing with I.F.D. is mistaken parentage, either through fostering or errors in recording. In cattle data derived from a blood group laboratory, many of these will be detected by the blood typing. In most other situations, however, the only way of detecting such errors will be through segregation at the locus under investigation. As with complete family data not all will be detected. If parents and/or offspring are interchanged at random, the expected proportion of all errors occurring which will be detected in I.F.D. is

$$2p^2(1-p)^2$$

which has a maximum value $1/8$ at $p = 1/2$, a value which is not greatly affected if the Hardy-Weinberg expectations are not strictly valid for the population. It seems unlikely that any departure from expectation could be due to misclassification unless there is a large number of errors in a large body of data, in which

case an estimate of the percentage of errors could be obtained from the data themselves, i.e., at least one will be detected showing that the data are not completely reliable.

One last point should be mentioned. The foregoing method of analysis depends upon the fact that among the progeny of parents of known genotype, there are expected ratios, i.e., from the rows of Table 1, one can obtain expected ratios. It is also true that analogous expected ratios can be obtained from the columns and diagonals of Table 1.

Specifically they are

- (1) A_1A_2 produced by A_1A_1 and A_2A_2 parents : A_1A_2 produced by A_1A_2 parents = 1:1
 A_1A_2 produced by A_1A_1 parents : A_1A_1 produced by A_1A_2 parents = 1:1
 A_2A_2 produced by A_1A_2 parents : A_1A_2 produced by A_2A_2 parents = 1:1
- (2) A_1A_1 produced by A_1A_1 parents : A_1A_1 produced by A_1A_2 parents = $p:1-p$
 A_1A_2 produced by A_1A_1 parents : A_1A_2 produced by A_2A_2 parents = $p:1-p$
 A_2A_2 produced by A_1A_2 parents : A_2A_2 produced by A_2A_2 parents = $p:1-p$

The progeny data contain only four degrees of freedom and we have used all of these in the analyses in Table 2. But even if this were not the case the column ratios would have to be treated with circumspection. The relationships obtained from the rows of Table 1 give a measure of the viability components of selection upon the progeny, while the column ratios are measures of viability (of parents, not offspring) and of the average number of offspring from each genotype. This last may not always be the same thing as fertility. The column ratios do not take into account the number of parents of each genotype. Thus, a χ^2 test on the column 1:1 ratio may allow one to conclude, quite validly, that more A_1A_2 progeny come from the A_1A_2 parental genotype than expected; but if there is only one A_1A_2 parent, conclusions about the fertility of A_1A_2 parents in general are quite out of order. The correct way of testing for fertility (or parental viability) differences is to estimate the mean and variance of the number of offspring produced by each genotype and then make the appropriate significance tests.

DISCUSSION

The ideas outlined above are elementary and not particularly new. No originality is claimed for them. What is claimed is that their significance for the study of selection in natural and domestic populations has been almost entirely overlooked. The basic dilemma in using the Hardy-Weinberg expectations applied to population data is that departures due to population structure cannot be separated from those due to selection. It should be apparent from this paper that incomplete family data (I.F.D.) allow these effects to be assessed separately. Thus, one of the principal uncertainties in dealing with the operation of natural and artificial selection can be removed by collecting I.F.D. instead of population data.

The usefulness of I.F.D. does not end at this point. Though a full treatment of the subject has not been attempted here, it is worth noting that I.F.D. possess sufficient degrees of freedom to allow estimates not only of the gene frequency

but also of V , the variance in gene frequency between subpopulations, and selection coefficients. The best statistical method for that purpose would be the maximum likelihood technique (see FISHER 1946; HOLM 1966; MATHER 1951; and BAILEY 1961 for general treatments of the method for genetical work). The three parental and seven progeny classes of Table 1 have between them six degrees of freedom, making it possible to estimate six parameters. If the third moment (Table 1) is assumed to be 0, then these six degrees of freedom can be used to estimate p , V , and the selection coefficients in parents and progeny of genotypes A_1A_1 and A_2A_2 relative to the heterozygote in each case.

Such estimations allow explicit tests to be made of the null hypotheses that the last five parameters are in accord with no subdivision and no selection. This represents a slight advantage over the χ^2 tests suggested above, where a just significant value for one selection coefficient might not give a significant χ^2 because these test several parameters jointly. Furthermore, estimation will allow the comparison of different bodies of data as they accumulate in the literature. Tests of heterogeneity for the different estimates of each parameter can be carried out. Again we may compare population data where each body of data must be judged essentially on its own.

No attempts have been made in the present paper to treat the possibility of using I.F.D. for the cases where multiple alleles, sex linkage or sex limited polymorphism are found. The extension of the idea should be fairly obvious for these cases.

The ideas outlined here have been used by COOPER and RENDEL (1968) in a study of blood groups in Swedish cattle and by NIECE and KRACHT (1967) in a study of ass transferrins. Much of the data accumulated by other cattle blood group laboratories could receive the same treatment.

The genetics of marsupials have been rather neglected. In several Australian laboratories some work is now in progress on protein polymorphisms and so I.F.D. could be exploited for these studies. COOPER and SHARMAN (1964) have drawn attention to this possibility and presented a small amount of I.F.D. It may also be remarked that the existence of such data to some extent compensates for the lack of an established laboratory marsupial suitable for genetic work.

I am indebted to PROFESSOR J. H. BENNETT who first drew my attention to the possibilities of I.F.D. I am also indebted to DRs. J. RENDEL and B. GAHNE who helped me clarify several drafts of the manuscript and to PROFESSORS W. H. STONE and J. F. CROW and DR. CARTER DENNISTON and MR. K. K. KIDD for critically examining the final draft. While in Uppsala the writer was supported by an Australian C.S.I.R.O. Overseas Studentship.

SUMMARY

In recent years, many genetically controlled protein and enzyme polymorphisms have been discovered. The genes involved usually do not exhibit dominance. This paper points out that selection and population structure for such polymorphisms can be easily and separately studied if incomplete family data (I.F.D.) are collected. Such data, records of one parent and offspring, are easily obtainable in populations of marsupials and placental mammals. They are to be

preferred to simple population data, in which the effects of population structure cannot be distinguished from those of selection.

LITERATURE CITED

- BAILEY, N. T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford.
- COOPER, D. W., 1966 A note on the examination of genotypic ratios in domestic animals using incomplete family data. *Animal Production* **8**: 511-513.
- COOPER, D. W., and J. RENDEL, 1968 Incomplete family data, selection and population studies of transferrins and blood groups in cattle. *Heredity* **23**: 49-66.
- COOPER, D. W., and G. B. SHARMAN, 1964 Transferrin variation in kangaroos. *Nature* **203**: 1094.
- FISHER, R. A., 1935 The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugenics* **6**: 13-25. — 1946 A system of scoring linkage data, with special reference to the pied factors in mice. *Am. Naturalist* **80**: 568-578.
- HOLM, G., 1966 Studies on chlorophyll mutations in barley: Methods of measuring linkage by single and combined maximum likelihood formulas solved by use of an automatic computer and the programming language Algol 60. *Hereditas* **55**: 77-120.
- KIMURA, M., 1955 Random genetic drift in multi-allelic locus. *Evolution* **9**: 419-435.
- MATHER, K., 1951 *The Measurement of Linkage in Heredity*. Methuen, London.
- NIECE, R. L., and D. W. KRACHT, 1967 Genetics of transferrins in burros (*Equus asinus*). *Genetics* **57**: 837-841.
- OWEN, A. R. G., 1953 A genetical system admitting of two distinct stable equilibria under natural selection. *Heredity* **7**: 97-102.
- RENDEL, J., 1958 Studies of cattle blood groups. IV. The frequency of blood group genes in Swedish cattle breeds, with special reference to breed structure. *Acta Agr. Scand.* **8**: 189-215.
- ROBERTSON, A., 1965 The interpretation of genotypic ratios in domestic animal populations. *Animal Production* **7**: 319-324.
- WAHLUND, S., 1928 Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre betrachtet. *Hereditas* **11**: 65-106.