S T U D I E S

I N

S T A T I S T I C A L G E N E T I C S

by

PREM NARAIN

M.Sc. (LUCKNOW)

Ph.D. (EDINBURGH)

Thesis presented for the Degree of
Doctor of Science of the University
of Edinburgh in the Faculty of Science

June, 1984

C O N T E N T S

STATEMENT i

ACKNOWLEDGEMENTS ii

ABSTRACT iv

INTRODUCTION 1

SECTION ONE 27

Studies in Statistical Genetics
- Published work

SECTION TWO 158

Studies in Statistical Genetics
- Unpublished work

APPENDIX 193

Twenty-three Reprints

STATEMENT

This is to certify that:

(a) the results published in the papers discussed at Serial Nos. 5, 6 and 13 in Section ONE of this thesis were earlier incorporated in my Ph.D. thesis RESPONSE TO SELECTION IN FINITE POPULATIONS submitted to Edinburgh University in 1969, and

(b) all the works reported in this thesis, except those jointly contributed, were entirely and independently contributed by me. In the jointly contributed works at Serial Nos. 3, 5, 11, 19, 21, 22 and 23 in Section ONE and at I in Section TWO, substantial contribution involving formulation of the problem, necessary mathematical derivations, extraction of results and writing of the papers were made by me. The collaborators provided support in the form of discussion and finalisation of the papers.

(PREM NARAIN)

ACKNOWLEDGEMENTS

I was inspired to work in the field of STATISTICAL GENETICS by several people at various stages.

To begin with, while working in a Drosophila Laboratory at Indian Veterinary Research Institute, Izatnagar (U.P.), India, I was encouraged by late Professor Theodosius Dobzhansky, Columbia University, New York, U.S.A. to take up a career in Population Genetics. Prof. Dobzhansky was then on a visit to India for the Indian Science Congress meeting in January, 1960.

It was Dr. P. Bhattacharya, the then Head of Division of Animal Genetics at I.V.R.I., Izatnagar who was instrumental in kindling my interest in this field.

At the Institute of Agricultural Research Statistics, New Delhi, which I joined in 1961 as Assistant Professor, I got inspiration from late Dr. V.G. Panse, the then Statistical Adviser for working on problems relating to theoretical aspects of Quantitative Genetics.

In 1967, I came to Institute of Animal Genetics, Edinburgh, as a Colombo Plan Scholar and worked, for my Ph.D., with Professor Alan Robertson, OBE, FRS. It was at this place that I got a systematic exposure to the subject which helped me considerably to broaden my knowledge and to pave the way for further researches on my return to India in 1969.

In February-March 1976, I got an opportunity to visit Department of Statistics, Iowa State University, Ames, U.S.A. as a Research Associate where I interacted fruitfully with Professors Oscar Kempthorne and Edward Pollak on statistico-genetic problems.

From November 1979 to March 1980, I was at the Centre for Demographic and Population Genetics, University of Texas at Houston, U.S.A. as a Visiting Professor. Here I worked with Professors Masatoshi Nei and Ranajit Chakraborty.

I take this opportunity to express my gratitude and indebtedness to the above mentioned people as well as several others with whom I came in contact while working in this field, for help and valuable discussions.

I am also grateful to Mr. Amar Ranjan Paul for drawing the figures.

It is my special pleasure to thank Mr. Subhash Chandra Gupta for his careful and cheerful work in typing the manuscript.

Last but not the least, I acknowledge and express sincere appreciation of the role played by my wife Krishna in extending moral support to me over the last about two and a half decades thereby enabling me to devote whole-heartedly to the subject.

ABSTRACT

1. The contributions published during 1962 to 1984 in statistical genetics were presented. While a brief survey of the work involving applications to Drosophila, plants, animals and man was made in Introduction, twenty-three papers pertaining to theoretical statistical genetics were selected for detailed presentation in Section ONE. In addition, two unpublished works were presented in Section TWO.

2. The twenty-three publications selected pertain broadly to the three areas, genetic properties of population (4), stochastic processes in population genetics (10) and genetics of quantitative variability (9). In the first case, papers deal with effects of linkage on the homozygosity of a selfed population as well as of a population under mixed selfing and random mating, generalisation of heterozygote x homozygote mating and Fisher's Fundamental Theorem of Natural Selection. In the second case, six out of ten papers illustrate the use of Markov chain introducing a new concept of duration of response to selection, conditional Markov chain approach, the probability of fixation under random fluctuations in selection intensity and the average time until fixation for a tri-allelic locus. The rest four papers use diffusion approximation approach for similar problems, introduce conditioned diffusion equations, discuss the problem of average age of a mutant in

finite population using the conditional approach and give estimates of heterozygosity in the context of molecular theory of evolution based on conditional arguments. In third case, papers deal with multiple allelic component of phenotypic variance, generalised heritability and response to selection for several characters, optimum group size in progeny testing, phenotypic index with several auxiliary traits, sire index corrected for an auxiliary character, the use of auxiliary traits in combined selection and partial diallel crosses.

3. The work on Genetic differentiation of quantitative characters between populations discusses a new model involving optimal selection and discrete allelic states for mutation in an infinitely large population. The behaviour of the ratio of inter- to intra-population variances over time is found to be helpful in testing the hypothesis of neutrality. The work on Progeny testing with auxiliary traits deals with a general theory of progeny testing where several auxiliary traits are considered along with the main trait to predict the breeding value of a male for the main trait. With one auxiliary trait, the accuracy of the progeny test is always increased and the number of progeny required for a pre-assigned value of the accuracy gets reduced. These gains are found to be substantial when genetic and phenotypic correlations between the main and the auxiliary traits are of opposite signs.

I N T R O D U C T I O N

C O N T E N T S

	GENERAL	2
I	THEORETICAL STATISTICAL GENETICS	3
II	APPLIED STATISTICAL GENETICS	4
III	HANDBOOK OF STATISTICAL GENETICS	10
	LIST OF PUBLICATIONS	11

My major field of study since 1958 has been Statistical Genetics. In addition, I have contributed to the field of Statistical methods applied to agriculture. My contributions in the field of statistical genetics have been both in theory and applied areas. In the former area, I was conferred the Degree of Ph.D. on 25th October, 1969 on the dissertation entitled RESPONSE TO SELECTION IN FINITE POPULATIONS by the Edinburgh University. For this Degree, I worked under the supervision of Prof. Alan Robertson, F.R.S. at the Institute of Animal Genetics, Edinburgh, during 1967 to 1969.

I append a list of my scientific contributions which total 146 and which were published during 1962 to 1984. Their distribution according to the two fields is

Statistical Genetics	105
Statistical methods applied to agriculture	41

In the field of statistical genetics, their distribution according to theory and applied aspects is

Theoretical Statistical Genetics	60
Applied Statistical Genetics	45

I. THEORETICAL STATISTICAL GENETICS

Of the 60 publications in this field, I am the sole or the senior author in 41 cases. Out of these 41, I have selected 23 papers for presentation in this thesis. These are described in Section 1 entitled Studies in Statistical Genetics - published work. In the list of publications at pages 11-26 these are marked as asterisks (**). In addition, in Section 2 are presented two unpublished works in one of which I am the sole author while in the other, the senior author. One of the two papers has been accepted for publication. The other one was presented during the XV International Congress of Genetics at New Delhi (India) in December, 1983. It has since been submitted for publication.

II. APPLIED STATISTICAL GENETICS

In 30 of 45 publications in this field, I am the sole or the senior author. None of these has been included in this thesis. However, a brief survey in respect of 29 out of 45 publications is given below. In addition, the survey also gives references to 5 publications pertaining to the field of theoretical statistical genetics for completing the description. These contributions are grouped, according to the species of application, into four heads viz. Drosophila, plants, animals and man.

1. Drosophila

I worked at the Indian Veterinary Research Institute, Izatnagar, U.P. (India) during 1958 to 1961 in a research scheme 'Fundamental study of Quantitative Inheritance' in which Drosophila was used as a tool with the ultimate object of developing models for the inheritance of quantitative characters for ultimate use in large animal selection experiments. We published results on the response to selection for fecundity in Drosophila melanogaster (Narain, 1962; Narain et al 1962; Prabhu et al, 1964) and on the application of diallel crosses in such studies (Singh et al 1964; Prabhu et al 1968). Subsequently, we published results on the estimation of components of variation due to interaction between genotype and temperature for wing

length and bristle number (Narain and Reddy, 1970).

2. Plants

With the help of data on 7x7 diallel cross in wheat and 8x8 diallel cross in triticale collected from Indian Agricultural Research Institute, New Delhi (India), statistical techniques for analysing complete and partial diallel crosses involving several characters were investigated and the use of discriminant function in such studies was demonstrated (Subba rao and Narain, 1976; Batra and Narain, 1980).

3. Animals

Under this head, the contributions are further sub-grouped into (a) cattle and buffaloes, (b) sheep, and (c) poultry.

(a) Cattle and buffaloes: With the help of data on Indian cattle collected from different organised farms in the country, it was found that the genetic gain in the first lactation milk yield could be increased by about 3 to 11 per cent if selection is based on a 'phenotypic index' in which the main trait is expressed as deviations from the expected values predicted with the help of one or more auxiliary traits (Narain and Mishra, 1975). A new sire index based on this technique was found to be superior over others when applied to data on Sahiwal breed of Indian cattle (Kumar and Narain, 1980). Statistical

techniques were used for separating genetic from environmental trends using records maintained in dairy herds over several generations of selection. The average genetic change in a trait was estimated as twice the pooled intra-sire intra-generation regression coefficient of the weighted difference between the herd and the individual sire means on the years (Narain and Garg, 1972). Some aspects of yield survival relationship in dairy cattle were studied (Narain and Bhatia, 1979). The efficiency of indirect selection for life-time production was also studied (Narain et al, 1975). A plan for evolving a dairy breed making use of animals of different grades due to crossing of Friesian bulls with Sahiwal cows, available at Military Dairy Farms in India, was prepared (Narain, 1977). A proposal for undertaking cross-breeding among three important breeds of Indian buffaloes to combine the desirable characters of milk production and fat percentage was made and a corresponding breeding plan developed (Narain, 1980). Using the data on different grades of cross-bred animals available at Military Farms, the optimum level of exotic inheritance for stabilising the breed was studied in relation to milk production and calving interval (Narain and Garg, 1979). Lactation performance indices in Sahiwal and Hariana cattle were constructed and studied by maximising the variation for the index between animals relative to that within animals (Narain and Chand, 1980; Chand and

Narain, 1983). A series of studies on Sahiwal cattle and Murrah buffaloes located at Chak-Ganjaria Farm at Lucknow, U.P. (India) was also undertaken (Kumar and Narain, 1977, 1978a and b, 1979). Investigations were undertaken to study the association between the immunogenetic traits such as blood type and economic traits such as milk yield in cattle and buffaloes so that the performance of an animal could be predicted on the examination of its blood type and the decision for the selection of the animal could be made at an early stage (Singh et al, 1981).

(b) Sheep: With the help of data collected at Sheep Breeding Farm under the Scheme for improvement of sheep and wool by crossing Kashmiri ewes with Rambouillet rams, an investigation on the relationship between the retention of a sheep in the flock and its wool yield in the initial clip was undertaken (Bhatia and Narain, 1973). With the help of the same data, the use of discriminant function and D^2-statistic in a cross-breeding programme with sheep was demonstrated. It was found that the discriminating power of the index based on greasy-fleece weight, fibre diameter, fibre length and fleece density was much higher than those on the basis of the individual traits separately (Narain and Garg, 1975). For dealing with the case of unequal variance-covariance matrices with such data, an alternative linear procedure which minimises the probabilities of mis-classification and is a minimax procedure was later used (Narain and Malhotra, 1979).

(c) Poultry: The Government of India, in collaboration with State Governments, initiated a series of Coordinated Poultry Breeding Programmes for improvement of egg production in the country. This involved evolving a strain of poultry with high level of egg production by selecting birds on the basis of Osborne's index which combines in an optimum way, the information on the individual bird with the average performance of the sire and dam families to which the bird belongs. One of the Regional Poultry Farm at Bhopal who initiated this breeding experiment collaborated with me for operating this programme for about seven years during 1972 to 1979. With the help of data so collected, response to selection for rate of lay was studied (Narain et al, 1973 a and b; Malhotra et al, 1974). Hatch and pen effects on some performance traits in White Leghorn were also investigated (Babu et al, 1975).

4. Man

In human nutrition, one talks of either genotypic or environmental variance ignoring the interaction between the two. However, intra-individual variation in protein or energy intake of an individual is found to vary over time and to persist even when data are averaged over a week. A new genetic model for such studies was therefore developed emphasising the importance of the covariance term in the variance of the sum of genotype

and environmental effects over time in the same individual within the framework of ontogenic growth. Using such a model, the genetic significance of intra-individual variation in energy requirement and of the autoregulatory mechanism in nitrogen balance were studied (Sukhatme and Narain, 1982, 1983, 1984; Narain, 1982).

III. HANDBOOK OF STATISTICAL GENETICS

The methodology of statistical genetics was illustrated with the help of fully worked out examples from plants and animals, in the form of a text-book, titled as above, for the benefit of practising statisticians, agricultural scientists as well as post-graduate students. The book was published by Indian Agricultural Statistics Research Institute, New Delhi (Narain *et al*, 1979).

LIST OF PUBLICATIONS

1962

1. NARAIN, P. Effect of age of female on the rate of egg production in Drosophila melanogaster. Drosophila Information Service, 36: 96-97.

2. NARAIN, P., C. JOSHI and S.S. PRABHU. Response to selection for fecundity in Drosophila melanogaster. Drosophile Information Service, 36: 97-99.

3. GOSWAMI, S.B. and P. NARAIN. The effect of air temperature and relative humidity on some physiological indices of buffalo-bulls. Indian J. Veterinary Science and Animal Husbandry, 32: 112-118.

1964

4. PRABHU, S.S., P. NARAIN, S. NEGI, C. JOSHI and H.P. SINGH. Studies in Drosophila egg II. Two-way selection for egg production in Drosophila melanogaster. J. Animal Morphology and Physiology, 11: 199-218.

5. SINGH, H.P., S.S. PRABHU, P. NARAIN and S. NEGI. Analysis of a 6x6 diallel cross for egg production in Drosophila melanogaster. J. Genetics, 59: 58-68.

1965

**6. NARAIN, P. The description of gene action and interaction with multiple alleles in continuous variation. Genetics, 52: 43-53.

7. SARMA, T.C.S.R. and P. NARAIN. Components of continuous variation in presence of epistasis and genotype x environment interactions. J. Indian Society of Agricultural Statistics, 17: 182-198.

1966

** 8. NARAIN, P. Homozygosity in a selfed population with an arbitrary number of linked loci. Genetics, 59: 254-266.

** 9. NARAIN, P. Effect of linkage on homozygosity of a population under mixed selfing and random mating. Genetics, 54: 303-314.

10. NARAIN, P. and V.N. AMBLE. An application of a confounded factorial design in studies on metabolic behaviour of spermatozoa. Indian J. Veterinary Science and Animal Husbandry, 36: 183-188.

1967

11. RAO, A.V. and P. NARAIN. Genetic variance and assortative mating. J. Indian Society of Agricultural Statistics, 19: 31-45.

1968

12. PRABHU, S.S., P. NARAIN, H.P. SINGH and S. RAWAT. Analysis of a 6x6 diallel cross for hatchability in Drosophila melanogaster. J. Genetics, 60: 17-23.

1969

** 13. NARAIN, P. and A. ROBERTSON. Limits and duration of response to selection in finite populations: the use of transition probability matrices. Indian J. Heredity, 1: 79-97.

1970

** 14. NARAIN, P. A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. Genetical Research, 15: 251-255.

15. NARAIN, P. and K.M. REDDY. A note on the estimation of components of variation due to interaction between genotype and temperature for wing length and bristle number in Drosophila ananassae. J. Genetics, 60: 178-183.

16. NARAIN, P. Influence of breeding on the efficiency of livestock production. Proc. Seminar on 'Methodology of studying productivity in Indian Agriculture'. I.A.R.S., New Delhi: 348-370.

1971

**17. NARAIN, P. Average number of generations required to attain limits of genetic improvement. SABRAO Newsletter, 3: 135-142.

18. ROBERTSON, A. and P. NARAIN. Survival of recessive lethals in finite populations. Theoretical Population Biology, 2: 24-50.

**19. NARAIN, P. The use of transition probability matrices in studies on limits of response to selection. J. Indian Society of Agricultural Statistics, 23: 48-60.

20. NARAIN, P. The incompatibility load due to Rh factor in small populations. Proc. International Symposium on 'Human Genetics', Andhra University, Waltair, 1-11.

21. NARAIN, P. Experimental designs for selection studies in dairy cattle. Proc. Seminar on 'Dairy Cattle Breeding', N.D.R.I.(U.N.D.P.), Karnal: 1-7.

1972

22. REDDY, K.M. and P. NARAIN. The effect of incomplete penetrance of a gene on the genetical consequences of assortative mating. Indian J. Genetics, 32: 37-49.

**23. NARAIN, P. and K.M. REDDY. A note on the generalisation of homozygote x heterozygote matings. Indian J. Heredity, 4: 6-10.

24. NARAIN, P. Statistical problems in improving cattle population through breeding. In Silver Jubilee Souvenir, Indian Society of Agricultural Statistics, I.A.R.S., N. Delhi: 51-59.

25. NARAIN, P. Construction of selection index and its use in poultry breeding. Proc. Seminar on 'Poultry Breeding and Genetics', I.V.R.I.(U.N.D.P.), Izatnagar (U.P.), 45-48.

26. NARAIN, P. Methods of analysing diallel crosses in poultry. Proc. Seminar on 'Poultry Breeding and Genetics', I.V.R.I.(U.N.D.P.), Izatnagar (U.P.), 53-62.

27. NARAIN, P. and L.K. GARG. Estimation of genetic change in Indian herds of cattle. Indian J. Animal Production, 3: 143-153.

1973

28. NARAIN, P. Processing and evaluation of animal breeding experiments on computer. Proc. Symposium on 'Design for agricultural experiments and computerisation of their analysis'. J. Indian Society of Agricultural Statistics, 25: 139-141.

29. BHARGAVA, P.N., P. NARAIN, K.G. ANEJA and ASHA PRADHAN. A study of the occurrence of rainfall in Raipur district with the help of Markov Chain model. J. Indian Society of Agricultural Statistics, 25(2): 197-204.

30. NARAIN, P., K.L. MATI and D.K. BISWAS. Selection index for rate of lay in chickens. Indian Poultry Review, 4: 371-373.

31. NARAIN, P., K.L. MATI, D.K. BISWAS and L.B. SHRIVASTAVA. Response to selection on the basis of an index for rate of lay in chickens. Indian Poultry Review, 4: 769-771.

32. BHATIA, V.K. and P. NARAIN. Relationship between the retention of a sheep in the flock and its wool yield in the initial clip. Indian J. Animal Science, 43: 996-1002.

1974

33. NARAIN, P. Techniques for measuring genetic improvement in livestock breeding. Prod. 2nd General Congress, SABRAO, N. Delhi. Indian J. Genetics, 34A: 966-970.

34. NARAIN, P. Some aspects of statistical genetics in relation to animal improvement. Chapter 22 In Advances in Agriculture, 4: U.P. Institute of Agricultural Sciences, Kanpur: 213-221.

35. NARAIN, P. and J.P. JAIN. Some aspects of a sampling design for measurements of responses in Intensive Cattle Development Programme. In Symp. Vol. 'Statistical Assessment of I.C.D.P.', I.S.A.S., N. Delhi: 99-104.

36. MALHOTRA, P.K., P. NARAIN and K.L. MATI. Estimation of the genetic parameters for economic traits in chickens selected on the basis of an index for rate of lay. Indian Poultry Review, 5: 575-577.

37. NARAIN, P. Statistical techniques in animal breeding. Proc. 1st Conference of Agricultural Research Statisticians, I.A.R.S., N. Delhi: 98-108.

38. NARAIN, P. Teaching of Statistics in agricultural institutes and universities. Proc. 1st Conference of Agricultural Research Statisticians, I.A.R.S., N. Delhi: 116-124.

**39. NARAIN, P. The conditioned diffusion equation and its use in population genetics. J. Royal Statistical Society, London, B36: 258-266.

40. JAIN, J.P. and P. NARAIN. The use of population generation matrix in dairy herds. J. Indian Society of Agricultural Statistics, 26(2): 71-92.

41. BHARGAVA, P.N., P. NARAIN and ASHA PRADHAN. A study of the recurrence of rainfall deficiency in relation to rice crop. JNKVV Research J., 8: 206-209.

**42. NARAIN, P., C. SUBBA RAO and A.K. NIGAM. Partial diallel crosses based on extended triangular association scheme. Indian J. Genetics, 34: 309-317.

43. JAIN, J.P. and P. NARAIN. The accuracy in predicting the breeding value on the basis of the individual's own merit and that of its relatives in inbred populations. Indian J. Animal Science, 44: 939-946.

1975

44. NARAIN, P. and L.K. GARG. A possible use of discriminant function and D^2-statistic for comparing different grades of sheep in a cross-breeding programme. Indian J. Animal Science, 45: 243-247.

45. BABU, M. PRAKASH, D.K. BISWAS, S.C. MOHAPATRA and P. NARAIN. Effect of non-genetic sources of variation on some performance traits in White Leghorn. Indian J. Animal Science, 45: 762-766.

46. NARAIN, P., R. SINGH and A.K. MISHRA. The efficiency of indirect selection for life-time production in Red Sindhi Cattle. Indian J. Animal Production, 6: 11-18.

**47. NARAIN, P. and A.K. MISHRA. Efficiency of selective breeding based on a phenotypic index. J. Genetics, 62: 69-76.

48. NARAIN, P. On the theory of limits to artificial selection and average time taken to attain them. Proc. Symposium on 'Recent developments in statistical genetics in relation to plant and animal breeding', J. Indian Society of Agricultural Statistics, 27(1): 83-85.

49. GEORGE, K.C. and P. NARAIN. Parent-offspring correlations under full-sib and parent-offspring mating systems. J. Indian Society of Agricultural Statistics, 27(2): 51-70.

50. NARAIN, P. Theoretical aspects of mean time for gene substitution in human evolution. Proc. Symposium on 'Theoretical models used in Human Genetics'. 2nd Annual Conference, Indian Society of Human Genetics, Calcutta: 1-16.

51. NARAIN, P. Factor analysis in social research. Indian J. Extension Education, 11: 63-70.

52. ABRAHAM, T.P., P. NARAIN and L.K. GARG. Application of discriminant function technique for making fertilizer recommendations with the help of soil test values. Indian J. Agriculture Science, 45: 430-433.

1976

53. RAO, C. SUBBA and P. NARAIN. Effect of sampling the diallel cross on the estimation of some genetic parameters. Indian J. Genetics, 36: 278-288.

54. MATHUR, S.N. and P. NARAIN. Some optimal plans for partial diallel crosses. Indian J. Genetics, 36: 301-308.

**55. NARAIN, P. On Fisher's Fundamental Theorem of Natural Selection with non-overlapping generations. Current Science, 45: 800-801.

1977

56. NARAIN, P. Theoretical statistics and improvement in production of plants and animals. Sympo. Vol. 'Basic Sciences and Agriculture'. Indian National Science Academy, N. Delhi: 230-238.

57. NARAIN, P. Livestock data collection techniques in India. Proc. International Workshop on 'Inventory of Livestock Resources in Asia', Philippines: 1-15.

58. NARAIN, P. A plan for evolving a dairy breed utilising cross-bred cattle available in Military Farms. Military Farms J., 9: 41-47.

** 59. NARAIN, P. and E. POLIAK. On the fixation probability of a gene under random fluctuations in selection intensities in small populations. Genetical Research, 29: 113-121.

** 60. NARAIN, P. The Conditional Markov Chain in a genetic context. J. Genetics, 63: 49-62.

61 NARAIN, P. A population geneticist's view of caste. Proc. Seminar on 'Genetics and Social Organisation', Centre for Theoretical Studies, Indian Institute of Science, Bangalore.

62. BHARGAVA, P.N., P. NARAIN, D. SINGH and ASHA SAXENA. Statistical Studies on the behaviour of rainfall in a region in relation to a crop. Monograph I.A.R.S., N. Delhi: 1-67.

63. KUMAR, D. and P. NARAIN. Inbreeding in Sahiwal herd and its impact on economic characters. Indian Veterinary Medical J., 1: 27-34.

64. NARAIN, P., L.K. GARG and J.P. JAIN. The effect of including individual's egg weight in a selection index for rate of lay in chickens. Indian Poultry Review, 8: 17-23.

65. PURI, P.D., A.K. NIGAM and P. NARAIN. Supplemented Block Designs. Sankhya: The Indian J. Statistics, B39: 189-195.

66. ARYA, A.S. and P. NARAIN. Partial diallel crosses based on some association schemes with three and four associate classes. Sankhya: The Indian J. Statistics, B39: 394-399.

** 67. NARAIN, P. On the average age of a mutant in finite populations. Proc. 41st Session I.S.I., N. Delhi. Bulletin, International Statistical Institute, 47(2): 165-170.

68. NARAIN, P. Discussion on the Meeting: 'Empirical studies on inheritance of rare genetic traits'. Proc. 41st Session I.S.I., N. Delhi. Bulletin, International Statistical Institute, 47(1): 713-714.

1978

69. JAIN, J.P., K.P.S. NIRMAN, K.G. ANEJA and P. NARAIN. Impact of milk supply schemes on the rural economy in milk collection areas of Delhi Milk Scheme. Technical Research Bulletin, I.A.R.S., N. Delhi: 1-215.

70. KUMAR, D. and P. NARAIN. Lactation correction factors and repeatability of economic characters in Sahiwal cows and Murrah buffaloes. Indian J. Dairy Science, 31: 220-225.

71. NARAIN, P. Statistical basis of designing experiments for residue analysis with special reference to sampling techniques. Proc. Workshop on 'Pesticides Residues with special reference to sampling techniques', I.A.R.I., N. Delhi: 22-24.

72. NARAIN, P. Strategy for livestock breeding. Proc. 3rd Conference of Agricultural Research Statisticians, I.A.S.R.I., N. Delhi: 67-68.

73. SINGH, H.P., B.C. SAXENA, P. NARAIN and S.P. VERMA. Impact of milk supply scheme on rural economy in milk collection areas of Madhavaram Milk Supply Scheme (Bench-Mark Survey). Technical Research Bulletin, I.A.S.R.I., N. Delhi: 1-160.

** 74. NARAIN, P. Optimum designs for progeny testing with minimum costs. Indian J. Animal Science, 48: 708-711.

** 75. NARAIN, P. Average time until fixation of a mutant at a tri-allelic locus in a finite population. Indian J. Genetics, 38: 57-62.

76. NARAIN, P. Statistical techniques for studying constraints - Factor analysis. Proc. Workshop-cum-Training Course in Methodology of Constraint Analysis, I.A.S.R.I., N. Delhi: 45-62.

77. KUMAR, D. and P. NARAIN. Effect of season of calving on lactation yield and length in Red Sindhi, Jersey and their crosses. Cherion - Tamil Nadu J. Veterinary Science and Animal Husbandry, 7: 190-194.

78. JAIN, J.P., B.C. SAXENA, K.G. ANEJA and P. NARAIN. Growth of milk producers' cooperatives in Mehsana. Indian Dairyman, 30: 549-552.

1979

** 79. NARAIN, P. On the statistical properties of the conditional equilibrium distribution under steady flux of mutations. Proc. Indian National Science Academy, B45: 239-246.

80. NARAIN, P., V.K. BHATIA and P.K. MALHOTRA. A Handbook of Statistical Genetics. I.A.S.R.I., N. Delhi: 1-365.

81. JAIN, J.P. and P. NARAIN. A further study on the use of population generation matrix in dairy herds. J. Indian Society of Agricultural Statistics, 31: 63-79.

82. NARAIN, P., J.P. JAIN, L.K. GARG, R.K. JAIN and P.S. RANA. Herd strengths at periodic intervals and annual genetic gain in alternative progeny testing programme. Military Farms J., 11: 9-16.

83. KUMAR, D. and P. NARAIN. Estimation of genetic changes in Sahiwal herd. Indian Veterinary J., 56: 844-848.

84. NARAIN, P. and L.K. GARG. Milk production and calving interval of cross-bred cows. Indian J. Dairy Science, 32: 193-195.

85. SAKSENA, ASHA, P.N. BHARGAVA and P. NARAIN. Rainfall pattern and crop planning. Indian J. Agricultural Research, 13: 208-214.

86. NARAIN, P. and J.C. MALHOTRA. Data analysis in animal breeding. Proc. 4th Conference of Agricultural Research Statisticians, Palampur (H.P.): 58-67.

87. JAIN, J.P., B.C. SAXENA, K.G. ANEJA and P. NARAIN. Impact of milk supply schemes on the rural economy in milk collection areas of Dudhsagar Dairy, Mehsana (Gujarat). Technical Research Bulletin, I.A.S.R.I., N. Delhi: 1-210.

88. SINGH, H.P., B.C. SAXENA, P. NARAIN, S.R.S. ARYA and J.P. JAIN. Impact of milk supply schemes on the rural economy in milk collection areas of Greater Calcutta Milk Supply Scheme (Bench-Mark Survey). Technical Research Bulletin, I.A.S.R.I., N. Delhi: 1-180.

89. NARAIN, P., J.P. JAIN, H.P. SINGH and B.C. SAXENA. The status of Rural milk producers in milk collection areas of four major dairy plants. Indian Dairyman, 31: 161-174.

90. NARAIN, P. Progeny testing in goats. Proc. Summer Institute on 'Goat Production', Central Institute for Research on Goats, Makhdoom (Mathura): 95-98.

91. NARAIN, P., R.K. KHOSLA, D.S. ANEJA and R.S. KHATRI. Ed. ICAR Golden Jubilee (1929-79). IASRI Souvenir Vol., I.A.S.R.I., N. Delhi: I-180.

92. NARAIN, P., P.N. BHARGAVA and ASHA SAXENA. The use of a Markov chain model for crop planning in rainfed areas. ICAR Golden Jubilee (1929-79), IASRI Souvenir Vol.: 75-84.

93. NARAIN, P. The use of partial diallel crosses in plant improvement. ICAR Golden Jubilee (1929-79), IASRI Souvenir Vol.: 183-193.

94. NARAIN, P. and J.P. JAIN. A survey of the development of research in Statistical Genetics at the IASRI. ICAR Golden Jubilee (1929-79), IASRI Souvenir Vol.: 266-291.

95. NARAIN, P. and K.V. SATHE. Analysis of animal breeding data with multiple characters on computer. ICAR Golden Jubilee (1929-79), IASRI Souvenir Vol.: 357-367.

**96. NARAIN, P. A new sire index for milk production corrected for an auxiliary trait. Indian J. Animal Genetics and Breeding, 1: 20-22.

**97. NARAIN, P. Certain generalisations in response to genetic selection. In 'Contributions to Statistics', S.C. Das Memorial Vol.: Utkal University, Bhubaneswar: 71-79.

98. NARAIN, P. The state of Agricultural Statistics. Commerce, Annual Number on Agriculture, 139: 131-135.

99. NARAIN, P. and V.K. BHATIA. Some aspects of relationship between yield in first lactation and survival to culling in successive lactations. Technical Report, Training and Research, I.A.S.R.I., N. Delhi, 4/79: 1-23.

100. NARAIN, P. Strategy of cross-breeding in livestock and poultry. Symposium on 'Assessment of cross-breeding programmes', J. Indian Society of Agricultural Statistics, 31(2): 94-97.

101. NARAIN, P., L.K. GARG, J.P. JAIN, J.C. PURI, P. LAL and P.S. RANA. Statistical methodology for developing efficient selection procedures in poultry breeding. Technical Report, I.A.S.R.I., N. Delhi: 1-115.

1980

102. MALHOTRA, P.K. and P. NARAIN. A study on the distribution of breeding values for rate of lay in poultry. Indian J. Poultry Science, 15: 37-42.

103. NARAIN, P. and LAL CHAND. Lactation performance index for Sahiwal and Hariana Cattle. Indian J. Animal Genetics and Breeding, 2: 16-20.

104. KUMAR, D. and P. NARAIN. Different methods of sire evaluation. Indian J. Dairy Science, 33: 468-472.

105. NARAIN, P. Livestock Breeding. Chapter 5 in 'Methodology for Improvement of Data Base on Livestock Resources', APHCA - IASRI, N. Delhi: 93-116.

106. NARAIN, P. Statistical problems in evolving new breeds of dairy cattle. Proc. Summer Institute on 'Advanced Statistical Methodology as applied to Animal Sciences', I.A.S.R.I., N. Delhi: 193-204.

107. NARAIN, P. Cross-breeding in buffaloes - A breeding plan. Indian J. Animal Genetics and Breeding, 2: 47-50.

108. BATRA, M.S. and P. NARAIN. Discriminatory analysis in Wheat and Triticale with the help of diallel cross data. Annals Agricultural Research, 1: 182-188.

1981

109. SINGH, H.P., P. NARAIN and N.D. KHANNA. Association between blood group antigenic factors and quantitative traits in some Indian herds of cattle and buffaloes. Indian J. Animal Science, 51: 745-751.

**110. NARAIN, P. and A.S. ARYA. Truncated Triangular Association Scheme and Related Partial Diallel Crosses. Sankhya, The Indian J. Statistics, B43: 93-103.

111. NEI, M., W.H. LI, F. TAJIMA and P. NARAIN. Polymorphism and Evolution of the Rh Blood Groups. Japanese J. Human Genetics, 26: 263-278.

112. JAIN, J.P., L.K. GARG, R.K. JAIN and P. NARAIN. Optimum progeny testing plans in dairy herds of various sizes. Technical Report, I.A.S.R.I., N. Delhi: 1-217.

113. NIGAM, A.K., V.K. GUPTA and P. NARAIN. Supplemented Block Designs. Monograph, I.A.S.R.I., N. Delhi: 1-26.

114. NARAIN, P. and R.K. KHOSLA. Statistics in Pest Management. Proc. National Seminar on 'Strategies of Pest Management', I.A.R.I., N. Delhi: 1-14.

1982

115. JAIN, J.P., P. NARAIN and J.C. MALHOTRA. Some aspects of progeny testing of dairy bulls under field conditions. Indian J. Animal Science, 52: 628-633.

116. SINGH, S. and P. NARAIN. Disequilibrium due to linkage in self-fertilized populations. Indian J. Genetics, 42: 129-133.

117. NARAIN, P. Genetic Components in Man and Environment. In Newer Concepts in Nutrition and their implications for Policy - Ed. P.V. Sukhatme, M.A.C.S. Research Institute, Pune: 117125.

118. NARAIN, P. and R.K. KHOSLA. Statistical Methodology for the estimation of post-harvest foodgrain losses. Proc. Workshop on 'Post-harvest losses and Small Farmer Storage', N. Delhi. Food Production and Rural Development Division, Commonwealth Secretariat, London, Vol.II: 131-137.

119. NARAIN, P., R.K. PANDEY and SHANTI SARUP. Perspective Plan for Foodgrains. Commerce, 145: 184-191.

120. BHARGAVA, P.N., P. NARAIN and J.K. KAPOOR. Effect of Weedicides on rice and wheat. Pesticides, 15: 27-31.

121. NARAIN, P. Genetic interpretation of the auto-regulatory mechanism in nitrogen balance. Proc. Golden Jubilee Conference on 'Human Genetics and Adaptation', I.S.I., Calcutta: 1-10.

122. NARAIN, P. and R.K. KHOSLA. Post Harvest Estimation of Losses of Agricultural Products due to biotic and abiotic factors. Proc. Golden Jubilee Conference on 'Frontiers of Research in Agriculture', I.S.I., Calcutta: 411-427.

123. SUKHATME, P.V. and P. NARAIN. A possible genetic interpretation of auto-regulatory mechanism in models for protein deficiency. Proc. Indian National Science Academy, B48: 748-754.

1983

124. NARAIN, P. Stochastic Problems and Methods in Population Genetics. Presidential Address, Section of Statistics, 70th Indian Science Congress, Proc. Pt.II: 1-32.

125. NARAIN, P. and J.C. MALHOTRA. Cost-benefit evaluation of progeny testing programme under field conditions. Proc. National Symposium on 'Breeding livestock for economic efficiency', Haryana Agricultural University, Hissar: 1-13.

126. NARAIN, P., R.K. KHOSLA and R.K. PANDEY. Agricultural Statistics. Seminar on 'Agricultural Development in Madhya Pradesh', Pachmarhi (M.P.). Technical Paper No.5(DC), I.A.S.R.I., N. Delhi: 1-13.

127. NARAIN, P. Statistical aspects of genotype x environment interactions in plant breeding. Valedictory Address, Summer Institute on 'Advances in breeding methodologies for crop improvement', Haryana Agricultural University, Hissar: 1-11.

128. NARAIN, P. Testing of neutrality of genes in sub-divided Populations. Proc. Summer Institute on 'Immunogenetics and Biochemical Polymorphism in Domestic Animals'. I.V.R.I., Izatnagar (U.P.), 1-18.

129. NARAIN, P. Some aspects of research in dry farming areas. Presidential Address - Section of Agriculture Science, 8th Indian Social Science Congress, Hyderabad: 1-17.

130. NARAIN, P. Some statistical aspects of research in dry farming areas. Technical Address - 7th Conference of Agricultural Research Statisticians, University of Agricultural Sciences, Bangalore: 1-10.

131. NARAIN, P. and O.P. KATHURIA. The Status of Fisheries Statistics in India. Agricultural Situation in India, 38: 259-266.

**132. NARAIN, P. Statistical aspects in optimising livestock breeding programmes. Proc. 44th Session, International Statistical Institute, Madrid(Spain), Vol.2: 833-836.

133. NARAIN, P. Status of animal experimentation in India and remedial measures for their improvement. Proc. Symposium on 'Design and Selection Procedures for Animal Experiments'. 37th Conference, Indian Society of Agricultural Statistics, Shimla: 1-11.

134. NARAIN, P., P.N. BHARGAVA and Mrs. A.K. BHATIA. Response to application of potash for wheat and rice. Fertilizer Marketing News, 14: 7-12.

135. NARAIN, P., R.K. PANDEY and SHANTI SARUP. Economic study of crop productivity in Indian Agriculture. Proc. 66th Annual Conference, Indian Economic Association, Bangalore: 88-97.

136. SUKHATME, P.V. and P. NARAIN. Intra-individual variation in energy requirement and its implications. Indian J. Medical Research, 78: 857-865.

137. NARAIN, P., P.N. BHARGAVA and H.C. JAIN. Profitability in Systems Approach. Fertilizer News, 28: 64-73.

**138. NARAIN, P., P.K. MALHOTRA and S.D. WAHI. The use of auxiliary traits in combined selection for poultry improvement. Indian J. Poultry Science, 18: 37-47.

139. CHAND, LAL and P. NARAIN. Genetic parameters of lactation performance index in Sahiwal Cattle. Indian J. Animal Science, 54: 238-239.

140. ARYA, A.S. and P. NARAIN. A computer programme for the analysis of partial diallel crosses based on general circulant designs. Indian J. Genetics, 43.

1984

141. NARAIN, P., P.N. BHARGAVA and ASHA SAXENA. A statistical study on incidence of drought in relation to agricultural production, Mausam, 35: 395-400.

142. NARAIN, P., J.P. JAIN and R.K. JAIN. A note on the increase in the coefficient of inbreeding in a progeny testing programme. Indian J. Animal Science, 54.

143. NARAIN, P. and R.K. KHOSLA. Estimation of foodgrain losses at different post-harvest stages. J. Indian Society of Agricultural Statistics, 36(1).

144. SUKHATME, P.V. and P. NARAIN. The genetic significance of intra-individual variation in energy requirement. In 'W.G. Cochran's Impact on Statistics', Ed. P.S.R.S. Rao and J. Sedransk, John Wiley and Sons, N. York.

145. NARAIN, P. On contributions of P.V. Sukhatme in the field of nutrition. In P.V. Sukhatme Felicitation Vol. Ed. P. Narain, I.S.A.S., N. Delhi.

146. NARAIN, P. Scope of Statistics in Anthropology. In 'Anthropology in Indian Context'. Ed. I.J.S. Bansal, Department of Human Biology, Punjabi University, Patiala.

S E C T I O N O N E

STUDIES
IN
STATISTICAL GENETICS

- PUBLISHED WORK

CONTENTS

	GENERAL	28
I	GENETIC PROPERTIES OF POPULATION	30
II	STOCHASTIC PROCESSES IN POPULATION GENETICS	45
III	GENETICS OF QUANTITATIVE VARIABILITY	103
	REFERENCES	154

It is now well recognised that the low productivity of plants and animals can be improved by conducting research in breeding. However, the foundations of the modern theory of breeding are based on the sciences of genetics and statistics. For instance, it is no use improving a character by breeding if the genetic considerations dictate that it has negligible heritable variation. Similarly, in the context of animal breeding, if statistical considerations demand that twenty progeny are needed for the progeny test of a sire, it is not desirable for a breeder to be content with two or three progeny per sire. The principles of genetics and statistics together constitute the scientific discipline which is often called '<u>Statistical Genetics</u>', the foundations of which were laid by Fisher, Haldane and Wright. To this list may also be added the names of Crow, Kimura, Robertson, Kempthorne, Nei and several others who have greatly advanced our knowledge in this field. This discipline has had a considerable impact on the practice of plant and animal breeding since the turn of this century. It has also helped in understanding the mechanics of organic evolution and in recent times has opened new vistas in the theory of molecular evolution.

The statistico-genetic approach to plant and animal improvement is not simply based on Mendelian principles although the laws of Mendel are fundamental in this approach. The manner in which a quantitative character as opposed to a qualitative one,

is controlled by heredity, is at the root of this approach. Further, while the approach of Mendel is on individual basis, the population or a random sample drawn from it, is the basis for any breeding programme. In order, therefore, to understand the techniques of breeding, it is very essential that the principles of genetics of population are properly understood. Such principles are also crucial for understanding the mechanics of evolution, particularly at the molecular level. The theoretical description and analysis of population genetic models as well as their use in breeding and evolution, however, often require advanced mathematical and statistical techniques. Some of the basic investigations conducted by me in these areas during 1965 to 1983 and documented in 23 publications are described in this Section. The section is divided broadly into three groups; genetic properties of population, stochastic processes in population genetics, and genetics of quantitative variability consisting of 4, 10 and 9 research papers respectively.

I. GENETIC PROPERTIES OF POPULATION

Quantitative measures of the intensity of inbreeding and degree of relationship under various systems of mating were first given by Wright (1921) with the aid of path coefficients. The work of Malécot (1948) resulted in essentially the same formulae as that of Wright but his approach was to make use of the probabilities of genes being identical by descent at a locus, the coefficient of inbreeding (F) of an individual being defined as the probability that the two genes possessed by that individual at a locus are identical by descent. Schnell (1961) generalised this approach for an arbitrary number of linked loci and defined an in-breeding function (ϕ) as the probability that the individual possesses genes which are identical by descent for a given set of linked loci. Narain (1 and 2) as described below used this method to study the effect of linkage on the homozygosity of a selfed population as well as of a population under mixed selfing and random mating for an arbitrary number of linked loci.

In a genetic incompatibility model, only certain specific types of matings out of all the possible types, produce viable offspring. However, there could be situations in which the only possible type of matings is between homozygotes and heterozygotes. Finney (1952) introduced such incompatibility models with respect to a single locus and two alleles whereas Scudo(1964)

described such models as a basis for polygenic sex-determination. In the paper discussed herein Narain and Reddy (3) generalised the models to situations in which several loci are segregating independently so that there are now heterozygotes at each of the loci as well as double heterozygotes.

The Fundamental Theorem of Natural Selection first given by Fisher (1930) broadly appears in two forms for the case of non-overlapping generations. According to one form, the change in the average fitness of a population is equal to the genotypic variance in fitness. The other form, which includes the effect of a mating system, states that for random mating population, with two-allele system, the rate of increase in average fitness at any time is equal to its additive genetic variance at that time. Thus, in the absence of dominance in fitness values, the two forms are identical. However, the interpretation of dominance in a two-allele system is essentially that of an interaction between the two alleles, the three fitness values attached to the three genotypes forming an arithmetic series in the absence of dominance. However, if these fitness values form a geometric series, there would be no dominance on the logarithmic scale but on the given scale some partial dominance would be exhibited resulting in two different forms of the theorem. As would be described shortly, Narain (4) showed that even in such cases of dominance due to scale effects the two

forms of the theorem could be identical.

1. Narain, P. (1966). Homozygosity in a selfed population with an arbitrary number of linked loci. Journal of Genetics, 59: 254-266.

Consider two individuals X and Y having genotypes $\frac{a_1 a_2 \ldots a_r}{b_1 b_2 \ldots b_r}$ and $\frac{c_1 c_2 \ldots c_r}{d_1 d_2 \ldots d_r}$ respectively where r is the number of loci and the horizontal line indicates that the genes above it lie on one chromosome and those below it lie on the other homologous chromosome. For a given locus say i-th, the coefficient of relationship between X and Y, denoted by ρ_i^{XY} is defined in terms of probabilities $P(a_i=c_i)$ that a random gene a_i from X is identical by descent with a random gene c_i from Y at the i-th locus. For a pair of loci i and j with a recombination value p_{ij}, we define the coefficient of relationship (ρ_{ij}^{XY}) in terms of probabilities $P(a_i=c_i ; a_j=c_j)$ that a random gene a_i from X is identical by descent with a random gene c_i from Y at the i-th locus as well as a random gene a_j from X is identical by descent with a random gene c_j from Y at the j-th locus. This definition is generalised to a set of r linked loci and denoted by $\rho_{123\ldots r}^{XY}$.

For developing recurrence relations, use is made of the result that, for a given set of loci, the inbreeding function

$\phi^{X\times Y}$ of an offspring from the mating of two individuals X and Y is given by

$$(1.1) \qquad \phi^{X\times Y}_{123\ldots r} = P^{XY}_{123\ldots r}$$

with

$$(1.2) \qquad \phi^{X}_{123\ldots r} = P\left(\begin{array}{l} a_1=b_1 \\ a_2=b_2 \\ a_3=b_3 \\ \vdots \\ a_r=b_r \end{array}\right)$$

as the inbreeding function of X for the set of r loci and similarly we have the inbreeding function of Y.

In the case of the system of mating involving only self-fertilisation, an individual X is mated with itself so that in the above formula, we have to replace Y by X. For the case of two loci i and j, we then get the recurrence relation connecting ϕ-functions between n-th and (n+1)-th generation,

$$(1.3) \qquad \phi^{(n+1)}_{ij} = \tfrac{1}{4}(1+C^2_{ij})(1+\phi^{(n)}_{ij}) + \tfrac{1}{4}(1-C^2_{ij})(\phi^{(n)}_i + \phi^{(n)}_j)$$

where $C_{ij} = 1-2p_{ij}$ and $\phi^{(n)}_i$ is the n-th generation inbreeding coefficient for the i-th locus.

The recurrence relations for the panmictic function π_{ij}, however, assume a more simpler form given by

$$(1.4) \qquad \pi_{ij}^{(n+1)} = (\frac{k_{ij}}{2})\pi_{ij}^{(n)}$$

where $k_{ij} = (1+C_{ij}^{2})/2$.

For a set of r linked loci, provided there is no interference, the recurrence relation for $\pi_{123\ldots r}$ gets generalised to

$$(1.5) \qquad \pi_{123\ldots r}^{(n+1)} = (\frac{k_{12}k_{23}\ldots k_{(r-1)r}}{2})\pi_{123\ldots r}^{(n)}$$

If initially we start with an individual heterozygous at each of the r loci and assume no interference, we get

$$(1.6) \qquad \phi_{123\ldots r}^{(n)} = 1-r(\tfrac{1}{2})^{n} + \Sigma_1(\frac{k_{ij}}{2})^{n} - \Sigma_2(\frac{k_{ij}k_{jl}}{2})^{n} + \Sigma_3(\frac{k_{ij}k_{jl}k_{lm}}{2})^{n}$$

$$\ldots\ldots + (-)^{r}(\frac{k_{12}k_{23}\ldots k_{(r-1)r}}{2})^{n},$$

where Σ_1 is summation over $\binom{r}{2}$ values of $k_{12}, k_{23}, \ldots k_{13}, k_{14}, \ldots k_{(r-1)r}$. Similarly Σ_2 is summation over $\binom{r}{3}$ pairs of k values depending upon the three loci selected out of r. Similar considerations hold for other summations.

The mean (m) and variance (v) of the distribution of the number of loci homozygous by descent in any population turn out to be, for r=2,3,...,

$$(1.7) \qquad m = r\phi_i^{(n)}$$

$$(1.8) \qquad v = r\phi_i^{(n)} (1-r\phi_i^{(n)}) + 2\Sigma\phi_{ij}^{(n)}$$

where Σ denotes summation over $\binom{r}{2}$ values of loci which are distinct.

The proportion of residual lines which become completely homozygous by descent in each generation depends on the ϕ-function by the formula

$$(1.9) \qquad \Delta^{(n)} = (\phi^{(n)} - \phi^{(n-1)})/(1-\phi^{(n-1)})$$

thus giving,

$$(1.10) \qquad \Delta_{123\ldots r}^{(n)} \simeq (\tfrac{1}{2})[1-\tfrac{1}{r}\Sigma_1 (k_{ij}^{n-1} - k_{ij}^{n})]$$

The properties of populations with respect to three loci during five generations of self-fertilization for $p_{12} = 0.3$ and $p_{23} = 0.5$ with no interference, giving $p_{13} = 0.8$ are given in Table 1.1.

TABLE 1.1 Values of inbreeding and panmictic functions under five generations of self-fertilization for three loci, the mean and variance of the number of loci homozygous by descent and the proportion of the residual lines that become completely homozygous by descent in each generation.

	GENERATIONS					
	0	1	2	3	4	5
π_{123}	1	0.1972	0.0389	0.0077	0.0015	0.0003
ϕ_{123}	0	0.1928	0.4784	0.6986	0.8361	0.9139
m	0	1.5000	2.2500	2.6250	2.8125	2.9064
v	0	1.0300	0.7221	0.3970	0.2025	0.1004
Δ_{123}	-	0.1928	0.3538	0.4222	0.4562	0.4747

It is apparent that the effect of linkage is to retard the rate of approach to homozygosity. While with one locus 50 per cent of the residual lines become completely homozygous by descent in each generation starting from the first, with two or three loci, this rate is much less in the first generation but increases with further generations of selfing depending upon the recombination parameters. For the three loci case with results given in Table 1.1, it becomes about 47 per cent by fifth

generations. Asymptotically, however, these rates approach the 50 per cent limit.

Citation

1. On pp. 240, 242 and 245 in
Shikata, M. (1967). Journal of Theoretical Biology, 17: 240-245.

2. On pp. 629, 639 and 640 in
Cockerham, C.C. and Weir, B.S. (1968). Genetics, 60: 629-640.

3. On pp. 933 and 940 in
Weir, B.S. and Cockerham, C.C. (1969). Genetics, 61: 923-940.

4. On pp. 290 and 291 in
Shikata, M. (1969). Journal of Theoretical Biology, 24: 279-293.

5. On pp. 242, 243 and 498 in
Wright, S. (1969). Evolution and the Genetics of Populations. Vol.2. The Theory of Gene Frequencies. University of Chicago Press, Chicago.

6. On p. 557 in
Crow, J.F. and Kimura, M. (1970). An Introduction to Population Genetics Theory. Harper and Row, New York.

7. On pp. 323 and 330 in
Cockerham, C.C. and Weir, B.S. (1973). Theoretical Population Biology, 4: 300-330.

2. Narain, P. (1966). Effect of linkage on homozygosity of a population under mixed selfing and random mating. Genetics, 54: 303-314.

Consider two loci case with i-th locus A-a and j-th locus B-b. Let there be constant probability s of selfing and (1-s) of mating at random in an initially random mating population in equilibrium. Then an individual in the n-th generation can possess the genes which are not identical by descent at both the loci if either it is an offspring resulting from the randomly mating individuals or it is an offspring of such a selfed individual in the (n-1)-th generation which possesses genes which are also not identical by descent at both the loci. In other words, the recurrence relation for π_{ij} under mixed selfing and random mating would be given by

$$(2.1) \qquad \pi_{ij}^{(n)} = s\left(\frac{k_{ij}}{2}\right)\pi_{ij}^{(n-1)} + (1-s)$$

This gives the following solution

$$(2.2) \qquad \pi_{ij}^{(n)} = \frac{2(1-s)}{(2-sk_{ij})}\left\{1-\left(\frac{sk_{ij}}{2}\right)^{n}\right\} + \left(\frac{sk_{ij}}{2}\right)^{n}\pi_{ij}^{(o)}$$

The function of inbreeding $\phi_{ij}^{(n)}$ is then calculated from the relation

$$(2.3) \qquad \phi_{ij}^{(n)} = 1-\pi_{i}^{(n)} - \pi_{j}^{(n)} + \pi_{ij}^{(n)}$$

where $\pi_i^{(n)}$ or $\pi_j^{(n)}$ is, by similar arguments, given by

$$(2.4) \qquad \pi_i^{(n)} = \frac{2(1-s)}{(2-s)}\{1 - (\tfrac{s}{2})^n\} + (\tfrac{s}{2})^n \pi_i^{(o)}$$

As n tends to infinity, we get the limiting values of π_{ij} and ϕ_{ij} as

$$(2.5) \qquad \lim_{n\to\infty} \pi_{ij}^{(n)} = 2(1-s)/(2-sk_{ij})$$

$$(2.6) \qquad \lim_{n\to\infty} \phi_{ij}^{(n)} = 1 - \frac{4(1-s)}{(2-s)} + \frac{2(1-s)}{(2-sk_{ij})}$$

When the population is initially random mating i.e. $\pi_{ij}^{(o)} = 1$, the rate at which residual lines become completely homozygous by descent is given by

$$(2.7) \qquad \Delta_{ij}^{(n)} = [\phi_{ij}^{(n)} - \phi_{ij}^{(n-1)}] \Big/ \{\lim_{n\to\infty} \phi_{ij}^{(n)}\} - \phi_{ij}^{(n-1)}$$

$$= 1 - (\tfrac{s}{2}) d_{ij}^{(n)} / d_{ij}^{(n-1)}$$

where $d_{ij}^{(n)} = (\frac{4}{2-s}) - 2(\frac{2-k_{ij}}{2-sk_{ij}})k_{ij}^{(n-1)}$

The above results for two loci were extended to three loci and given, in general, for an arbitrary number of loci, assuming

no interference. The effect of linkage was studied numerically for the two loci situation, involving only one recombination parameter. The effect was found to increase with increase in s for the first two generations but thereafter it increases upto a certain value of s and decreases subsequently. The interaction between the phenomena of linkage and selfing was clearly demonstrated but in so far as rate of inbreeding is concerned, the presence or absence of linkage, practically makes little difference after four to five generations of mixed selfing and random mating.

Citation

1. On p. 498 in

Wright, S. (1969). Evolution and the Genetics of Population. Vol.2. The Theory of Gene Frequencies. University of Chicago Press, Chicago.

2. On p. 557 in

Crow, J.F. and Kimura, M. (1970). An Introduction to Population Genetics Theory. Harper and Row, New York.

3. On pp. 323 and 330 in

Cockerham, C.C. and Weir, B.S. (1973). Theoretical Population Biology, 4: 300-330.

4. On pp. 248, 251 and 262 in

Weir, B.S. and Cockerham, C.C. (1973). Genetical Research, 21: 247-262.

3. Narain, P. and K.M. Reddy (1972). A note on the generalisation of homozygote x heterozygote matings. Indian Journal of Heredity, 4(1): 6-10.

With two loci segregating independently in a population, there are 9 genotypes AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb and aabb. Suppose these 9 genotypes are divided into two groups, one group consisting of the four homozygotes and the double heterozygote and the other consisting of the four single heterozygotes. The first group is divided into two sets, one consisting of the four homozygotes and the other consisting of the only double heterozygote. The other group is also divided into two sets, one consisting of the two genotypes heterozygous at the first locus and the other consisting of the other two genotypes heterozygous at the other locus. The matings are now allowed only between the two sets within each of the two groups. This ensures a mechanism for matings only between homozygotes and heterozygotes at each of the two loci separately as well as simultaneously. The eight types of matings produced from the arrangement would give rise to the 9 genotypes in the offspring generation with frequencies which can be related to their frequencies in the previous generation. These recurrence relations are solved to determine the equilibrium of the population.

It is found that in so far as heterozygotes - single or double - are concerned, the population attains equilibrium just after one generation but so far as the homozygotes are concerned, it takes an infinitely large number of generations for the attainment of equilibrium. These results have been generalised to an arbitrary number of independently segregating loci. With the help of Computer, it was shown that as the number of loci increases, the number of generations required to attain the equilibrium also increases. The equilibrium genotypic frequencies are found to depend on certain ratios between the initial genotypic frequencies and can be easily found by multiplying the equilibrium genotypic frequencies expected at each of the loci separately.

4. Narain, P. (1976). On Fisher's fundamental theorem of natural selection with non-overlapping generations. Current Science, 45(22): 800-801.

Let the relative fitness of the three genotypes AA, Aa and aa be respectively w_2, w_1 and w_o in a random mating population with gene frequencies p for A and q for a with p+q=1. The average fitness of such a population can be expressed as

(4.1) $$W = pw_2 + qw_o - pq(w_2 - 2w_1 + w_o)$$

where $(w_2-2w_1+w_o)$ expresses the degree of dominance on the arithmetic scale. After the operation of natural selection, the increase in average fitness, ΔW, is

$$(4.2) \qquad \Delta W = \sigma_w^2/W$$

where σ_w^2 is genotypic variance in fitness values and is the sum of additive genetic (σ_A^2) and dominance (σ_D^2) variances given by

$$(4.3) \qquad \sigma_A^2 = 2pq[W(w_2-2w_1+w_o) + (w_1^2 - w_2w_o)]$$

$$(4.4) \qquad \sigma_D^2 = p^2q^2[w_2-2w_1+ w_o]^2$$

It is then possible to show that ΔW can, alternatively, be expressed as

$$(4.5) \qquad \Delta W = \frac{\sigma_A^2}{W}\left[1 + \frac{pq(w_2-2w_1+w_o)}{2(W+\alpha)}\right]$$

where $\alpha = (w_1^2-w_2w_o)/(w_2-2w_1+w_o)$

Taking into account the round of random mating in addition to the effect of natural selection, the change in the average fitness of the population, denoted by ΔW^*, is

$$(4.6) \qquad \Delta W^{*} = \frac{\sigma_A^2}{W}\left[1 + \frac{pq(w_2 - 2w_1 + w_o)}{2W}\right]$$

Comparing the expressions for ΔW and ΔW^{*}, we find that even if $(w_2 - 2w_1 + w_o)$ is not zero, the two would be identical provided $\alpha = 0$ or in other words $w_1^2 = w_2 w_o$ indicating no dominance on the geometric scale.

II. STOCHASTIC PROCESSES IN POPULATION GENETICS

Population genetics concerns with the genetical make up of living populations over time maintained under either natural or artificial conditions. While in the former case, we are concerned with the process of organic evolution in wild species, in the latter, it is the process of affecting genetic improvement in the economic traits of domesticated crop plants and livestock. In either case, the basic variable of study is gene frequency or proportion of a given gene in a population. A fundamental problem in population genetics is then to describe the changes in the frequency of a gene over time due to systematic forces like selection, mutation and migration. When the size of the given population is very large and the individuals of the population mate at random, the change in the gene frequency is deterministic and can be easily studied by simple algebraic principles as shown by Fisher (1922), Haldane (1924) and Wright (1931). In actual practice, however, the populations are small so that the gene frequency is also subject to fluctuations over time due to random forces created either by the random sampling of gametes in reproduction or by the random fluctuations in systematic forces or else by both. The change in gene frequency over time is then a <u>stochastic process</u> and we have to employ mathematical and statistical methods to study such processes.

When the gene frequency undergoes a random change from generation to generation, a certain form of distribution of gene frequency is realised. According to Wright (1931), such a distribution can be regarded either as the distribution of frequencies at equivalent loci in one population or as the distribution of frequencies at a single locus replicated in many equivalent populations. As time proceeds, this distribution gets broadened with reversible fixation (or loss) of genes leading to a state of steady decay when the distribution curve attains a constant form. The height of the curve then decreases at a constant rate and becomes zero in the limit unless there are systematic forces which could stabilise the distribution of gene frequencies. The expected frequency of the given allele in the limit is known as the probability of its fixation (Kimura, 1957). This probability can be regarded either as the proportion of equivalent loci which would be expected to be fixed in the limit in any line or as the proportion of replicate selected lines in which an individual gene could be expected to be fixed in the limit. Further, the fixation or loss of genes occur after a variable number of generations. Hence the distribution of time to fixation of a given allele disregarding the cases in which it is lost, can better describe the life of a gene until it is fixed as would be discussed in Narain and Robertson (5).

The stochastic process of the change in gene frequency involves Markov property in the sense that changes in the gene frequency in a given generation depend solely on their frequencies in the immediately preceding generation and are totally independent of the past history of the population. Such a Markov process can be studied through two approaches. In one approach, known as diffusion approach, the Markov process is approximated as continuous in gene frequency as well as in time parameter. Using this approach, Kimura (1962) obtained a formula for the probability of fixation of a mutant gene in a population as well as Kimura and Ohta (1969) derived an expression for the average number of generations until fixation of a neutral mutant gene. As discussed herein Narain (6), following this approach obtained an expression for the variance and the coefficient of variation of the number of generations until fixation of a neutral mutant gene in a finite population.

The second approach to study the change in gene frequency is to describe the process exactly by a finite Markov Chain with discrete time parameter and gene frequency as a discrete random variable changing by steps between zero and one, depending on the population size. This is known as transition matrix approach and it involves the evaluation of transition matrices for moderate population size on a computer. In the context of the limits of artificial selection, this approach was used by

Narain and Robertson (5) and Narain (7 and 8) as would be discussed shortly.

The underlying stochastic process of the change in gene frequency can result in either fixation or loss of an allele at a locus within finite length of time provided there are no balancing forces to prevent such fixation or loss to occur. In the language of stochastic process, the fixation or loss of an allele corresponds to absorption in one or other of the two possible absorbing states. All the sample paths, as realisation of the given stochastic process can, therefore, be divided into two categories. One category of sample paths would all end up with absorption in one of the boundaries whereas the other category of sample paths would all absorb in the other boundary. It is then more appropriate to consider only such sample paths that lead to absorption in one of the two boundaries disregarding those in which the absorption occurs in the other boundary. This is made possible by invoking a conditional stochastic process with the help of the probability of absorption in one of the boundaries. Conditioned diffusion equations which are parallel to Fokker-Planck diffusion equations used in physics, were therefore, introduced as would be discussed in Narain (9). Based on similar considerations but using transition matrix approach, conditional Markov chains were adopted in

genetic studies as described in Narain (10). The transition matrix approach was subsequently adopted as in Narain and Pollak (11) to derive a formula for the probability of fixation of an allele at a single locus when the selection intensities fluctuate randomly over time.

Investigations by Kimura and Ohta (1973) and Maruyama (1974) showed that the average age of mutants at particular frequencies segregating in a finite population can be quite old. They adopted the diffusion approach to arrive at these results but ignored the possibility of quasi-fixation particularly when $4N_e v < 1$ where N_e is effective population size and v is the mutation rate per locus per generation and the number of possible allelic states is so large that whenever mutation occurs it leads to a new i.e. not a pre-existing allele. As shown in Narain (12), when $4N_e v < 1$, a conditional diffusion approach to the problem, wherein only those sample paths are considered which lead to the loss of the mutant allele by random drift, is required to be adopted.

The theory of distribution of time until fixation, mostly applicable to the case of a single locus with two alleles, was extended to the case of a single locus with <u>three</u> alleles as shown in Narain (13) using transition matrix approach. The numerical results obtained therein were further studied graphically in Narain (1983) in which the diffusion approach to the

problem of k-allele at a locus was also discussed by introducing multi-dimensional conditioned diffusion equations.

In the context of neutral theory of molecular evolution advocated by Kimura (1968), several mathematical population genetic models for the mutational production of new alleles have been introduced. In particular, 'infinite sites model' of Kimura (1969) assumes that the number of nucleotide sites for mutation is sufficiently large while the mutation rate per site is very low so that whenever a mutation occurs it represents a new site in which no mutant forms are segregating within the population. For such a model and using diffusion approach, Kimura (1969) discussed the statistical properties of the equilibrium distribution under steady flux of mutations. In Narain (14), however, a conditional diffusion approach to the problem by taking into account only those sample paths of the process that lead to the loss of the mutant forms from the populations, is adopted. In Narain (1983) the conditional distribution function itself was derived and the important statistical properties of the distribution were graphically exhibited.

5. Narain, P. and Robertson, A. (1969). Limits and duration of response to selection in finite populations: the use of transition probability matrices. The Indian Journal of Heredity, 1(1): 79-98.

The theory of limits of response to selection in finite populations was developed by Robertson (1960) primarily on the basis of diffusion approximation to the distribution of gene frequency. In the exact description of the stochastic process, we have a corresponding Finite Markov Chain for determining the expected change in the gene frequency by a given generation as well as in the limit in any genetic situation. In addition, we can also study the statistical properties of the distribution of time to fixation of the favoured allele, given that it is ultimately attained. This new approach of studying the duration of response to selection in finite population using transition matrix approach was attempted for the first time in this paper.

In a population of diploid individuals of constant size N with a single locus with two alleles A_1 and A_2, consider the gene frequency of A_1 as a discrete random variable x_i taking a finite set of (2N+1) values (i/2N), $i = 0,1,2,\ldots,2N$. This corresponds to a finite absorbing Markov chain with two absorbing states E_o and E_{2N} representing respectively the status of the populations with fixation of A_2 and A_1. The rest (2N-1)

EX BIBL. UNIV. EDINBURGEN

states $E_1, E_2, \ldots, E_{2N-1}$ are transient states any one of which representing the status of a population with mixtures of A_1 and A_2 genes. Assuming the process to be time homogeneous and p_{ij} as the transition probability for the population to move from E_i to E_j in one-step, the transition probability matrix $\underset{\sim}{P}$, of order $(2N+1) \times (2N+1)$ is given by

$$\underset{\sim}{P} = \begin{bmatrix} 1 & \underline{0}' & 0 \\ \underline{p}_0 & \underset{\sim}{Q} & \underline{p}_{2N} \\ 0 & \underline{0}' & 1 \end{bmatrix} \tag{5.1}$$

where $\underset{\sim}{Q}$ is a matrix of order $(2N-1) \times (2N-1)$ representing one-step transition probabilities amongst transient states only, $\underline{p}_0$ and $\underline{p}_{2N}$ are column vectors of order $(2N-1) \times 1$ representing one-step transition probabilities from a transient state to E_0 and E_{2N} respectively and $\underline{0}'$ is a row vector of order $(2N-1) \times 1$ consisting of zero elements.

Let $\underline{U}(t)$ denote the column vector of the fixation probability by t-th generation so that $\underline{U}(1)$, being the last column of $\underset{\sim}{P}$, give the fixation probability in one-step. Further, let $\underline{r}(t)$ denote the column vector of the expected changes in the gene frequency of A_1 at the t-th generation given that initially the population had frequency of A_1 as $(i/2N)$ for

$i=1,2,\ldots,(2N-1)$. Also, let $\underset{\sim}{\Delta p}$ denote the column vector of the changes in gene frequency of A_1 in a single step. Then it is shown that

$$(5.2) \qquad \underset{\sim}{U}(t) = (\underset{\sim}{I}-\underset{\sim}{Q}^{t})(\underset{\sim}{I}-\underset{\sim}{Q})^{-1}\underset{\sim}{U}(1)$$

$$(5.3) \qquad \underset{\sim}{r}(t) = (\underset{\sim}{I}-\underset{\sim}{Q}^{t})(\underset{\sim}{I}-\underset{\sim}{Q})^{-1}\underset{\sim}{\Delta p}$$

This gives matrix formulae for the eventual fixation probability vector $\underset{\sim}{U}$ and the vector $\underset{\sim}{r}$ of the expected change in the frequency of A_1 in the limit i.e. selection limit by letting $t \to \infty$,

$$(5.4) \qquad \underset{\sim}{U} = (\underset{\sim}{I}-\underset{\sim}{Q})^{-1}\underset{\sim}{U}(1)$$

$$(5.5) \qquad \underset{\sim}{r} = (\underset{\sim}{I}-\underset{\sim}{Q})^{-1}\underset{\sim}{\Delta p}$$

The fundamental matrix

$$(5.6) \qquad \underset{\sim}{T} = (\underset{\sim}{I}-\underset{\sim}{Q})^{-1}$$

gives the expected total number of times the population spends in the different transient states on the way to eventual fixation or loss of A_1 from a given state E_i for $i=1,2,\ldots,(2N-1)$. As the proportion of times that a population goes from a particular state to the fixation of A_1 is given by the elements of

the vector $\underline{U}$, the vector

$$(5.7) \qquad \underline{m} = (\underline{I}-\underline{Q})^{-1}\underline{U} = (\underline{I}-\underline{Q})^{-2}\underline{U}(1)$$

gives the expected total number of steps in the process which end up with the fixation of A_1 only. The average time until fixation of A_1 is therefore given by the ratio of the elements of vectors $\underline{m}$ and $\underline{U}$. The second moment of the distribution of time until fixation of A_1 is given by the ratio of the elements of vectors $\underline{v}$ and $\underline{U}$ where v is given by

$$(5.8) \qquad \underline{v} = [2(\underline{I}-\underline{Q})^{-3} - (\underline{I}-\underline{Q})^{-2}]\underline{U}(1)$$

The variance can then be obtained by subtracting the square of the mean from the second moment.

The theory of transition matrix so developed is applied to the selection process where one of the two alleles at the locus is favoured at the cost of the other. For the probability of transition from one state to another, we use binomial sampling model. Consider the processes of selection and sampling as occurring sequentially in that order, selection being at the gametic stage in favour of A_1 with a selective value of $(1+s/2)$ where s is small. Now the gene frequency in the different lines is distributed binomially with mean $2Np_i'$ and index 2N where p_i' is

the gene frequency of A_1 after selection and Δp_i is

$$(5.9) \qquad \Delta p_i = p_i' - p_i$$

$$= (s/2)p_i(1-p_i) \Big/ (1+p_i s/2)$$

Then p_{ij} becomes

$$(5.10) \qquad p_{ij} = \binom{2N}{j} (p_i')^j (1-p_i')^{2N-j}, \quad i,j=0,1,\ldots,2N$$

This transition probability, for transient states being a function of selective value of s and initial frequency p_i, could be expanded as a power series in s, and expressed in matrix terms as functions of $\underset{\sim}{Q}_0$, the transition matrix with $s = o$. Likewise, the fundamental matrix $\underset{\sim}{T}$, being a function of $\underset{\sim}{Q}$, can also be expressed as a series expansion in terms of $\underset{\sim}{Q}_0$ and $\underset{\sim}{T}_0$. The column vector $\underline{\Delta p}$ of the changes in gene frequency in the initial one generation is expressed as

$$(5.11) \qquad \underline{\Delta p} = s/2\,(1-s/4)\,\underline{x}_1 + (s^2/8)\,\underline{x}_2 + O(s^3)$$

where $\underline{x}_1$ and $\underline{x}_2$ are the first two right-hand eigen-vectors of $\underset{\sim}{Q}_0$ viz.

$$(5.12) \qquad \underline{x}_1 = [\, p_i(1-p_i) \,]$$

$$(5.13) \qquad \underline{x}_2 = [\, p_i(1-p_i)(1-2p_i) \,]$$

This enables us to compute an expression for the elements of vector $\underline{r}$ of the total expected change in the gene frequency of A_1 in the limit by operating functions of $\underset{\sim}{Q}_0$ onto the vectors of $\underset{\sim}{Q}_0$. This gives

$$(5.14) \qquad \underline{r} = \alpha_{10}\underline{x}_1 + \alpha_{20}\underline{x}_2$$

where α_{10} and α_{20} are functions of N and s but take the following limiting values when N tends to infinity and s tends to zero such that Ns tends to a finite limit

$$(5.15) \qquad \alpha_{10} \to Ns[1-(Ns)^2/15]$$

$$(5.16) \qquad \alpha_{20} \to (Ns)^2/3$$

For given initial frequency p, this approximation gives an explicit expression for the expected selection limit (L) as

$$(5.17) \qquad L = Nsp(1-p) + ((Ns)^2/3)p(1-p)(1-2p) + \ldots\ldots$$

Since the expected selection limit is the difference between the probability of fixation u(p) of the favoured allele and the initial frequency p, we get the formula for u(p) as

$$(5.18) \qquad u(p) = p + Nsp(1-p) + (1/3)(Ns)^2 p(1-p)(1-2p) + \ldots\ldots$$

which is the same as the series expansion of the formula for u(p) given by Kimura (1964) valid for $|Ns| < \pi$. This result connects the two approaches viz. diffusion and transition matrix for obtaining the probability of fixation of a gene which depends on initial frequency p and Ns instead of N and s separately.

In addition to the above results the paper also gives numerical results on the average and coefficient of variation of the time to fixation for various values of Ns and p for additive genes. It was found that both the mean time and coefficient of variation decrease as Ns increases at a given p. For a fixed Ns, the mean time is highest(lowest) at low(high) gene frequency but the coefficient of variation is highest(lowest) at high(low) gene frequency. When Ns=1 and p=1/2, the mean time is about 2.25N with a coefficient of variation of about 70 per cent.

A comparison of the average times to fixation with those of loss as well as to homozygosity indicated that when Ns=o and p=1/2, all the three coincide to give a value of 2.55N. But for p greater than or less than 1/2, the three differ, the time to homozygosity being always between the other two. For p less than 1/2, the mean time until homozygosity is much nearer to the mean time until loss than to that until fixation. A similar comparison holds true for the coefficient of variation as well.

N.B. The results published in this paper were earlier incorporated in the Ph.D. thesis RESPONSE TO SELECTION IN FINITE POPULATIONS submitted by me to Edinburgh University in 1969.

6. Narain, P. (1970). A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. Genet. Res. Camb., 15: 251-255.

Consider a mutant allele A_2 with frequency p, the normal allele A_1 having therefore frequency (1-p), in a diploid population of N individuals with variance effective number N_e. Let u(p,t) be the probability that mutant A_2 gets fixed by the t-th generation starting with frequency p at t=o. Let

$$(6.1) \qquad S_1(p) = \int_0^\infty t^2 \frac{\partial u(p,t)}{\partial t}\, dt$$

Then $V_1(p) = S_1(p)/u(p)$ represents the second moment about origin of the length of time until fixation of A_2 excluding the cases in which it is lost from the population. If $M_{\Delta p}$ and $V_{\Delta p}$ represent the mean and variance of the rate of change in the frequency of A_2 per generation, then the differential equation for $S_1(p)$ is given by

$$(6.2) \qquad \frac{1}{2} V_{\Delta p} \frac{d^2 S_1(p)}{dp^2} + M_{\Delta p} \frac{dS_1(p)}{dp} + 2T_1(p) = 0$$

where $T_1(p) = \int_0^\infty t \frac{\partial u(p,t)}{\partial t} dt = M_1(p)u(p)$ with $M_1(p)$ as the average length of time until fixation of A_2. An expression for $M_1(p)$ for neutral mutant was obtained by Kimura and Ohta (1969) as

$$(6.3) \qquad M_1(p) = -4N_e\left(\frac{1-p}{p}\right)\log_e(1-p)$$

The differential equation for $V_1(p)$ is obtained by differentiating $S_1(p) = V_1(p).u(p)$ twice and substituting in the differential equation for $S_1(p)$. For neutral genes, $M_{\Delta p}=0$, $V_{\Delta p} = p(1-p)/2N_e$, $u(p)=p$, $G(p)=1$. Using the formula for $M_1(p)$ for neutral genes, this gives the differential equation for $V_1(p)$ as

$$(6.4) \qquad \frac{d^2V_1(p)}{dp^2} + \frac{2}{p}\frac{dV_1(p)}{dp} - \frac{32N_e^2}{p^2}\log_e(1-p) = 0$$

Solving this differential equation subject to $\lim_{p\to 0} V_1(p) =$ finite and $V_1(1) = 0$, we get an expression for $V_1(p)$ as

$$(6.5) \qquad V_1(p) = 32N_e^2\left[\left(\frac{1-p}{p}\right)\log_e(1-p) + \frac{\pi^2}{6} - \sum_{k=1}^{\infty} p^k/k^2\right]$$

The expression for variance $V(p)$, obtained from $\{V_1(p)-[M_1(p)]^2\}$ indicates that the coefficient of variation, $\sqrt{V(p)}/M_1(p)$, would

be independent of the population size. As $p \to o$, $M_1(p)$ tends to $4N_e$ and $V(p)$ tends to $16N_e^2[(\pi^2/3)-3]$, giving a coefficient of variation of $[(\pi^2/3)-3]$, being about 54 per cent. Thus an originally rare neutral mutant gene, in a population of effective size N_e, takes $4N_e$ generations on an average, with a standard deviation of about $2N_e$ generations, until it spreads in the whole population.

A comparison of the mean and the standard deviation obtained as above by the diffusion approach with the exact values obtained by the transition matrix approach for a population of size 8, is shown in Table 6.1.

TABLE 6.1 Exact values and diffusion approximation (D.A.) for mean and standard deviation of the number of generations until fixation of A_2 ($N = 8$).

	Mean			Standard deviation		
p	Exact	D.A.	Difference	Exact	D.A.	Difference
0.125	28.30	29.92	1.62	16.58	17.19	0.61
0.250	26.06	27.60	1.54	16.46	17.04	0.58
0.375	23.57	25.04	1.47	16.19	16.76	0.57
0.500	20.76	22.16	1.40	15.73	16.27	0.54
0.625	17.50	18.80	1.30	14.96	15.48	0.52
0.750	13.60	14.80	1.20	13.65	14.14	0.49
0.875	8.55	9.52	0.97	11.14	11.60	0.46

It is apparent from Table 6.1 that a larger mean is associated with a larger standard deviation. The diffusion approximation overestimates both the mean and the standard deviation. However, while the overestimation for the mean is on an average, about one generation, it is only about half-generation for the standard deviation.

Citation

1. On pp. 432 and 557 in
Crow, J.F. and Kimura, M. (1970). An Introduction to Population Genetics Theory. Harper and Row, New York.

2. On pp. 161 and 163 in
Maruyama, T. and Kimura, M. (1974). Evolution, 28: 161-163.

3. On pp. 104 and 270 in
Nei, M. (1975). Molecular Population Genetics and Evolution. North-Holland Publishing Company, Amsterdam.

N.B. The results published in this paper were earlier incorporated in the Ph.D. thesis RESPONSE TO SELECTION IN FINITE POPULATIONS submitted by me to Edinburgh University in 1969.

7. Narain, P. (1971). Average number of generations required to attain limits of genetic improvement. SABRAO NEWSLETTER, 3(2): 135-142.

The selective capacity of a stock handled by a breeder depends upon the amount of genetic variability present in the stock. The response to selection is, therefore, predicted with the help of an estimate of the genetic variability and selection differential. However, this prediction of response to selection is only for a short term in the sense that it cannot predict the response after several generations of selection. For this purpose, it is necessary to know the factors which affect the limits of response to selection and also how many generations of selection are required to achieve the limit. Robertson (1960) developed a theory of limits in artificial selection for predicting the limits of response to selection. As already discussed, Narain and Robertson (5) developed this theory further so as to enable a prediction of the average number of generations needed to attain the selection limits. In this paper, simple applications of this theory, for the use of breeders, are given.

The dependence of the average number of generations required for attainment of limits (fixation of the favoured allele) divided by population size, on Ns and initial gene frequency (q) for additive, dominant and recessive genes is shown graphically in Figure 7.1 (a, b and c). It is apparent from

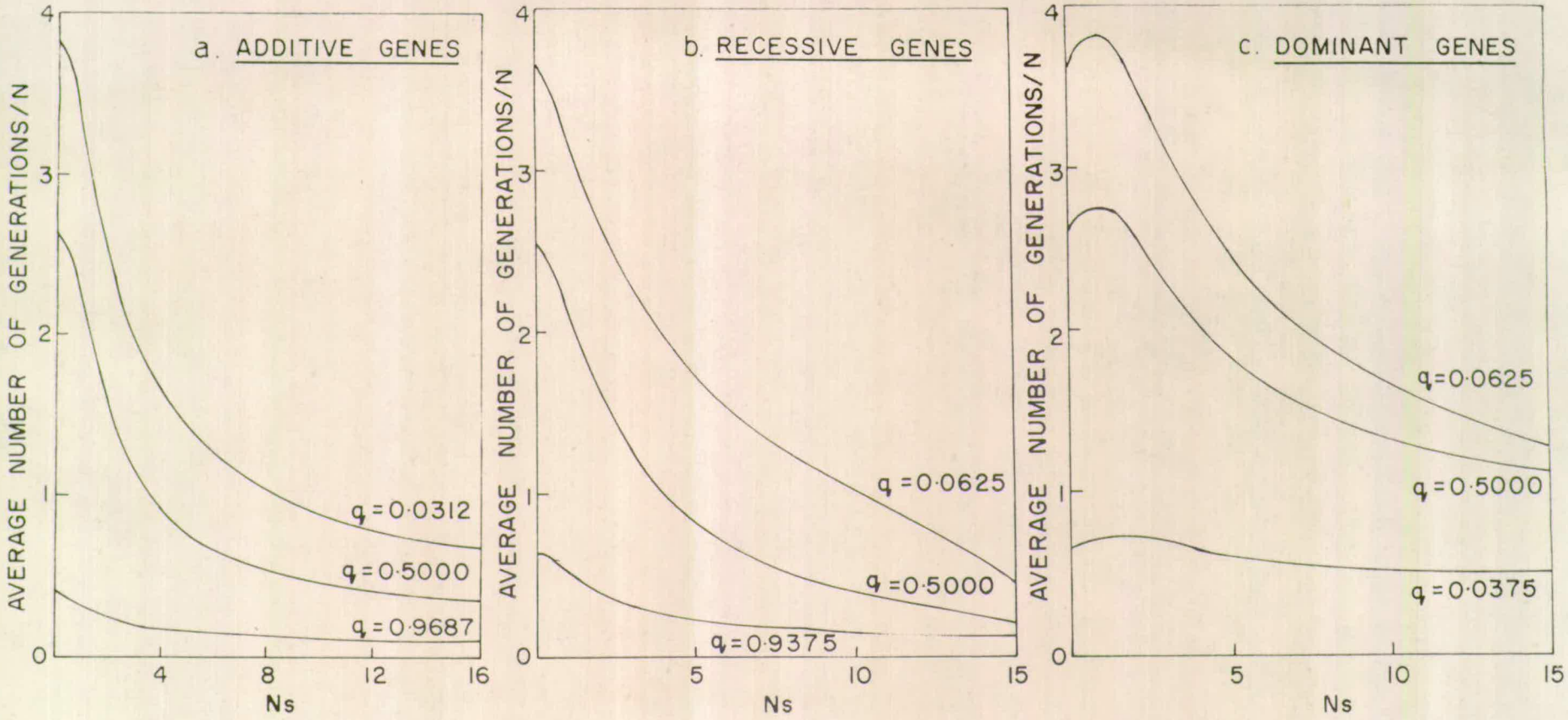

Fig.7.1 Average number of generations required for fixation divided by population size, for different values of Ns and initial gene frequency.

these graphs that the average number of generations decreases as Ns increases in so far as an additive or recessive gene is concerned. But for a dominant gene, the mean time increases initially for small values of Ns and decreases thereafter. In this case the maximum occurs at Ns=1 if the initial gene frequency is 1/2. For a gene with low initial frequency, however, the maximum occurs at a value of Ns less than 1 and for high initial gene frequency it occurs at a value of Ns greater than 1. When a value of Ns is fixed, the average number of generations is highest (lowest) at low (high) initial gene frequency. When Ns is 1, and the initial gene frequency is 1/2, the average number of generations is about 2.25N, 2.14N and 2.77N respectively for additive, recessive and dominant genes. A gene therefore, takes less time to reach fixation when it is recessive than when it is dominant.

The estimates of genetic progress per generation obtained by dividing the selection limit by the average time needed to attain it are presented in Table 7.1 for N=8 along with the corresponding values when N is infinite for comparison. The latter values pertain to the change in gene frequency after one generation of selection and depend on the selective value s and gene frequency q.

TABLE 7.1 Estimates of genetic progress per generation

Initial gene frequency	Value of s	Recessive		Additive		Dominant	
		N=8	N=∞	N=8	N=∞	N=8	N=∞
0.0625	0.125	0.0017	0.0005	0.0026	0.0037	0.0036	0.0069
	0.250	0.0039	0.0009	0.0070	0.0073	0.0093	0.0137
0.5000	0.125	0.0130	0.0156	0.0119	0.0156	0.0109	0.0156
	0.250	0.0274	0.0312	0.0233	0.0312	0.0193	0.0312
0.9375	0.125	0.0100	0.0069	0.0084	0.0037	0.0059	0.0005
	0.250	0.0187	0.0137	0.0135	0.0073	0.0091	0.0009

It is apparent from these results that the estimates of genetic change per generation in finite population are either lower or higher than the genetic change expected in an infinite population depending only on the initial gene frequency and the type of gene action and does not depend on the value of s. For initial gene frequency equal to half, the finite population estimates are lower but for high initial gene frequencies, they are greater than the corresponding infinite population values regardless of the type of gene action. For rare genes, on the other hand, the finite population estimates are greater than the infinite population values only in the case of recessive

genes. For rare additive and dominant genes, the lowering of estimates due to finite population is about 5 to 47 per cent.

The process of random change in gene frequency due to finite size of the population introduces variation in gene frequencies among lines but the average of gene frequencies over lines remains the same as expected in an infinite population. The genetic change per generation in a finite population is therefore expected to be the same as in an infinite population. The above results, however, show considerable bias which is understandable since we are only considering the average of the distribution of time to fixation which is positively skewed (Kimura, 1970).

8. Narain, P. (1971). The use of transition probability matrices in studies on limits of response to selection. Journal of Indian Society of Agricultural Statistics, 23(2): 48-60.

Let the expected frequency of allele A by the t-th generation be denoted by $E[q_i(t)]$ when the initial population had its frequency as $q_i(o) = \frac{i}{2N}$. The expected response in the gene frequency by the t-th generation is then

$$(8.1) \qquad E[R_i(t)] = E[q_i(t)] - q_i(o)$$

In vector notations, it is

$$E[\underline{R}(t)] = E[\underline{q}(t)] - \underline{q}(o) \tag{8.2}$$

Now the expected frequency by the t-th generation can be obtained by finding the mean of the variate $x_j=j/2N$ for the distribution given by the i-th row of matrix $\underset{\sim}{P}$ given in Narain and Robertson (5) raised to power t. This gives, in matrix notations,

$$E[\underline{q}(t)] = \underset{\sim}{Q}^{t}\,\underline{q}(o) + \underline{U}(t) \tag{8.3}$$

Similarly, if the expectation of response vector in the first generation is denoted by $E[\underline{\Delta q}]$, we have

$$E[\underline{\Delta q}] = \underset{\sim}{Q}\underline{q}(o) + \underline{U}(1) - \underline{q}(o) \tag{8.4}$$

After some simplification, we get the result

$$\begin{aligned} E[\underline{R}(t)] &= (\underset{\sim}{I} - \underset{\sim}{Q}^{t})(\underset{\sim}{I} - \underset{\sim}{Q})^{-1} E(\underline{\Delta q}) \\ &= (\underset{\sim}{I} - \underset{\sim}{Q}^{t})E(\underline{R}) \end{aligned} \tag{8.5}$$

as we obtained in Narain and Robertson (5) from heuristic considerations.

Analytically, we can evaluate these quantities provided we know the eigen-roots and vectors of $\underset{\sim}{Q}$. Let $\alpha_1, \alpha_2, \ldots, \alpha_{2N-1}$ be eigen-roots of $\underset{\sim}{Q}$ with $\underline{x}_i'$ and $\underline{y}_i$ as the right- and left-hand eigen-vectors corresponding to the root α_i. Then

$$(8.6) \qquad E[\underline{R}(t)] = \sum_{i=1}^{2N-1} (1-\alpha_i^t)\underline{x}_i\underline{y}_i' \, E(\underline{R})$$

$$(8.7) \qquad E[\underline{R}] = \sum_{i=1}^{2N-1} (1-\alpha_i)^{-1} \underline{x}_i\underline{y}_i' \, E(\underline{\Delta q})$$

The simplest application to demonstrate the power of the transition matrix approach in getting analytical expressions for a genetic situation is a mating system known as self-fertilization, mostly practised in plants. In this case, males and females are necessarily of the same genotype. This reduces the number of all possible types of mating under random mating to only those between identical genotypes and corresponds to a situation when N=1. The population gets divided into an infinite number of lines from each of which two gametes are drawn to form one mature individual. With two loci each with two alleles A-a and B-b on the same chromosome with a cross-over probability r, there would therefore be ten types of lines corresponding to 10 states of the system as given below:

	AB	Ab	aB	ab
E_1	1	0	0	0
E_2	0	1	0	0
E_3	0	0	1	0
E_4	0	0	0	1
E_5	1/2	1/2	0	0
E_6	1/2	0	1/2	0
E_7	0	1/2	0	1/2
E_8	0	0	1/2	1/2
E_9	s/2	r/2	r/2	s/2
E_{10}	r/2	s/2	s/2	r/2

where s=(1-r) and E_9 and E_{10} would have only double heterozygotes of the type AB/ab (coupling) and Ab/aB (repulsion) respectively. There are now four absorbing states AB, Ab, aB and ab. Following the theory outlined above, we get analytical expressions for the probability of fixation of gametes AB, Ab, aB and ab from the six initial states as follows, where v = 1/(1+2r).

Initial States

	E_5	E_6	E_7	E_8	E_9	E_{10}
AB	1/2	1/2	0	0	v/2	rv
Ab	1/2	0	1/2	0	rv	v/2
aB	0	1/2	0	1/2	rv	v/2
ab	0	0	1/2	1/2	v/2	rv

As expected, the linkage can only have effect if the population starts in E_9 or E_{10} and that the probability of fixation of gamete AB (gamete Ab) is the same as that of gamete ab (gamete aB). The effect of linkage is to increase or decrease the probability of fixation of a coupled gamete (AB or ab) according as the initial population consists of a coupling or repulsion heterozygote respectively. As expected, the reverse is found to be true for the repulsion gametes (Ab or aB).

9. Narain, P. (1974). The conditioned diffusion equation and its use in population genetics. Journal of the Royal Statistical Society, B36: 258-266.

The problem of studying the statistical properties of the distribution of time to fixation (extinction) of a particular gene in a finite population disregarding the cases in which it is lost (fixed) attempted by Kimura and Ohta (1969) as well as by Narain (1970) using the diffusion approach can be tackled in a more general way by invoking a conditioned diffusion approach as introduced in this paper. It essentially gives the same results as obtained earlier but the methodology used is new and more general, capable of being applied to fields other than genetics.

Let the conditional density function of a random variable x at time t, given that its final value at $t=\infty$ will be 1, be denoted by $f_{c1}(x;t)$ that is

$$(9.1) \qquad f_{c1}(x;t) = \lim_{t_1 \to \infty} f_{c1}(x;t,1;t_1), \quad t < t_1$$

as against the unconditional density function f(x;t) which satisfies the forward Kolmogorov (or Fokker-Planck or Ficks) equation,

$$(9.2) \qquad \frac{f(x;t)}{t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}[v(x)f(x;t)] - \frac{\partial}{\partial x}[m(x)f(x;t)]$$

where m(x) and v(x) are the instantaneous drift and diffusion coefficients respectively. We have to first know the probability $u(p) = \lim_{t\to\infty} u(p;t)$ that the boundary x=1 is eventually reached before the boundary x=0, given that the process starts with the value of x as p. The probability u(p;t) that x=1 during a time interval t_1 given that initially it takes the value p, however, satisfies the backward Kolmogorov equation

$$(9.3) \qquad -\frac{\partial u(p,t)}{\partial t} = L\,u(p;t)$$

where L is a second order partial differential operator given by

$$(9.4) \qquad L = \frac{1}{2}\,v(p)\frac{\partial^2}{\partial p^2} + m(p)\frac{\partial}{\partial p}$$

By setting $\frac{\partial u(p,t)}{\partial t} = 0$ and solving the differential equation subject to $u(0)=0$, $u(1)=1$, we can get $u(p)$.

The conditional and unconditional density functions are then, applying Baye's theorem, connected as

$$u(p)f_{c1}(x;t) = f(x;t)u(x) \tag{9.5}$$

This enables us to transform the forward equation of Kolmogorov to the following conditioned forward diffusion equation relative to the event of the process attaining absorption at $x=1$,

$$\frac{\partial f_{c1}(x;t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}[v(x)f_{c1}(x;t)] - \frac{\partial}{\partial x}[m_1^*(x)f_{c1}(x;t)] \tag{9.6}$$

where

$$m_1^*(x) = m(x) + v(x)\frac{d}{dx}[\log u(x)] \tag{9.7}$$

showing thereby that while the diffusion coefficient remains the same in the conditioned process, the drift coefficient changes and can be determined if we know the form of the probability of fixation function. The adjoint of the conditioned forward diffusion equation, termed as conditioned backward diffusion equation relative to the event of the process attaining absorption at $x=1$ is then given by

$$(9.8) \qquad -\frac{\partial u_{c1}(p;t)}{\partial t} = L_{c1}\, u_{c1}(p;t)$$

where L_{c1}, a second order differential operator, is

$$(9.9) \qquad L_{c1} = \frac{1}{2}v(p)\frac{\partial^2}{\partial p^2} + m_1^*(p)\frac{\partial}{\partial p}$$

and $u_{c1}(p;t)$ is the probability that x=1 during a time interval t, given that initially it takes the value p and relative to the event of the process attaining absorption at x=1. In a similar manner, we can treat the case when the conditioning of the process is relative to its absorption at x=o.

We now define the mean time until absorption at x=1 by

$$(9.10) \qquad M_{c1}(p) = \int_o^\infty t\, u_{c1}(p;t)dt$$

The corresponding differential equations along with boundary conditions are found to be

$$(9.11) \qquad L_{c1}M_{c1}(p) + 1 = o$$

subject to $\lim_{p\to o} M_{c1}(p) =$ a finite quantity; $M_{c1}(1) = o$.

The integration of this differential equation gives

$$(9.12) \qquad M_{c1}(p) = \int_p^1 I(p)u(p)[1-u(p)]dp + \frac{1-u(p)}{u(p)}\int_o^p I(p)\{u(p)\}^2 dp$$

$$(9.13) \qquad \lim_{p \to o} M_{c1}(p) = \int_o^1 I(p)u(p)[1-u(p)]dp$$

where $I(x) = 2\int_o^1 G(x)dx \Big/ [v(x)G(x)]$ and $G(x) = \exp[-2\int\{m(x)/v(x)\}dx]$

In a similar manner we can get the expression for the second moment of the distribution of time until absorption in the given boundary and using it along with that of mean time we can get the variance of the length of time until absorption in that boundary.

These general results are applied, in the paper, to the problem of pure random drift, giving essentially the same results for mean and variance of time until fixation as well as of time until extinction of a particular gene as obtained earlier by Kimura and Ohta (1969) and Narain (6).

Citation

1. On pp. 628 and 629 in
Aase, K.K. (1977). Journal of Applied Probability, 14: 626-629.

2. On pp. 184 and 196 in
Watterson, G.A. (1977). Theoretical Population Biology, 12: 179-196.

3. On pp. 218 and 225 in
Littler, R.A. and Good, A.J. (1978). Theoretical Population Biology, 13: 214-225.

10. Narain, P. (1977). The conditional Markov chain in a genetic context. Journal of Genetics, 63(2): 49-62.

We consider the Markov chain set up already described in Narain (5 and 8). We necessarily assume the absence of mutation so that there are two absorbing states for the single locus with two alleles A and a. The vectors of the eventual probabilities of fixation of A and a, denoted by $\underset{\sim}{U}$ and $\underset{\sim}{L}$ respectively are obtained by operation $(\underset{\sim}{I}-\underset{\sim}{Q})^{-1}$ on $\underset{\sim}{p}_{2N}$ and $\underset{\sim}{p}_{o}$ respectively. Now consider a finite absorbing Markov chain conditional to the eventual absorption in E_{2N}. We then have only one absorbing state E_{2N} and (2N-1) transient states $E_1, E_2, \ldots, E_{2N-1}$ from which absorption is possible only in E_{2N}. Let $p_{ij}^{(cl)}$ be the one-step transition probability for the system to move from E_i to E_j relative to the event of ultimate absorption in E_{2N}. If i-th element of $\underset{\sim}{U}$ is U_i, then

$$(10.1) \qquad p_{ij}^{(cl)} = p_{ij}\, U_j/U_i$$

with $U_{2N}=1$. This gives the transition probability matrix $\underset{\sim}{P}^{(cl)}$ as

$$(10.2) \qquad \underset{\sim}{P}^{(cl)} = \begin{bmatrix} \underset{\sim}{Q}^{(cl)} & \underset{\sim}{P}_{2N}^{(cl)} \\ 0 & 1 \end{bmatrix}$$

where $\underset{\sim}{Q}^{(c1)}$ is of order $(2N-1)\times(2N-1)$, giving one-step transition probabilities amongst transient states only, conditional to fixation in E_{2N} and $\underline{P}_{2N}^{(c1)}$ is a column vector of order $(2N-1)\times 1$ representing the one-step transition probability from a transient state to E_{2N} relative to the eventual absorption in E_{2N}. Using the Chapman-Kolmogorov (Feller, 1951) equations, we get

$$(10.3)\qquad \underset{\sim}{Q}^{(c1)}(t) = [\underset{\sim}{Q}^{(c1)}]^{t} = \underset{\sim}{D}_1^{-1}\,\underset{\sim}{Q}^{t}\,\underset{\sim}{D}_1$$

$$(10.4)\qquad \underline{U}^{(c1)}(t) = [\underset{\sim}{I}-\underset{\sim}{Q}^{(c1)}(t)][\underset{\sim}{I}-\underset{\sim}{Q}^{(c1)}]^{-1}\,\underline{P}_{2N}^{(c1)}$$

$$= [\underset{\sim}{I}-\underset{\sim}{D}_1^{-1}\,\underset{\sim}{Q}^{t}\,\underset{\sim}{D}_1]\underline{e}$$

giving the probabilities of fixation of A by t-th generation in the conditional process, where $\underset{\sim}{D}_1 = \mathrm{diag}(U_1, U_2, \ldots U_{2N-1})$. Similarly, we can treat the case for a finite absorbing Markov chain conditional to the eventual absorption in E_o.

The vector of conditional expected response due to selection by the t-th generation is obtained as

$$(10.5)\qquad E[\underline{R}^{(c)}(t)] = [\underset{\sim}{I}-(\underset{\sim}{Q}^{(c1)})^{t}]E[\underline{R}^{(c)}]$$

$$= [\underset{\sim}{I}-\underset{\sim}{D}_1^{-1}\,\underset{\sim}{Q}^{t}\,\underset{\sim}{D}_1][\underline{e}-\underline{p}(o)]$$

relative to the eventual fixation of A regarded as a desirable allele where $\underset{\sim}{p}(o)$ is the vector of the frequency of desirable allele in the initial population.

The conditional Markov chain approach can also be used to derive the probability generating function of the distribution of time until fixation of a particular allele say A_1, from which the first and second moments could be obtained. It can be shown that the vector of the probability generating function $\underset{\sim}{\pi}^{(c1)}(z)$, is given by

$$(10.6) \qquad \underset{\sim}{\pi}^{(c1)}(z) = z\underset{\sim}{D}_1^{-1}(\underset{\sim}{I}-z\underset{\sim}{Q})^{-1}(\underset{\sim}{I}-\underset{\sim}{Q})\underset{\sim}{U}$$

This gives the vectors of the first and second moments as

$$(10.7) \qquad E(T_1) = \underset{\sim}{D}_1^{-1}(\underset{\sim}{I}-\underset{\sim}{Q})^{-1}\underset{\sim}{U}$$

$$(10.8) \qquad E(T_1^2) = \underset{\sim}{D}_1^{-1}[2(\underset{\sim}{I}-\underset{\sim}{Q})^{-2}-(\underset{\sim}{I}-\underset{\sim}{Q})^{-1}]\underset{\sim}{U}$$

as we got earlier by direct methods in Narain and Robertson (5).

For the unconditional case, analytical results on some of the properties of Markov chain could be obtained by making use of the eigen-roots and eigen-vectors of the matrices $\underset{\sim}{P}$ and $\underset{\sim}{Q}$ when we specify the transition probability p_{ij} as following the

binomial law as in Narain and Robertson (5). In such a case, we made use of the eigen-roots and eigen-vectors of $\underset{\sim}{P}$ with no selection, as derived by Feller (1951). For the conditional case also the same procedure could be used to derive the eigen-roots and vectors, assuming no selection. In this case,

$$(10.9) \qquad p_{ij}^{(c1)} = \binom{2N}{j} p_i^{\,j} (1-p_i)^{2N-j} (p_j/p_i) \qquad \begin{matrix} i=1,2,\ldots,2N \\ j=1,2,\ldots,2N \end{matrix}$$

where $p_i = i/2N$. The eigen-roots are found to be

$$(10.10) \qquad \alpha_r^{(c)} = (1-r/2N)\binom{2N}{r} \lfloor r / (2N)^r \qquad r=0,1,\ldots,(2N-1)$$

For $r=1,2,\ldots,(2N-1)$, these roots happen to be the same as that of $\underset{\sim}{Q}^{(c1)}$ and similarly as that of $\underset{\sim}{Q}^{(co)}$ which are same as that of $\underset{\sim}{Q}$ viz.

$$(10.11) \qquad \alpha_r = \binom{2N}{r} \lfloor r / (2N)^r \qquad r=2,3,\ldots,(2N)$$

Following Feller's procedure, the eigen-vectors for the conditional case can also be obtained. We ultimately get, for $j=1,2,\ldots,2N$,

$$(10.12) \qquad x_{j1} = (1-p_j) \quad \text{corresponding to } \alpha_1^{(c)} = (1-1/2N)$$

(10.13) $x_{j2} = (1-p_j)(1-2p_j)$ corresponding to

$$\alpha_2^{(c)} = (1-1/2N)(1-2/2N)$$

(10.14) $x_{j3} = (1-p_j)[\frac{2N-1}{1oN-6} - p_j(1-p_j)]$ corresponding to

$$\alpha_3^{(c)} = (1-1/2N)(1-2/2N)(1-3/2N)$$

We thus find that the elements of the vectors are $(1/p_j)$ times those of in the unconditional case.

For the conditional case the expected value of $j_{(v)}=j(j-1)$, $\ldots(j-v+1)$ is found to be

(10.15) $$E(j_{(v)}) = [(2N)_{(v+1)}p_i^{v} + v(2N)_{(v)}p_i^{v-1}]/2N$$

Taking v=1 and 2, we get expressions which ultimately lead to

(10.16) $$E(\Delta p_i) = (1-p_i)/2N$$

(10.17) $$V(\Delta p_i) = (1/2N)(1-1/2N)p_i(1-p_i)$$

If we take population so large that $(1/2N)^2$ can be neglected, we find that the mean and variance of the change in gene frequency due to random drift, in the conditional process, are $(1-p_i)/2N$ and $p_i(1-p_i)/2N$ as against o and $p_i(1-p_i)/2N$ respectively in the

unconditional case. This is exactly what we got from the diffusion approach for the pure drift case as shown in Narain (9). It is also interesting to note that if $(1/2N)^2$ is not negligible, the variance decreases along with increase in the mean in the exact conditional process. The latter has been termed as a fictitious drift by Nagasawa and Maruyama (1979).

In this paper, the conditional Markov chain approach is also applied to the problem of the effect of linkage on the mean and variance of time until fixation of a gamete in selfed populations. The problem of determining the probability of fixation of a gamete, in such a case, has already been discussed in Narain (8). Taking the P-matrix of the process and U_{AB} vector of the probabilities of fixation of gamete AB given therein, the $\underset{\sim}{P}^{(c11)}$ and hence $\underset{\sim}{Q}^{(c11)}$ matrices for the process conditional to absorption in AB/AB are obtained. This ultimately gives the vectors of mean as well as second moment about origin of time until fixation of AB as

$$(10.18) \qquad E[T'_{AB}] = [2, 2, 0, 0, \alpha^{(c)}_{AB}, \alpha^{(r)}_{AB}]$$

$$(10.19) \qquad E[T^{2'}_{AB}] = [6, 6, 0, 0, \beta^{(c)}_{AB}, \beta^{(r)}_{AB}]$$

where $\alpha^{(c)}_{AB}$, $\beta^{(c)}_{AB}$ corresponding to the situation when the population is initially in the coupling phase, are given by

$$(10.20)\qquad \alpha_{AB}^{(c)} = (1+2r)(1+4rs)/(1+2rs) + (1-2r)/(1+2r)$$

$$(10.21)\qquad \beta_{AB}^{(r)} = (1+2r)(3+26rs+24r^2s^2)/(1+2rs)^2 + (1-2r)(3-2r)/(1+2r)^2$$

and $\alpha_{AB}^{(r)}$, $\beta_{AB}^{(r)}$ corresponding to the situation when the popula- is initially in the repulsion phase, are given by

$$(10.22)\qquad \alpha_{AB}^{(r)} = (1+2r)(1+4rs)/2r(1+2rs) - (1-2r)/2r(1+2r)$$

$$(10.23)\qquad \beta_{AB}^{(r)} = (1+2r)(3+26rs+24r^2s^2)/2r(1+2rs)^2 - (1-2r)(3-2r)/2r(1+2r)^2$$

In a similar manner, we get the corresponding vectors for moments of time until fixation of Ab, aB and ab. As expected, results for time until fixation of ab are the same as that of AB. It is further found that when the population is initially in the coupling phase, the means and second moments about origin of time until fixation of Ab as well as aB are the same as $\alpha_{AB}^{(r)}$ and $\beta_{AB}^{(r)}$ respectively whereas when the initial popula- tion is in repulsion phase, these are correspondingly $\alpha_{AB}^{(c)}$ and $\beta_{AB}^{(c)}$.

The numerical results about the effect of recombination fraction r on the mean and the standard deviation of time until fixation for the case when the initial population is in the coupling phase are given in Table 10.1. The results for the case when the initial population is in repulsion phase are obtainable from this Table by interchanging either A and a or B and b.

TABLE 10.1 Mean and standard deviation (s.d.) of the number of generations until fixation of a gamete for the initial population with heterozygotes in coupling phase.

r	AB or ab		Ab or aB	
	Mean	s.d.	Mean	s.d.
0.0000	2.0000	1.4142	4.0000	2.0000
0.0625	2.0208	1.4337	3.7244	1.8622
0.1250	2.0743	1.4753	3.4973	1.7682
0.1875	2.1506	1.5213	3.3103	1.7078
0.2500	2.2424	1.5635	3.1516	1.6709
0.3125	2.3441	1.5959	3.0120	1.6486
0.3750	2.4513	1.6187	2.8873	1.6382
0.4375	2.5602	1.6306	2.7734	1.6354
0.5000	2.6666	1.6329	2.6666	1.6329

It is apparent from Table 10.1 that linkage decreases (increases) or increases (decreases) the average and standard deviation of

time to fixation of a coupled (repulsed) gamete according as the initial population consists of a coupling or repulsion heterozygote respectively.

11. Narain, P. and E. Pollak (1977). On the fixation probability of a gene under random fluctuations in selection intensities in small populations. Genetical Research, 29: 113-121.

In this paper, the problem of determining the probability of fixation of a particular allele in a finite population under fluctuation of selection intensities is tackled using Markov chain methods. All along it is assumed that there are N monoecious individuals having two alleles A and a and the population reproduces in discrete generations. If in generation n, the fitness coefficients of A and a are respectively $(1+s_1)$ and $(1+s_2)$, the selection intensities s_1 and s_2 are assumed to fluctuate over time in a random manner with identical distribution functions in all generations and independently between generations with parameters $E(s_1) = \bar{s}_1$, $E(s_2) = \bar{s}_2$, $Var(s_1) = v_1$, $Var(s_2) = v_2$, $Cov(s_1, s_2) = r$, $|r| \leq \sqrt{v_1 v_2}$. This model is virtually the same as that of Karlin and Levikson (1974) but the parameters are not taken to be of the order of (1/2N). The frequency of the A-gene in generation n, given the frequency before selection as $p_i = i/2N$ is then

$$(11.1) \qquad p_i^{(n)} = p_i + p_i(1-p_i)(s_1-s_2)/[1+s_2+(s_1-s_2)p_i]$$

Employing the standard Wright-Fisher binomial model with parameters $(2N, p_i^{(n)})$, the transition probability p_{ij} is a function of (s_1, s_2). We can expand it, by Taylor's expansion, as a function of two variables and follow the analytical procedure of Narain and Robertson (5) for treating such a process. The expectation of p_{ij} with respect to the distribution of s_1 and s_2 is then found to be

$$(11.2) \qquad E(p_{ij}) = p_{ij}(o)[1+2N\{(\bar{s}_1-\bar{s}_2)(1-\bar{s}_2)+v_2-r\}(p_j-p_i) + N\{(\bar{s}_1-\bar{s}_2)^2+v_1+v_2-2r\}\{(2N-1)(p_j-p_i)^2 - p_j(1-p_j)-2p_i(p_j-p_i)\}]$$

Because of the assumption of (s_1, s_2) having the same distribution in each generation and independently between generation, $E(p_{ij})$ is a one-step transition probability for any n. For such a Markov chain, homogeneous in time, the fixation probability of the A-gene, denoted by $u(p_i)$, is obtained by determining the vector of the total expected change in the frequency at the limit, $L = (L_1, L_2, \ldots, L_{2N-1})$ where $L_i = u(p_i)-p_i$. This in turn is obtained by operating the matrix $\underset{\sim}{T} = (\underset{\sim}{I}-\underset{\sim}{Q})^{-1}$ on to the vector of the initial expected change in the frequency vector $E(\underline{\Delta p})$.

In a manner similar to that used for getting $E(p_{ij})$, we can get $E_{s_1,s_2}(\underset{\sim}{T})$ in terms of the parameters of the distribution of (s_1,s_2) and matrices $\underset{\sim}{T}_o$, $\underset{\sim}{Q}_o$ and certain related matrices for $s_1=s_2=o$. Also, $E(\underline{\Delta p})$ in terms of vector notation, is found to be

$$(11.3) \qquad E(\underline{\Delta p}) = (1-\frac{\bar{s}_1+\bar{s}_2}{2})(\bar{s}_1-\bar{s}_2)\underline{x}_1+\frac{1}{2}(\bar{s}_1-\bar{s}_2)^2\underline{x}_2 + \frac{1}{2}(v_1+v_2-2r)\underline{x}_2-\frac{1}{2}(v_1-v_2)\underline{x}_1$$

where $\underline{x}_1$ and $\underline{x}_2$ are the vectors as in Narain and Robertson (5) corresponding to the first two eigen values of $\underset{\sim}{Q}_o$. Following the results given in Narain and Robertson (5), the operation of $\underset{\sim}{T}$ on $E(\underline{\Delta p})$ gives finally

$$(11.4) \qquad \underline{L} = \underline{L}_o - N(v_1-v_2)\underline{x}_1 + \frac{N^2}{3N-1}(v_1+v_2-2r)\underline{x}_2$$

where L_o denotes the corresponding vector for the non-random selection model with $v_1=v_2=r=o$, given by

$$(11.5) \qquad \underline{L}_o = 2N(\bar{s}_1-\bar{s}_2)(1-\frac{\bar{s}_1+\bar{s}_2}{2})\underline{x}_1 + \frac{N^2(4N-1)}{3N-1}(\bar{s}_1-\bar{s}_2)^2\underline{x}_2$$

For large values of N and small values of $\bar{s}_1,\bar{s}_2,v_1,v_2$ and r such that $N(\bar{s}_1-\bar{s}_2)$, Nv_1, Nv_2 and Nr remain constant, these

formulae result in approximate expressions for $u(p_i)$ and $u_o(p_i)$. If we define

$$(11.6) \qquad \alpha = (2v_2-v_1-r)/(v_1+v_2-2r)$$

the expression for the probability of fixation $u(p_i)$ is finally given by

$$(11.7) \qquad u(p_i) = p_i + 2\{N(\bar{s}_1-\bar{s}_2)\}p_i(1-p_i) + \frac{4}{3}\{N(\bar{s}_1-\bar{s}_2)\}^2 p_i(1-p_i)(1-2p_i) - \frac{2}{3}\{N(v_1+v_2-2r)\}p_i(1-p_i)(\alpha-p_i)$$

In general, therefore, the fixation probability of the A-gene, under random fluctuations in selection intensities, either decreases or increases over its value with non-random selection, depending upon whether p_i, the initial gene frequency is less or greater than α.

Citation

1. On p.5817 in
Takahata, N. and Kimura, M. (1979). Proceedings National Academy of Science, U.S.A., 76: 5813-5817.

12. Narain, P. (1977). On the average age of a mutant in finite populations. Bulletin International Statistical Institute, 47(2): 165-170.

If $f_{co}(p,x;t)$, as already discussed in the method given by Narain (9), is the conditional probability density function, the i-th moment (i=o,1,2,....) of t, the time interval in generations for an allele to have frequency x, starting with p and relative to the event that it eventually disappears from the population is given by

$$(12.1) \qquad T_{co}^{(i)}(p,x) = \int_o^\infty t^i f_{co}(p,x;t)dt$$

Then $T_{co}^{(i)}(p,x)$ is found to satisfy the differential equation

$$(12.2) \qquad \frac{d^2}{dx^2}\left[\frac{x(1-x)}{4N_e}T_{co}^{(i)}(p,x)\right] - \frac{d}{dx}\left[\{m(x) + \frac{x(1-x)}{2N_e}\frac{d}{dx}\log_e(1-u(x)\}\right.$$
$$\left. T_{co}^{(i)}(p,x)\right] + i\, T_{co}^{(i-1)}(p,x) = 0$$

This differential equation is subject to boundary condition that as x approaches zero, $T_{co}^{(i)}(p,x)$ approaches a finite quantity. In this equation, if we put i=o, we get

$$(12.3) \qquad T_{co}^{(o)}(p,x) = \int_o^\infty f_{co}(p,x;t)dt$$

which is nothing else but the expected total number of visits to a particular frequency x in respect of those sample paths which end up with the loss of the allele from the population. This is often termed as 'conditional sojourn time' and its counterpart, in the transition matrix approach, is $[I-Q^{(co)}]^{-1}$, the fundamental matrix of the conditional Markov chain as we have already discussed. By putting i = 1 and 2 and solving the differential equation successively, we get $T_{co}^{(1)}(p,x)$ and $T_{co}^{(2)}(p,x)$ respectively. The average age and second moment about origin of the age denoted respectively by $E[t_{co}(p,x)]$ and $E[t_{co}^2(p,x)]$ are then determined by dividing these quantities by $T_{co}^{(o)}(p,x)$. It is important to note that for solving the differential equation for $T_{co}^{(1)}(p,x)$, the boundary conditions are such that as x approaches o, the average age approaches the mean time until loss of the allele. Similarly, the boundary conditions for the differential equation for $T_{co}^{(2)}(p,x)$ are such that as x tends to o, the second moment about origin tends to the mean square time until loss of the allele.

In the neutral case, m(x) = o, u(x) = x and we get

(12.4) $T_{co}^{(o)}(p,x) = 4N_e$, being independent of p and x

(12.5) $$E[t_{co}(p,x)] = -4N_e\left[\left(\frac{p}{1-p}\right)\log_e p + \left(\frac{1-x}{x}\right)\log_e(1-x) + 1\right]$$

In the expressions, if we let p tend to unity, we get average time until fixation of an allele with initial frequency x. This is quite expected since the ultimate value of x is zero.

For the model of infinite iso-alleles with $4N_e v < 1$, $m(x) = -vx$, $u(x) = 1-(1-x)^F$, $F = 1-4N_e v$ and we get

$$(12.6) \qquad T_{co}^{(o)}(p,x) = \frac{4N_e}{F}\left[\frac{1-(1-x)^F}{x}\right], \quad \text{being independent of } p \text{ but not of } x$$

$$(12.7) \qquad E[t_{co}(p,x,v)] = \frac{4N_e}{F}\left[\int_x^1 \frac{1-(1-x)^F}{x}\,dx + \frac{(1-x)^F}{1-(1-x)^F}\int_0^x \frac{\{1-(1-x)^F\}^2}{x(1-x)^F}\,dx - (1-p)^{-F}\int_p^1 \frac{(1-p)^F-(1-x)^F}{x}\,dx\right]$$

If we let p approach unity, the average age coincides with the average time until fixation of an allele with frequency x. On the other hand, if we let x approach zero, we get the mean time until loss of the allele with initial frequency p.

Citation

1. On pp. 254 and 265 in
Tavaré, S. (1979). Theoretical Population Biology, 16: 253-266.

13. Narain, P. (1978). Average time until fixation of a mutant at a tri-allelic locus in a finite population. Indian Journal of Genetics and Plant Breeding, 38(1): 57-62.

Consider a finite population of gametes derived from a population of diploid individuals of constant size N and a single locus with three alleles A_1, A_2 and A_3. The sampling of gametes due to finite size of population causes a random change, from generation to generation, in the frequency of a given allele. The population gets sub-divided into several lines with different gene frequencies but with random breeding within lines. There would be seven types of lines out of which three would be homozygous for each of three alleles, three would be such that any two out of three alleles would be segregating and one would be such that all the three alleles would be segregating. If $p_{(i_1 i_2)(j_1 j_2)}$ represents the conditional probability that there are j_1, A_1 genes, j_2, A_2 genes out of 2N genes in the line, given that there were i_1, A_1 genes and i_2, A_2 genes in this line in the previous generation, the $(N+1)(2N+1) \times (N+1)(2N+1)$ transition probabilities determine a transition matrix $\underset{\sim}{P}$ given by

$$(13.1) \qquad \underset{\sim}{P} = \begin{bmatrix} \underset{\sim}{I} & \underset{\sim}{Q} \\ \underset{\sim}{R} & \underset{\sim}{S} \end{bmatrix}$$

where $\underset{\sim}{I}$ is a 3x3 unit matrix, $\underset{\sim}{Q}$ is $3 \times (N+2)(2N-1)$ matrix of

zeros, $\underset{\sim}{R}$ is a $(N+2)(2N-1) \times 3$ matrix of one-step fixation probabilities of A_1, A_2 and A_3 and $\underset{\sim}{Q}$ is a $(N+2)(2N-1) \times (N+2)(2N-1)$ matrix given by

$$(13.2) \qquad \underset{\sim}{Q} = \begin{bmatrix} \underset{\sim}{Q}_{12} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{Q}_{13} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{Q}_{23} & \underset{\sim}{0} \\ \underset{\sim}{D}_{12} & \underset{\sim}{D}_{13} & \underset{\sim}{D}_{23} & \underset{\sim}{Q}^{*} \end{bmatrix}$$

where $\underset{\sim}{Q}_{ij}$ gives the transition probabilities between the transient states possible between any two out of three alleles and $\underset{\sim}{D}_{ij}$ gives the transitions from the states with all the three alleles present to any of those with only two alleles present. $\underset{\sim}{Q}^{*}$ represents the $(N-1)(2N-1) \times (N-1)(2N-1)$ matrix of transitions between states with all the three alleles present. The nature of fundamental matrix $\underset{\sim}{T} = (\underset{\sim}{I}-\underset{\sim}{Q})^{-1}$ in such a case takes the form

$$(13.3) \qquad \underset{\sim}{T} = \begin{bmatrix} \underset{\sim}{T}_{12} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{T}_{13} & \underset{\sim}{0} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{T}_{23} & \underset{\sim}{0} \\ \underset{\sim}{T}^{*}\underset{\sim}{D}_{12}\underset{\sim}{T}_{12} & \underset{\sim}{T}^{*}\underset{\sim}{D}_{13}\underset{\sim}{T}_{13} & \underset{\sim}{T}^{*}\underset{\sim}{D}_{23}\underset{\sim}{T}_{23} & \underset{\sim}{T}^{*} \end{bmatrix}$$

where $\underset{\sim}{T}_{12} = (\underset{\sim}{I}-\underset{\sim}{Q}_{12})^{-1}$, $\underset{\sim}{T}_{13} = (\underset{\sim}{I}-\underset{\sim}{Q}_{13})^{-1}$, $\underset{\sim}{T}_{23} = (\underset{\sim}{I}-\underset{\sim}{Q}_{23})^{-1}$, $\underset{\sim}{T}^{*} = (\underset{\sim}{I}-\underset{\sim}{Q}^{*})^{-1}$.

The fixation probability vector $\underline{U}_1$ for A_1 consists of four component vectors. $\underline{U}_{12}$, $\underline{U}_{13}$, $\underline{U}_{23}$ and $\underline{U}^*$ and are obtained by operating $\underline{T}$ on the corresponding component of $\underline{R}$. The mean time until fixation of the allele A_1, disregarding the cases in which it is lost, is then given by the ratio of elements of vector $\underline{M}_1$ and $\underline{U}_1$ where $\underline{M}_1$ consists of four component vectors $\underline{M}_{12}$, $\underline{M}_{13}$, $\underline{M}_{23}$ and $\underline{M}^*$ which are $\underline{M}_{12} = \underline{T}_{12}\underline{U}_{12}$, $\underline{M}_{13} = \underline{T}_{13}\underline{U}_{13}$, $\underline{M}_{23} = \underline{0}$, $\underline{M}^* = \underline{T}^*(\underline{P}_{12}\underline{M}_{12} + \underline{P}_{13}\underline{M}_{13} + \underline{U}^*)$.

The transition probabilities are assumed to follow trinomial sampling model given by

$$(13.4) \qquad P_{(i_1 i_2)(j_1 j_2)} = \frac{\lfloor (2N)}{\lfloor j_1 \; \lfloor j_2 \; \lfloor (2N-j_1-j_2)} (q'_{i_1})^{j_1} (q'_{i_2})^{j_2} (1-q'_{i_1}-q'_{i_2})^{(2N-j_1-j_2)}$$

where q'_{i_1} and q'_{i_2} are the respective frequencies of A_1 and A_2 after one generation of selection, starting with frequencies in the previous generation as $q_{i_1} = i_1/2N$, $q_{i_2} = i_2/2N$.

If the relative selective advantages of alleles A_1, A_2 and A_3 with frequencies q_1, q_2 and $q_3 = 1 - q_1 - q_2$ are respectively $(1+s_1)$, $(1+s_2)$ and 1; the changes in the gene frequencies due

to selection can be expressed as functions of q_1, q_2, $\bar{s}$ and V_m (mean and variance in superiority of A_1 over A_2 and A_3 taken together). The numerical results on mean time to fixation of A_1 were obtained for N=6 and all possible gene frequencies. These were obtained for the three combinations of frequencies: $q_1=q_2=0.0833$; $q_1=0.0833$ and $q_2=0.4167$; $q_1=q_2=0.3333$. Here we present them graphically in Figure 13.1 a, b, c respectively showing their dependence on $N\bar{s}$ and N^2V_m as well as q_1 and q_2 as discussed in Narain (1983).

It is apparent from the figure that when both the alleles are at low frequencies, the mean time decreases with an increase in $N\bar{s}$ at all values of N^2V_m but when one or both the alleles are at intermediate or higher frequencies, this behaviour changes for values of N^2V_m beyond 1, the mean time increasing initially, attaining a maximum and then decreasing. It can be inferred from these results that the introduction of a third allele in an otherwise two allele system decreases the fixation time, on the average, provided there is variability in the effect of the third allele over its two pre-existing alleles.

N.B. The results published in this paper were earlier incorporated in the Ph.D. thesis RESPONSE TO SELECTION IN FINITE POPULATIONS submitted by me to Edinburgh University in 1969.

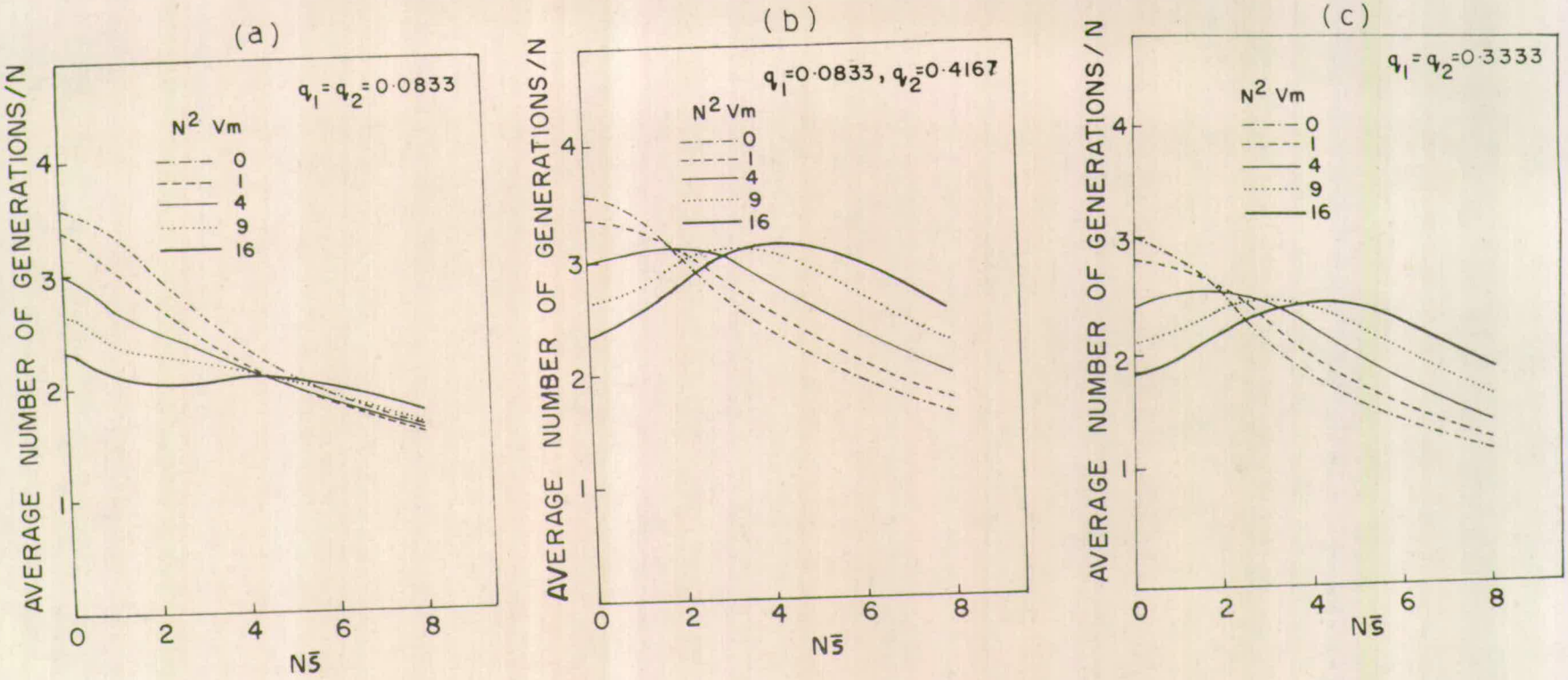

Fig. 13.1 Average number of generations required for fixation divided by population size, for different values of $N\bar{s}$, N^2V_m and initial gene frequencies.

14. Narain, P. (1979). On the statistical properties of the conditional equilibrium distribution under steady flux of mutations. Proceedings Indian National Science Academy, Series B, 45(3): 239-246.

Using infinite sites model for mutational production of new alleles, Kimura (1969) estimated heterozygosity on the basis of diffusion approach which takes into account all the sample paths of the process that lead either to the fixation or loss of the mutant from the population. Although it is not possible to know whether the mutant, at its initial occurrence, with frequency p, is going to be eventually lost or fixed, we do know that the probabilities of these two eventualities are [1-u(p)] and u(p) respectively. Using this probability, one can invoke the conditioned process in which the loss of the allele is made certain. The recurrence of new mutants is then balanced by their loss only and not by both fixation as well as loss. Such a situation would occur if the population size is very large and the selection forces are weak as is normally encountered in practice. u(p) would then be small. However, if the event of fixation does occur, the conditional expectation would be much larger than if the gene is lost. In such a case, unconditional expectation of Kimura (1969) would give too heavy a weight to sample paths that would rarely occur. These considerations suggest that it is more appropriate to consider only those sample

paths of the process that lead to the loss of the mutant from the population disregarding those in which they are fixed.

We assume that, on an average, in each generation mutant forms appear in the population in v_m nucleotide sites so that mutation rate per gamete is $v = v_m/2N$. We can then envisage a conditional stable distribution of the mutant frequencies in different sites considering only those sites in which mutants are not lost. Since v_m is the number of sites in which new mutants appear in the population in each generation, $v_m f_{co}(p,x;t)dx$ represents the contribution made by mutants which appeared t generations earlier with initial frequency p to the present frequency class in which the mutant frequencies are in the range from x to x+dx. Thus, considering all the contributions made by mutants in the past, the expected number of sites in which the mutants are in frequency range x to x+dx in the present generation conditional to their loss from the population, is $\phi_{co}(p,x)dx$ where

$$(14.1) \qquad \phi_{co}(p,x) = v_m \int_0^\infty f_{co}(p,x;t)dt, \qquad (0 \leq x < 1)$$

is the conditional stable distributions under steady flux of mutations. The expectation of an arbitrary function g(x) with respect to this distribution is then given by

$$(14.2) \qquad I^g_{co}(p) = \int_0^1 g(x)\phi_{co}(p,x)dx$$

This functional is differentiable upto the second order at p and can be shown to satisfy the ordinary differential equation

$$(14.3)\qquad \frac{V_{\Delta p}}{2}\frac{d^2 I^{g}_{co}(p)}{dp^2} + \left[M_{\Delta p} + V_{\Delta p}\frac{d}{dp}\{\log(1-u(p))\}\right]\frac{dI^{g}_{co}(p)}{dp} + v_m g(p) = o$$

subject to the boundary conditions $I^{g}_{co}(o) = o$ and $\underset{p\to 1}{\text{Lim}} I^{g}_{co}(p) = \text{finite}$.

The various statistical properties of the conditional distribution can be studied by assigning various forms to $g(p)$ and solving the resulting differential equation.

Distribution function: Here $g(p) = \delta(x-p)$ where $\delta(\cdot)$ is Dirac delta function. The solution of the differential equation is given by

$$(14.4)\qquad \phi_{co}(p,x) = 2v_m u(p)[1-u(x)]\Big/[1-u(p)]V_{\Delta x}G(x), \quad p \le x < 1$$

$$= 2v_m u(x)[1-u(x)]\Big/V_{\Delta x}G(x), \quad o < x \le p$$

where $G(x) = du(x)/dx$. In particular, for pure random drift, $M_{\Delta x} = o$, $V_{\Delta x} = x(1-x)/2N_e$, $u(p) = p$, $G(p) = 1$, giving

$$(14.5)\qquad \phi_{co}(p,x) = \frac{4N_e v_m p(1-x)}{x(1-p)}, \quad p \le x < 1$$

$$= 4N_e v_m, \quad o < x \le p$$

Total number of segregating sites in the population: In this case, $g(p) = 1$ and if we assume that the mutant is neutral, we get

$$(14.6) \qquad I_{co}(p) = -4N_e v_m \left(\frac{p}{1-p}\right) \log_e p$$

For $p = 1/2N$ and $(N_e/N) = 0.5$, we get approximately

$$(14.7) \qquad I_{co}(1/2N) \approx 4N_e v \log_e 4N_e$$

In the unconditional case, Kimura (1969) correspondingly obtains,

$$(14.8) \qquad I(1/2N) \approx 4N_e v[1 + \log_e 4N_e]$$

A comparative picture of the total number of segregating sites in the two cases for varying N_e with $v = 2 \times 10^{-6}$ is shown in Figure 14.1(a).

Average number of heterozygous sites per individual: Here, we take $g(x) = 2x(1-x)$ and assume neutral mutants, giving

$$(14.9) \qquad H_{co}(p) = (4/3)N_e v_m p(2-p)$$

When $p = 1/2N$ and $(N_e/N) = 0.5$, this gives approximately

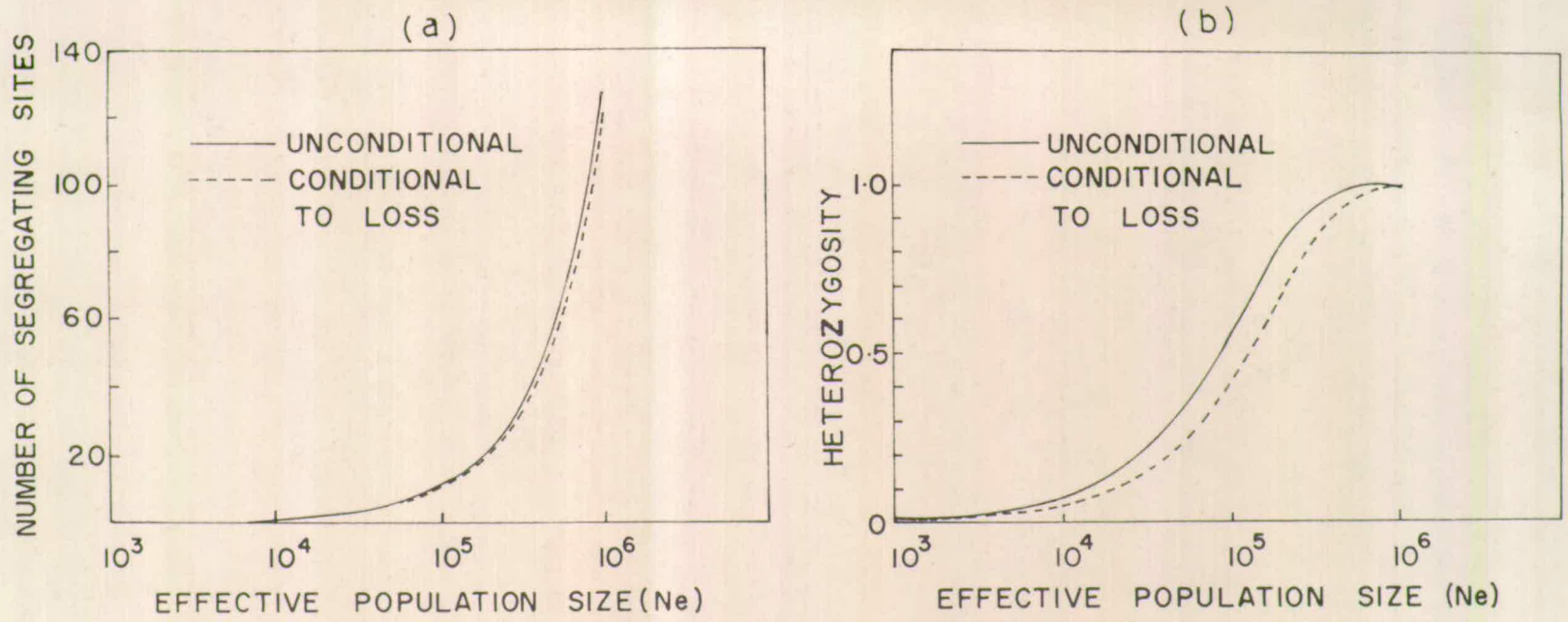

Fig. 14.1 Statistical properties of equilibrium distribution under steady flux of mutations for different values of effective population size.

(14.10) $H_{co}(1/2N) \simeq (8/3)N_e v$

Heterozygosity: The probability that a cistron (gene) is heterozygous at one or more sites, usually referred to as heterozygosity (h) is connected with the average number of heterozygous segregating sites per individual by the approximate relation

(14.11) $h(p) \simeq 1 - \exp[-H(p)]$

For the conditional neutral case and for $p = 1/2N$, $(N_e/N) = 0.5$, we then have

(14.12) $h_{co} \simeq 1 - \exp[-(8/3)N_e v]$

whereas correspondingly for unconditional case,

(14.13) $h \simeq 1 - \exp[-4N_e v]$

A comparative picture of heterozygosity in the two cases with $v = 2 \times 10^{-6}$ and for varying N_e is shown in Figure 14.1(b).

For intermediate effective population size of the order of 10^5, the heterozygosity in the conditional case could be substantially smaller. For instance, if we assume that molecular

neutral mutants occur at the rate of 2 per gamete per generation so that $v_m=4N=8N_e$, the heterozygosity in a population of effective size 10^5 would be about 12 per cent, much nearer to the reported estimate of 10 per cent as against 18 per cent obtained by the unconditional approach of Kimura (1969).

Substitutional Load: The amount of selective elimination that accompanies the process of substituting one allele for another by natural selection is known as substitutional load (Kimura and Maruyama, 1969). In the additive selection scheme, the mean fitness of the population is less by $s(1-x)$ as compared to the fitness of the optimum genotype so that the load in this population is $s(1-x)$. The expected value of the sum total of the load from time $t=o$ to time $t=\infty$, in the conditional case, denoted by $L_{co}(p)$ can therefore, be obtained by taking $g(x)=s(1-x)$ and solving the differential equation with appropriate boundary conditions. For the situation, $p=1/2N$ and an advantageous mutant such that $2N_es \gg 1$ but $(N_e/N)s \ll 1$, we get

$$(14.14) \qquad L_o(1/2N) \approx 4(N_es)v[1-\gamma-\log_e(N_e/N)s]$$

where $\gamma = 0.57721...$ is Euler's constant. If both $2N_es$ as

well as $(N_e/N)s$ are much smaller than unity, we have

$$(14.15) \qquad L_o(1/2N) \approx 2v[2(1-N_e s) - (\gamma + \log_e 2N_e s)/N_e s]$$

If we take $v_m=1$ with $N=2\times10^5$, $(N_e/N) = 0.5$ and $s=0.01$, we use the first formula to get $L_o(1/2N) = 0.0572$, whereas if $s=10^{-6}$, $N_e=10^5$ so that $N_e s=0.1$ and $(N_e/N)s = 0.5\times10^{-6}$, we use the second formula to get $L_o(1/2N) = 15\times10^{-6}$. Such considerably smaller substitutional loads for the case when the favoured mutant is destined to be lost from the population, indicate that there may not be any limit to the rate of gene substitution.

III. GENETICS OF QUANTITATIVE VARIABILITY

The description of gene control in characters showing continuous variation, first considered by Fisher (1918) and later elaborated in the form of a 'Biometrical Method' by Mather (1949) is usually in terms of a set of loci each having only two alleles at a locus. If one is dealing with a F_2 derived from two homozygous strains, this assumption is quite justified as only two alleles are then possible at each locus. But when several F_2 populations are considered or when one is dealing with a randomly breeding population, it is necessary to consider more than two alleles at each locus. However, if we adopt the technique of representing the three genotypes A_1A_1, A_1A_2 and A_2A_2, possible with two alleles A_1 and A_2, by a stochastic variable ϕ taking values -1, o and 1 respectively so that the heterozygote is located mid-way between the two homozygotes, we cannot do so with three or more alleles at a locus. As we would presently see this difficulty is overcome when, as in Narain (15), we represent the six genotypes possible with three alleles by a two-dimensional stochastic vector $\phi = (x,y)$ such that the homozygotes are placed at the vertices of an equilateral triangle and the heterozygotes are at the mid-point of the edges in a two-dimensional plane. This gives the effect of considering multiple alleles in continuous variation.

In studies on inheritance of quantitative characters, the 'heritability' of a character is often expressed as a regression coefficient of the breeding value of the individual on its phenotypic value for the given character where breeding value is a part of the phenotypic value. This concept helps in predicting the breeding value of the individual from the knowledge of heritability and phenotypic value. When extended to genetic selection programmes, where individuals with phenotypic values greater than a certain specified value are selected for breeding the next generation, it helps in predicting the response to selection. When several correlated characters are considered simultaneously, it is possible to generalise these concepts as we would see in Narain (16).

The genetic improvement in economic characters showing continuous variation depends upon the intensity with which the selection is applied as well as on the accuracy of the selection as measured by the correlation between the breeding value of the individual and the criterion of selection. Generally, there is a conflict between these two factors as, due to limited resources and limited size of family, increasing the accuracy of selection results in decreasing the intensity of selection. For instance, in a progeny testing programme in dairy cattle, the resources available to a breeder limit the total number (N) of daughters to be milk recorded so that the number (B) of young

bulls which can be tested each year becomes smaller if we increase the number (n) of daughters per bull, thereby increasing the accuracy of the progeny test. Since the number (S) of best tested bulls to be added to the sted each year for use in the artificial insemination is limited due to a tolerable amount of inbreeding, the proportion of sires to be selected increases due to smaller number of young bulls under test, thereby decreasing the intensity of selection. In such a case, an optimum strategy such that the expected genetic superiority of the sire is maximised for fixed resources was given by Robertson (1957). On the other hand, if a breeder is interested in minimising the cost of running a progeny testing programme for a given rate of genetic improvement, the optimum strategy would be according to a general procedure as would be discussed in Narain (17). Further, breeding strategies could also be devised which maximises the total returns on expenditure particularly when such returns tend to accumulate over long period of time as in the progeny tests. In Narain (18) we discuss how to determine the optimum intensity of selection in a progeny testing programme by taking the costs as well as returns into account.

The rate of genetic improvement in a quantitative trait due to selective breeding on the basis of its phenotypic values can be increased if the variation in this trait due to auxiliary

traits, particularly at the environmental level, are minimised as far as possible. In such a case selection is to be made on the basis of an index expressed as deviation of the phenotypic value of the trait from its expected value predicted with the help of the auxiliary traits. Such an index may be called a phenotypic index to distinguish it from the selection index introduced by Hazel (1943). The advantage of phenotypic index is that unlike the case of selection index, a knowledge of the estimates of genetic parameters is not necessary for its construction. As such it is easier to adopt. This technique was exploited by Rendel (1954), Osborne (1957), Purser (1960) and Searle (1965) but considering only one auxiliary trait. As we would shortly see, Narain and Mishra (19) developed a general procedure with several auxiliary traits and examined the conditions under which selection based on phenotypic index is more efficient than that without using the auxiliary traits.

Dairy sires are often selected on the basis of daughter-dam comparison using a corrected daughter average index which is based on correcting the daughters' average on the basis of the regression of daughters' performance on those of dams' for the unequal production levels of the dams mated to different sires. This regression, as we know, measures half the heritability of the character under consideration. In Narain (20) we develop a theory to further correct the daughters' average for the given

character for variations in another correlated auxiliary character using the technique of phenotypic index described in Narain and Mishra (19).

The genetic superiority of individuals for selection can be determined either on the basis of their own performance or the performance of their relatives such as full-sibs and half-sibs or else an optimum combination of several such information, all for the same trait. The latter is known as combined selection, involves construction of a selection index and is mostly used in poultry breeding. Narain et al (21) developed, as we would see presently, new selection indices which combine information on the individual performance for several auxiliary traits with the information on the trait under improvement for the individual as well as its relatives, such as full-sibs, half-sibs and dams. The necessity for developing theory for combined selection with auxiliary traits arose during the course of the operation of an Indian poultry breeding programme aimed at evolving suitable strains of egg type chickens. The selection criterion was the index developed by Osborne (1957) based on the character rate of lay. It was observed that while the average rate of lay improved, the average egg weight deteriorated due to correlated response to selection for rate of lay on the basis of the index. This raised the statistical problem of devising an index which includes the information on

the individual performance for an auxiliary trait such as egg weight in addition to combining the information about the main trait (rate of lay) on the individual bird with those of its full-sibs and half-sibs. A general theory for such indices with k auxiliary traits is therefore developed in this paper.

Plant Breeders often adapt diallel crossing technique in which all possible single crosses among a group of inbred lines are raised. When reciprocal crosses and parental inbreds are not included, there are $N = n(n-1)/2$ possible single crosses among a set of n lines which are tested in a suitably replicated randomised design. This number increases rapidly with increase in n. With facilities available for testing only a limited number of crosses, a diallel cross may therefore be possible only when n is relatively small. However, if only a small number of lines are included, the estimates of the variances of the general combining ability (g.c.a.) and specific combining ability (s.c.a.) among the whole population of potentially available lines are subject to large sampling errors and many potentially high yielding lines may be left completely untested. It is, therefore, necessary to have a large number of inbred lines but raise only a sample of all possible crosses among them. Such a diallel cross is known as <u>partial diallel cross</u> (PDC). Gilbert (1958), Kempthorne and Curnow (1961) and Curnow (1963) discussed problems of construction and analysis of PDCs. For

enabling the plant breeders to make use of PDC in their experimentation, it is necessary, in addition to giving them the method of construction and analysis, to indicate which of the several possible designs of PDC, in a given situation, is most efficient in the sense that it gives the least average variance of the difference between the g.c.a. effects of a pair of lines. Of the different ways of approaching this problem, in one the structure of a partially balanced incomplete block (PBIB) design is used in constructing the PDC. This is because there is a one-to-one correspondence between the PDC and partially balanced incomplete block (PBIB) design with 2-plot per block and two associate classes or three associate classes or in general 'm' associate classes with parameters n, r=s, b=ns/2, k=2, $\alpha_1, \alpha_2, \ldots, \alpha_m$ where α's take values either zero or one. In such a case, we have different variances for different comparisons. In particular, we may have balance over only a set of comparisons thus resulting in two-, three- or four-, variance samples based on PBIB designs. The efficiency of the PDC then depends on the average of the variances over all the comparisons, which can be expressed per unit error variance (σ^2). As we would see, Narain et al (22) constructed and analysed PDC when n is of the form p(p-1)(p-2)/6 where p is an integer greater than 3, using extended triangular (ET) scheme with three associate classes. Further, in Narain and Arya (23) we discuss a new association

scheme called truncated triangular (TT) scheme with five associate classes when n=p(p-2)/2 with p an even positive integer greater than or equal to 8 and use it to construct and analyse PDC.

15. Narain, P. (1965). The description of gene action and interaction with multiple alleles in continuous variation. Genetics, 52: 43-53.

Consider a panmictic population containing three alleles at a given locus. The six genotypes A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3 and A_1A_3 can be represented by a vector variable (X_1, X_2, X_3) where X_i is 0, 1 or 2 and denotes the number of A_i genes in the genotype. All these vectors lie on the plane $X_1+X_2+X_3=2$. Taking the origin at (1,1,0) and making all the vectors lie on the plane $X_2=0$, gives the two dimensional vector variables (1,0), (0,0), (-1,0), (-1,1), (-1,2) and (0,1) which represent respectively the six genotypes A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3 and A_1A_3. An individual can then be represented by $\phi=(x,y)$ which takes any one of the six pairs of values depending upon its genotype. It can be seen that the probabilities for an individual $\phi=(x,y)$ to produce gametes A_1, A_2 and A_3 are respectively $(1+x)/2$, $(1-z)/2$ and $y/2$, where $z=x+y$. When two individuals with genotypes $\phi_1=(x_1,y_1)$ and $\phi_2=(x_2,y_2)$ are crossed, the distribution of ϕ in the offspring can be obtained

by probability arguments. For instance, ϕ would take value (o,o) i.e. the individual would be A_1A_2 with probability $\frac{1}{4}[(1+x_1)(1-z_2)+(1+x_2)(1-z_1]$ where $z_1=x_1+y_1$, $z_2=x_2+y_2$. Such a distribution gives a mathematical representation of Mendel's law of segregation with three alleles and is generalised for an arbitrary number of alleles at a locus as per details in the paper. When the two individuals ϕ_1 and ϕ_2 happen to be members of a random mating population in equilibrium with frequencies p, q and r respectively for A_1, A_2 and A_3, it can be seen that $E(\phi_1) = E(\phi_2)$ i.e. $E(x_1) = E(x_2) = p-q-r$ and $E(y_1) = E(y_2) = 2r$ and in the next generation, the frequencies of six genotypes are found to be the same as in the previous generation thus confirming the Hardy-Weinberg law of equilibrium for random mating.

Now consider k loci, A,B,C,...,K each with three alleles controlling the quantitative character. The whole genotype can be represented by a set of k two-dimensional stochastic variable $G = (\phi_a, \phi_b, \ldots, \phi_k)$. We assume the absence of linkage so that $Cov(x_a,x_b) = Cov(x_a,y_b) = Cov(x_b, y_a) = Cov(x_b,y_b) = o$ etc. The relationship between the metric and the genotype is set up by taking a suitable polynomial function in x_a and y_b and summed over the loci given by

$$(15.1) \qquad \sum_a M(x_a,y_a) = \sum_a [\alpha_{1a}x_a+\alpha_{2a}y_a+\alpha_{3a}x_a^2+\alpha_{4a}y_a^2+\alpha_{5a}x_ay_a]$$

where non-allelic interactions are assumed to be absent and the values of α's are given by

$$
\begin{aligned}
\alpha_{1a} &= d_{1a} \\
\alpha_{2a} &= d_{2a}-h_{1a}+h_{2a}+h_{3a} \\
\alpha_{3a} &= -h_{1a} \\
\alpha_{4a} &= -h_{2a} \\
\alpha_{5a} &= -h_{1a}-h_{2a}+h_{3a}
\end{aligned}
\qquad (15.2)
$$

For each locus, therefore, five parameters are required to describe the metric completely. While the main effect d_{1a} measures the additive effect of A_1 relative to A_2, d_{2a} measures that of A_2 relative to A_3 and $(d_{1a}-d_{2a})$ therefore measures that of A_3 relative to A_1. The dominance deviations are then h_{1a}, h_{2a} and h_{3a} for the pairs A_1-A_2, A_2-A_3 and A_3-A_1 respectively.

With the above specification of the metric for the character, it is possible to derive various first and second degree statistics in the offspring of a cross $G_1 x G_2$ by taking the required expectations and summing over loci. The general results so obtained could be applied to any system of mating such as selfing or random mating. In particular, for a randomly breeding population in equilibrium, the metric values

would be functions of the 5k genetic parameters, gene frequencies p_a, q_a, r_a for each of the k loci and the genotypic average of the population given by

$$(15.3) \qquad \sum_a d_a = \sum_a (p_a - q_a - r_a) d_{1a} + 2\sum_a r_a d_{2a} - \sum_a (1 - 2p_a q_a) h_{1a} + 2\sum_a q_a r_a h_{2a} + 2\sum_a p_a r_a h_{3a}$$

Consider the situation in which we take A_2 as fixed and A_1 and A_3 with frequencies p_a and r_a are considered with respect to this allele at each of the k loci. Let

$$(15.4) \qquad d_{ia} = m_a + \alpha_{ia}, \qquad i=1,2$$

$$(15.5) \qquad h_{ia} = h_a + \beta_{ia}, \qquad i=1,2$$

with

$$(15.6) \qquad m_a = (p_a d_{1a} + r_a d_{2a})/(p_a + r_a)$$

$$(15.7) \qquad h_a = (p_a h_{1a} + r_a h_{2a})/(p_a + r_a)$$

at each of the k loci. Then $p_a\alpha_{1a} + r_a\alpha_{2a} = 0$, $p_a\beta_{1a} + r_a\beta_{2a} = 0$ at each of the locus. We then obtain the total genetic variance as

$$(15.8) \qquad {}_H V_R = \tfrac{1}{2} D_A + \tfrac{1}{4} H_A + \tfrac{1}{2}(D_B + H_B + H_C)$$

where

$$(15.9)\qquad D_A = \sum_a [4q_a(1-q_a)\{m_a+(2q_a-1)h_a+\frac{2r_a p_a}{1-q_a}h_{3a}\}^2]$$

$$(15.10)\qquad H_A = \sum_a [16q_a^2(1-q_a)^2\{h_a-\frac{r_a p_a}{(1-q_a)^2}h_{3a}\}^2]$$

$$(15.11)\qquad D_B = \sum_a [4p_a\{\alpha_{1a}+q_a\beta_{1a}+r_a h_{3a}\}^2+4r_a\{\alpha_{2a}+q_a\beta_{2a}+p_a h_{3a}\}^2]$$

$$(15.12)\qquad H_B = \sum_a [4q_a(1-q_a)\{p_a(\beta_{1a}-\frac{r_a h_{3a}}{1-q_a})^2+r_a(\beta_{2a}-\frac{p_a h_{3a}}{1-q_a})^2\}]$$

$$(15.13)\qquad H_C = \sum_a \frac{4r_a^2 p_a^2(1-3q_a)}{(1-q_a)^2}h_{3a}^2$$

We are thus able to partition the total genetic variance into three components viz. additive genetic component D_A, dominance component H_A and multiple allelic component $M_A = D_B+H_B+H_C$. The last component is a new component and is existant only when $d_{1a}\neq d_{2a}$, $h_{1a}\neq h_{2a}$ and $h_{3a}\neq o$ at each locus. When $d_{1a}=d_{2a}$ at each locus, the third main effect is zero. Similarly, when $h_{1a}=h_{2a}$ and $h_{3a}=o$ at each of the loci, we have only one dominance effect at each locus. The total genetic variance reduces, in such a case, to

$$(15.14)\qquad {}_H V_R = \frac{1}{2}D_A + \frac{1}{4}H_A$$

where

$$(15.15) \qquad D_A = \sum_a 4q_a(1-q_a)\{m_a+(2q_a-1)h_a\}^2$$

$$(15.16) \qquad H_A = \sum_a 16q_a^2(1-q_a)^2 h_a^2$$

which are the same as for di-allelic loci with m_a as main effect and h_a as dominance effect at each locus. The tri-allelic case can then be regarded as a di-allelic case with allelic frequencies as q_a and (p_a+r_a) for A_2 and (A_1,A_3) at each of the k loci. Thus we are able to conclude that the simultaneous vanishing of D_B, H_B and H_C, in other words, the vanishing of M_A would indicate that multiple allelism does not contribute to the total genetic variance in a randomly breeding population.

16. Narain, P. (1979). Certain generalisations in response to genetic selection. Contributions to Statistics, S.C. Das Memorial Vol., Utkal University, Bhubaneswar (India): 71-79.

Let the phenotypic measurements $P_1,P_2,\ldots,P_k$ of an individual for k characters be represented by a kxl column vector $\underline{P}$ whereas the corresponding conceptual breeding values by a kxl column vector $\underline{A}$. The corresponding non-additive and environmental effects together may be represented in the form

of same order column vector $\underline{R}$. We then have

(16.1) $\underline{P} = \underline{A} + \underline{R}$

assuming the absence of interaction between genotype and environment. All variables are expressed as deviations from the respective means and standardised to have variances as unity and therefore correlations equal to covariance. Let Σ_P, Σ_A and Σ_R denote the symmetric dispersion matrices for variables $\underline{P}$, $\underline{A}$ and $\underline{R}$ respectively. The diagonal elements of Σ_P are unity whereas those of Σ_A are h_{ii}, the heritability for i-th character, i=1,2,...,k. The non-diagonal elements of Σ_A are $h_{ij} = r_{ij}\sqrt{h_{ii}\, h_{jj}}$, the covariance of breeding values between i and j characters and r_{ij} the corresponding genetic correlation, $i \neq j = 1,2,\ldots,k$. It is easy to see that

(16.2) $\mathrm{Cov}(\underline{A},\underline{P}) = E(\underline{A}\,\underline{P}') = \Sigma_A$

(16.3) $\Sigma_P = \Sigma_A + \Sigma_R$

Then if Σ_P is non-singular,

(16.4) $$\underline{I} = \Sigma_A \Sigma_P^{-1} + \Sigma_R \Sigma_P^{-1}$$

$$= \underline{H} + \underline{E}$$

showing that, in some sense, the elements of kxk matrix $\underline{H}$

represent the parameters corresponding to the heritability in an univariate situation.

A related concept of heritability in the univariate case is in terms of the square of the correlation coefficient between A_i and P_i. This suggests that for multi-variate case we need examine the correlation between two vector-valued variables $\underline{A}$ and $\underline{P}$. For this we set up linear relations

$$(16.5) \qquad \underline{a}'\underline{A} = \sum_{i=1}^{k} a_i A_i$$

$$(16.6) \qquad \underline{a}'\underline{P} = \sum_{i=1}^{k} a_i P_i$$

and choose the coefficients in such a manner that the correlation between them is maximum. This correlation denoted by h(a) is given by

$$(16.7) \qquad h(a) = \sqrt{\frac{\underline{a}'\Sigma_A\underline{a}}{\underline{a}'\Sigma_P\underline{a}}}$$

The maximisation leads to the equation

$$(16.8) \qquad |H - h^2(a)I| = 0$$

indicating that the desired correlations usually known as canonical correlations, are the square-roots of the eigen-

roots of matrix $\underset{\sim}{H}$. We arrange the set of these correlations, in order of their magnitude, $h(a_1) \geq h(a_2) \ldots \geq h(a_k)$ with their corresponding pairs of canonical variates $(\underline{a}_1'\underline{A}, \underline{a}_1'\underline{P})$, $(\underline{a}_2'\underline{A}, \underline{a}_2'\underline{P}), \ldots, (\underline{a}_k'\underline{A}, \underline{a}_k'\underline{P})$. We can then take the two linear functions $\underline{a}_1'\underline{A}$ and $\underline{a}_1'\underline{P}$ with maximum correlation $h(a_1)$, as the breeding value and phenotypic value respectively for the character compounding the set of k characters. The square of this correlation $h^2(a_1)$ expresses the fraction of the phenotypic variance of the compound character which is due to additive genetic effects. It can therefore legitimately be regarded as a generalised concept of heritability. In other words, the largest eigen-root of $\underset{\sim}{H}$ can be termed as <u>generalised heritability</u>. It can be seen that for k=1, this reduces to the usual heritability h^2 but for k=2, we get

$$(16.9) \qquad h^2(a_1) = \frac{1}{2(1-\rho_{12}^2)}\left[(h_{11}+h_{22}-2\rho_{12}h_{12}) \pm \sqrt{(h_{11}-h_{22})^2 + 4(h_{12}-h_{11}\rho_{12})(h_{12}-h_{22}\rho_{12})}\right]$$

where ρ_{12} is the phenotypic correlation between the two characters. When it is zero, we get

$$(16.10) \qquad h^2(a_1) = \tfrac{1}{2}\left[(h_{11}+h_{22}) \pm \sqrt{(h_{11}-h_{12})^2 + 4h_{12}^2}\right]$$

When h_{12} is also zero, there being no genetic correlation also,

$h^2(a_1)$ reduces to h_{11} or h_{22} depending upon the sign of square-root. It is also interesting to note that when $h_{12}=h_{11}\rho_{12}$ i.e. $r_{12} = \rho_{12}\sqrt{\frac{h_{11}}{h_{22}}}$, $h^2(a_1)$ again reduces to h_{11} if we take positive sign of the square-root but with negative sign becomes

$$(16.11) \qquad h^2(a_1) = (h_{22}-h_{11}\rho_{12}^2)/(1-\rho_{12}^2)$$

The method of independent culling levels adopted when the criterion of selection is based on more than one character was discussed by Young and Weiler (1960) as well as by Finney (1962). In this paper, we have generalised it for the case of k characters and demonstrated how the expected response to selection, in such a case, reveals the multivariate analogue of the univariate formula.

17. Narain, P. (1978). Optimum designs for progeny testing with minimum costs. Indian Journal of Animal Science, 48(10): 708-711.

We assume that the size of the breeding unit i.e. the total number of milk recorded cows (N) is sufficiently large to ensure the efficiency of progeny testing over the use of sires selected on the basis of their dams' performance. We further consider only the genetic gain due to selection among

tested bulls (path of genetic improvement from sires to breed young bulls) as about 40 to 60 per cent of total genetic improvement is realised through this path. To take into account the costs involved in running the programme, we consider two parameters; firstly the cost (C_1) involved in securing a daughter having at least first lactation milk record and secondly the cost (C_2) involved in maintaining a bull till he is progeny tested. We then have a total of five parameters, N, B, S, C_1 and C_2 but, in effect, the optimisation leads to certain combination of parameters resulting in a fewer number.

If we define a cost ratio $r = C_1/C_2$, the total cost (C) is seen to be directly proportional to r but inversely proportional to $p = S/B$, the proportion of bulls selected i.e.

$$(17.1) \qquad C^* = Kr + 1/p$$

where C^* is $C/(C_2S)$ with S fixed in advance and $K = N/S$. If we denote the selection intensity in standard deviation units as i and the correlation between the breeding value of sire (A) and progeny average (I) as r_{AI}, the expected genetic superiority of the selected sires in units of additive genetic standard deviation ($\Delta G_S/\sigma_g$) is given by

$$(17.2) \qquad W = ir_{AI}$$

A fixed rate of genetic improvement (W) can then be obtained by various choices of i and r_{AI}. The optimum strategy would then be to determine the group size n = N/B such that C is minimised for a fixed value of W. If we assume that we are selecting from a large sample of sires and breeding values are normally distributed, i = z/p, where z is the ordinate of the normal curve at the point where the area cut off towards right is p and

$$(17.3) \qquad r_{AI} = \sqrt{p/(p+a/K)}$$

where $a = (4-h^2)/h^2$. The optimisation problem boils down to finding pairs of values of p and K which minimises C for fixed values of W and r. This gives

$$(17.4) \qquad ar = (i^2-W^2)^2 / W^2(i^2-2ix+W^2)$$

where x = dz/dp. The optimum p thus depends on W and 'ar' with restrictions on p as

$$(17.5) \qquad i \geq W \text{ and } i^2-2ix+W^2 > 0$$

to ensure a positive value for 'ar'. This dependence is shown in Figure 17.1. The optimum p decreases as 'ar' increases at a fixed value of W but increases as W decreases at a fixed

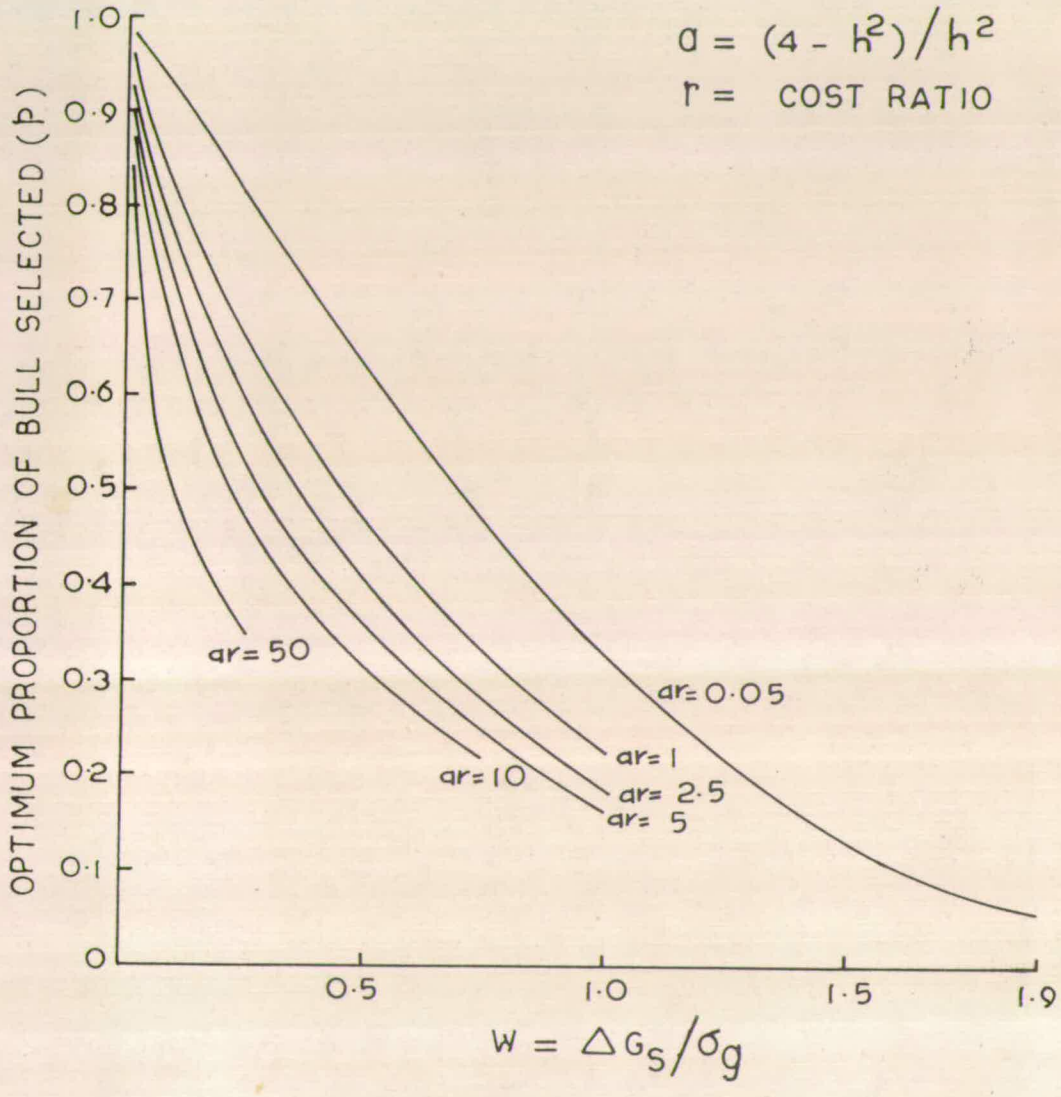

Fig. 17.1 Optimum proportion of bulls selected at given rate of genetic improvement for different values of ar.

value of 'ar'. The dependence of optimum value of (K/a) on W and ar is shown in Figure 17.2. It also decreases as 'ar' increases at a fixed value of W but also decreases as W decreases at a fixed value of 'ar'. Similarly, the dependence of the minimum cost on W and ar is shown in Figure 17.3. The minimised cost increases as 'ar' increases at a fixed value of W. It also increases as W increases at a fixed value of 'ar'.

The optimum value of n, when expressed in units of a, is found to be

$$(17.6) \qquad (n/a) = W^2/(i^2-W^2)$$

Such optimum values are presented in Table 17.1.

TABLE 17.1 Optimum values of (n/a) for different values of W and 'ar'.

W	ar							
	0.1	0.5	1.0	2.5	5.0	10.0	20.0	50.0
0.10	0.44	0.18	0.12	0.07	0.05	0.04	0.03	0.02
0.15	0.60	0.25	0.17	0.11	0.08	0.06	0.04	0.03
0.20	0.74	0.33	0.25	0.14	0.10	0.08	0.06	0.04
0.25	0.86	0.39	0.26	0.18	0.13	0.10	0.08	0.06
0.50	1.57	0.73	0.54	0.37	0.30	0.25	0.21	*
0.75	2.29	1.11	0.86	0.63	0.52	0.45	0.40	*
1.00	2.89	1.55	1.26	0.92	0.75	*	*	*
2.00	6.01	3.61	*	*	*	*	*	*

* Optimum values cannot be determined as the corresponding values of W are not admissible.

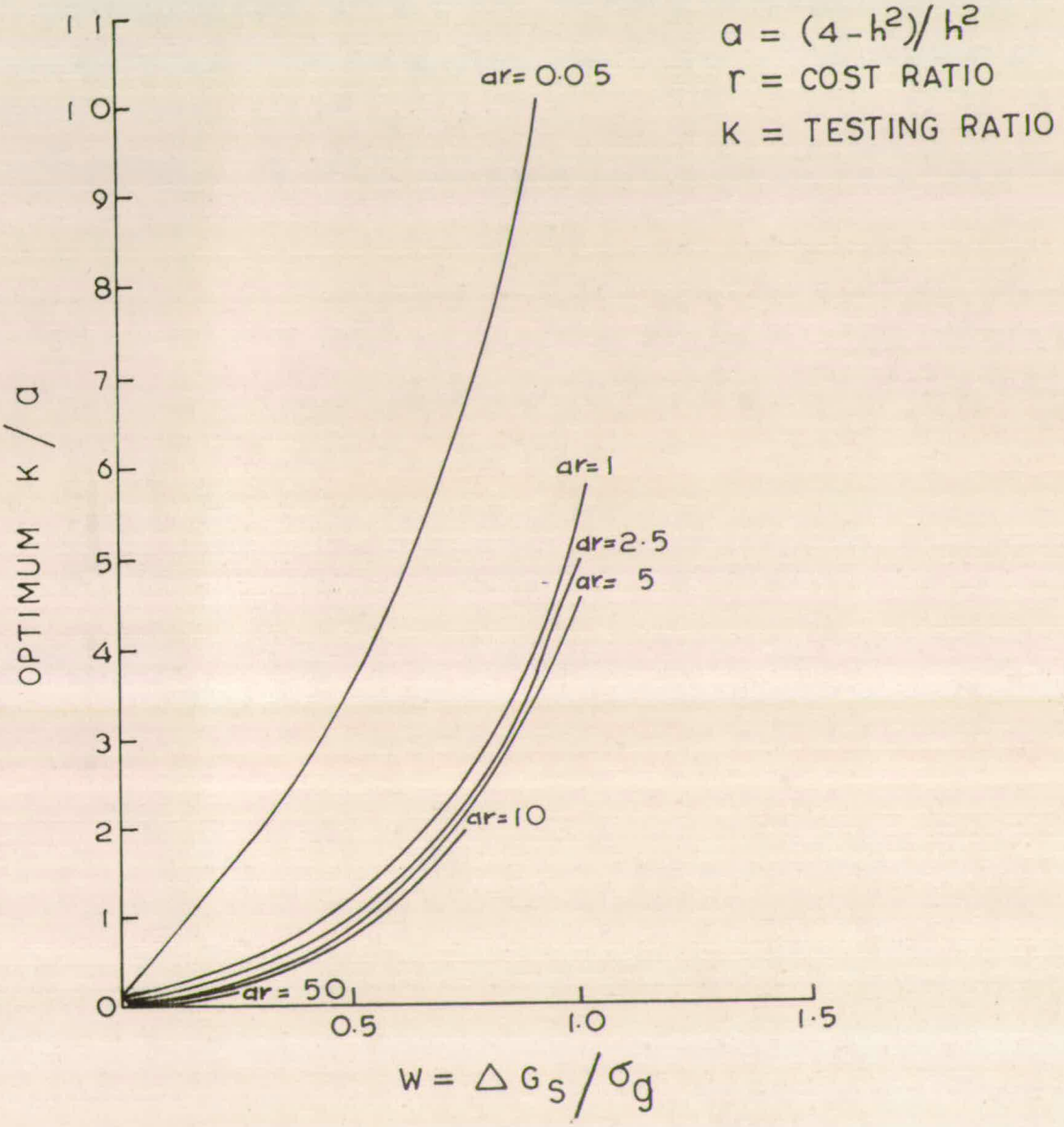

Fig. 17.2 Optimum K/a at given rate of genetic improvement for different values of ar.

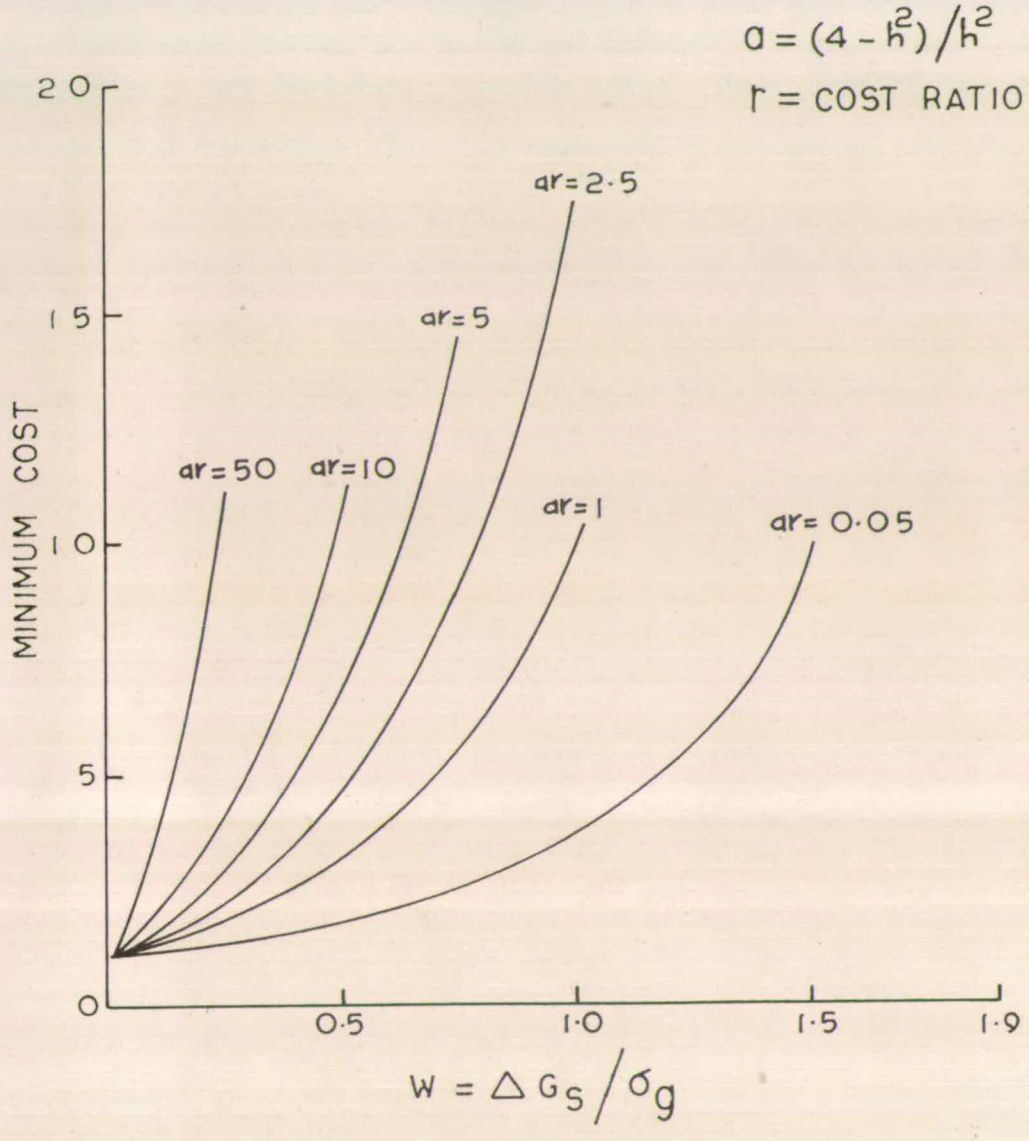

Fig. 17.3 Minimum cost of progeny testing at given rate of genetic improvement for different values of ar.

Thus, for instance, if the heritability is 0.25 giving a = 15 and the cost ratio r is about 1/3, we get the optimum n by multiplying the values in the table under the column for ar = 5 by 15 for different values of W. For W = 1.0, we find that about 11 daughters per sire are required.

18. Narain, P. (1983). Statistical aspects in optimising livestock breeding programmes. Proceedings of the 44th Session of the International Statistical Institute, Vol.2 (Contributed Papers): 833-836.

Consider a progeny testing programme in dairy herds where the pattern of breeding in successive rounds, consists of dividing the total number (N) of female population into two groups, one consisting of elite cows, (1-P) fraction of the total, to be mated to a given number (S) of proven bulls to secure future young males and replacement cows and the other, the remaining female population to be mated to a certain number (B) of young bulls which are sons of the best proven bulls of an earlier set, for testing. In such cases, we imply considering two paths of genetic improvement, from sire to sons and from sire to daughters which together constitute as much as 60 to 70 per cent of total expected genetic improvement. The optimisation problem is then to choose the number of daughters (n) per bull under test such that the profit (π) accruing from

the programme in terms of the present value of all future returns is maximised. Since $n = pK$ with $p = S/B$ and $K = N/S$, the problem boils down to optimise p for given values of K. The profit function is found to be of the form

$$(18.1) \qquad \pi = \alpha(z/p)\phi - \beta(C_o + 1/p)$$

when C_o is the component of total cost which is independent of p, ϕ is r_{AI} and

$$(18.2) \qquad \alpha = 100(2-P)M\sigma_A v/2mR(1+R)^{y-2}$$

$$(18.3) \qquad \beta = (1+R)/R$$

where M is the number of lactations in which the response is expressed, depending on the replacement rate and average herd life, σ_A is the additive genetic standard deviation, v is the monetary return for 1 per cent increase in the average milk production (m), y is the number of generations before returns start accruing and R is the interest rate per generation. For maximum profit, we have to solve a cubic in ϕ given by

$$(18.4) \qquad z\phi^3 - (2px - z)\phi - 2\beta/\alpha = o$$

This gives three roots for ϕ of which the positive real root

is taken as the solution from which optimum value of p, for a given value of K/a, is numerically obtained. These are shown in Figure 18.1 where (K/a) is taken on a log scale. The curve obtained in Robertson (1957) is also shown for the sake of comparison. It is seen from the figure that for a given value of $(2\beta/\alpha)$, the optimum p decreases as (K/a) increases. As this value tends to zero, we approach, as expected, the curve obtained by Robertson (1957) where the cost of the scheme is not taken into account. The limiting values of p as (K/a) tends to zero for different values of $(2\beta/\alpha)$ are shown graphically in Figure 18.2. As $(2\beta/\alpha)$ tends to be high, the optimum intensity of selection between tested sires tends to be lower than that obtained in Robertson (1957) for optimum running of the scheme.

19. Narain, P. and Mishra, A.K. (1975). Efficiency of selective breeding based on a phenotypic index. Journal of Genetics, 62(2): 69-76.

Consider n auxiliary traits $x_k (k=1,2,\ldots,n)$ related to the main trait y with phenotypic and breeding values, expressed as deviations from the respective means, respectively denoted by $P(x_k)$, P(y) and $A(x_k)$, A(y). Also, let the phenotypic values be standardised to have unit variances so that heritabilities $h^2(x_k)$ and $h^2(y)$ are the respective genetic variances. Consider

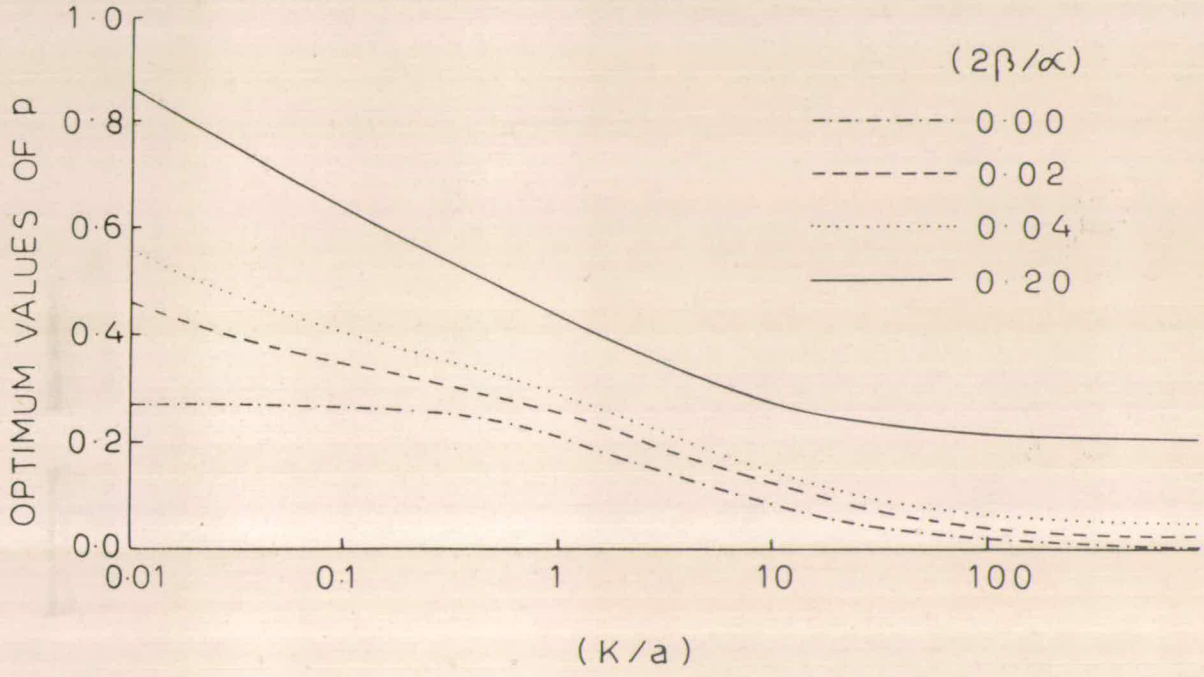

Fig. 18.1 Optimum proportion of bulls selected for different values of K/a and $2\beta/\alpha$.

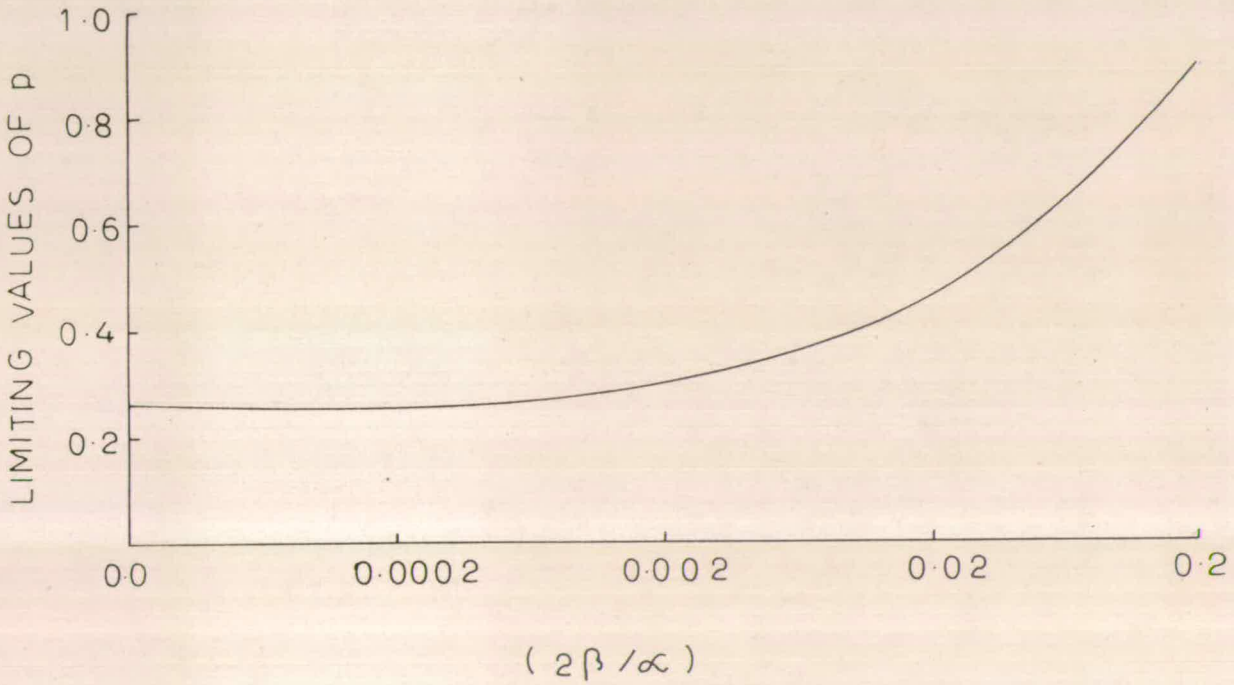

Fig. 18.2 Limiting proportion of bulls selected for optimum running of progeny testing scheme corresponding to different values of $2\beta/\alpha$.

now the selection based on the phenotypic index

$$(19.1) \qquad I_P = P(y) - \sum_{k=1}^{n} b_k P(x_k)$$

where b_k is partial regression coefficient of $P(y)$ on $P(x_k)$, $k=1,2,\ldots,n$. The expected genetic improvement in y is then

$$(19.2) \qquad \Delta G = i b_{A(y)I_P} \sigma(I_P)$$

where i is the intensity of selection, $b_{A(y)I_P}$ is regression coefficient of $A(y)$ on I_P and $\sigma(I_P)$ is the phenotypic standard deviation of the phenotypic index. For estimating $b_{A(y)I_P}$, we set up the relationship

$$(19.3) \qquad E[A(y)] = a_o P(y) + \sum_{k=1}^{n} a_k P(x_k)$$

and solve the resulting normal equations. We get

$$(19.4) \qquad b_{A(y)I_P} = a_o = h^2(y)\,(1-\underline{R}_o' \underset{\sim}{R}^{-1} \underline{C})/(1-\underline{R}_o' \underset{\sim}{R}^{-1} \underline{R}_o)$$

where $\underline{R}_o' = (R_{o1}, R_{o2}, \ldots, R_{on})$, R_{ok} being the phenotypic correlation between $P(y)$ and $P(x_k)$, $\underset{\sim}{R}$ is $n \times n$ correlation matrix of phenotypic correlation coefficients R_{kl} between the auxiliary traits and $\underline{C}' = (C_1, C_2, \ldots, C_n)$ C_k being $r_{ok} h(x_k)/h(y)$, r_{ok} being

the genetic correlation coefficient between y and x_k. Since $\sigma^2(I_p) = (1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R}_o)$, the efficiency of selection on the basis of phenotypic index relative to individual selection with the same intensity of selection is given by

$$(19.5) \qquad E_P = (1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{C})/(1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R}_o)^{1/2}$$

However, if we consider selection, with the same intensity, made on the basis of the usual selection index given by

$$(19.6) \qquad I_S = P(y) - \sum_{k=1}^{n} w_k P(x_k)$$

we have to choose optimum values of $\underset{\sim}{w}' = (w_1, w_2, \ldots, w_n)$ by maximising E_S given by

$$(19.7) \qquad E_S = (1-\underset{\sim}{w}'\underset{\sim}{C})/(1-2\underset{\sim}{w}'\underset{\sim}{R}_o+\underset{\sim}{w}'\underset{\sim}{R}\underset{\sim}{w})^{1/2}$$

It is found that E_S is maximum when

$$(19.8) \qquad \underset{\sim}{w} = \underset{\sim}{R}^{-1}(\underset{\sim}{R}_o-\underset{\sim}{C}K)$$

where

$$(19.9) \qquad K = (1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R}_o)/(1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{C})$$

When the genetic correlations r_{ok}'s are all zero, $\underset{\sim}{w}$ reduces to

$\underline{b}' = (b_1, b_2, \ldots, b_n)$ and I_S reduces to I_P. This means when all the auxiliary traits are related to the main trait only at the environmental level, the phenotypic index is optimal with efficiency

$$(19.10) \qquad E_S = E_P = 1/(1-\underline{R}_o' \underset{\sim}{R}^{-1} \underline{R}_o)^{1/2}$$

Even if all the r_{ok}'s are not zero, the phenotypic index could be used though its efficiency would then be less than maximal since, in general, we find

$$(19.11) \qquad E_P^2 = E_S^2 - \underline{C}' \underset{\sim}{R}^{-1} \underline{C}$$

However, it may happen that under certain conditions, E_P itself is greater than unity in which case it would be preferable to use E_P because of its simplicity in adoption. It is found that E_P is always more than unity whenever the vectors $\underline{R}_o$ and $\underline{C}$ are having elements with opposite signs. In other words, phenotypic and genetic correlations between y and x_k are of opposite signs. Even if it is not so i.e. $\underline{R}_o$ and $\underline{C}$ are of the same sign, E_P can be greater than unity provided $\underline{C} < \underline{e}$, $\underline{e}$ being vector of unities and

$$(19.12) \qquad [\underline{R}_o' - 2\underline{C}'/(1+\underline{C}' \underset{\sim}{R}^{-1} \underline{C})] \underset{\sim}{R}^{-1} \underline{C} > \underline{0}$$

When we assume that the auxiliary traits are themselves uncorrelated so that $\underset{\sim}{R} = \underset{\sim}{I}$, the efficiency is greater than unity if $\sum_{k=1}^{k} R_{ok}C_k$ is negative. When all C_k's are equal to C as well as all R_{ok}'s are equal to R_o, it is possible to see the effect of the number of auxiliary traits on the efficiency. It increases (decreases) with the increase in this number when R_o and C are having opposite (same) signs. The efficiency is affected seriously at higher values of n. For example, when n = 20, C = +0.2, the efficiency varies from 0.447 to 4.027 but when n = 1, it varies only from 0.980 to 1.061.

20. Narain, P. (1979). A new sire index for milk production corrected for an auxiliary trait. The Indian Journal of Animal Genetics and Breeding, 1(1): 20-22.

A sire index based on dam-daughter comparison and corrected for an auxiliary trait is given by

$$(20.1) \qquad S_I = A_I + \frac{2nW}{(n+a_y)}\left[(\bar{D}_I - A_I) - \frac{1}{2}h_I^2(\bar{M}_I - A_I)\right]$$

where $\bar{D}_I$ is the average performance of n daughters, $\bar{M}_I$ is the average performance of the corresponding dams, A_I is the herd average, and h_I^2 is the heritability of I where I is the phenotypic index (trait y corrected for auxiliary trait x)

(20.2) $I = P_y - bP_x$

where P_y and P_x are the phenotypic values of y and x expressed as deviations from the mean and 'b' is the regression coefficient of y on x. As in Narain and Mishra (19),

(20.3) $h_I^2 = h_y^2 (1-RC)^2/(1-R^2)$

where $R = bs_x/s_y$ and $C = rh_x/h_y$. W in the expression for S_I is given by

(20.4) $$W = \frac{1-r^2(n+a_{xy})/(n+a_x)}{1-r^2(n+a_{xy})^2/(n+a_x)(n+a_y)}$$

where $a_y = (4-h_y^2)/h_y^2$, $a_x = (4-h_x^2)/h_x^2$, $a_{xy} = (4R-rh_xh_y)/rh_xh_y$

h_x^2 and h_y^2 being the heritabilities of x and y respectively, R and r being the phenotypic and genetic correlations between x and y respectively and s_x and s_y are the phenotypic standard deviations of x and y respectively.

If we assume that the daughters and dams are having equal variabilities for each of the two characters and equal covariability between the two characters and if we neglect the sampling variance of h_y^2, we get

$$(20.5) \qquad V(S_I) \approx \left(\frac{2nW}{n+a_y}\right)^2 \frac{s^2_{D_y}}{n} (1-R^2)\left(1-\frac{1}{4}h_I^4\right)$$

where $s^2_{D_y}$ is the sampling variance of daughters for y. The efficiency of S_I relative to the corrected daughter average index (R=r=o) is then given by

$$(20.6) \qquad E_{S_I} = \left(1-\tfrac{1}{4}h_y^4\right) / W^2\left(1-\tfrac{1}{4}h_y^4 E^2 - R^2\right)$$

where $E=(1-RC)/\sqrt{1-R^2}$ is the efficiency of the phenotypic index selection relative to individual selection. If r = o, W = 1, C = o and E_{S_I} is clearly greater than unity. When n, the number of daughters per sire is adequately large, W tends to be 1 and E_{S_I} is greater than 1 whenever E is greater than 1. This happens whenever the phenotypic and genetic correlations are of opposite signs. When they are of the same sign, E is greater than 1 if C is less than 1 and R is greater than $2C/(1+C^2)$.

21. Narain, P., Malhotra, P.K. and Wahi, S.D. (1983). The use of auxiliary traits in combined selection for poultry improvement. Indian Journal of Poultry Science, 18: 37-47.

Let $I_{1k}(x_1, x_2, \ldots, x_k)$ denote the index which predicts the breeding value (G_y) of the individual for the main trait (y)

under improvement by combining, in an optimal manner, its own performance (P_y) for y, its performance $\underline{P}_x' = (P_{x_1}, P_{x_2}, \ldots, P_{x_k})$ for the set of k auxiliary characters $\underline{x}' = (x_1, x_2, \ldots, x_k)$, the average ($\bar{H}_y$) of the phenotypic values of n paternal half-sibs for y and the average ($\bar{F}_y$) of the phenotypic values of m full-sibs for y. It is then expressed as

$$(21.1) \qquad I_{1k}(x_1, x_2, \ldots, x_k) = \sum_{n=1}^{k} a_i P_{x_i} + b_1 P_y + b_2 \bar{H}_y + b_3 \bar{F}_y$$

where the coefficients $\underline{a}' = (a_1, a_2, \ldots, a_k)$, b_1, b_2 and b_3 are to be worked out in such a way that the accuracy in the prediction of G_y on the basis of I_{1k} is maximised. Similarly, let $I_{2k}(x_1, x_2, \ldots, x_k)$ denote the index which predicts G_y by combining, in an optimal manner, in addition to the information included in I_{1k}, the information supplied by the performance (D_y) of the individual's dam for the character y. That is,

$$(21.2) \qquad I_{2k}(x_1, x_2, \ldots, x_k) = \sum_{i=1}^{k} a_i' P_{x_i} + b_1' P_y + b_2' \bar{H}_y + b_3' \bar{F}_y + b_4' D_y$$

where the coefficients are to be similarly determined. The optimisation gives the multiple correlation coefficient ($r_{GI_{1k}}$) between G_y and I_{1k} and hence the efficiency of the index over that on individual performance for y as

$$(21.3) \qquad E_{1k} = r_{GI_{1k}}/h_y$$

$$= \left[\frac{\{h_y^2+\alpha(1-h_y^2)\}(1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R})-h_y^2(\underset{\sim}{R}_o-\underset{\sim}{r}_o+\alpha(\underset{\sim}{C}-\underset{\sim}{R}_o))'\underset{\sim}{R}^{-1}(\underset{\sim}{C}-\underset{\sim}{R}_o)}{h_y^2\{1+\alpha(1-h_y^2)\}(1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R}_o)-\alpha h_y^4(\underset{\sim}{C}-\underset{\sim}{R}_o)'\underset{\sim}{R}^{-1}(\underset{\sim}{C}-\underset{\sim}{R}_o)}\right]^{\frac{1}{2}}$$

where $\alpha = h_y^2(4M+N-MNh_y^2)/(16-4Mh_y^2-Nh_y^2)$, $N = n/(1+\frac{n-1}{4}h_y^2)$ and $M = m/(1+\frac{m-1}{2}h_y^2)$. When either $\underset{\sim}{R}_o = \underset{\sim}{r}_o = \underset{\sim}{o}$, $\underset{\sim}{C} = o$ or $\underset{\sim}{R}_o = \underset{\sim}{C}$, the efficiency reduces to the efficiency of Osborne index. But when only genetic correlations are zero i.e. $\underset{\sim}{r}_o = \underset{\sim}{o}$, $\underset{\sim}{C} = \underset{\sim}{o}$, we have

$$(21.4) \qquad E_{1k}(r_o=o) = \left[\frac{h_y^{*2}+\alpha(1-h_y^{*2})}{h_y^{*2}\{1+\alpha(1-h_y^{*2})\}}\right]^{\frac{1}{2}}$$

where $h_y^{*2} = h_y^2/(1-\underset{\sim}{R}_o'\underset{\sim}{R}^{-1}\underset{\sim}{R}_o)$ is the heritability of the trait y after corrections for the auxiliary traits as in Narain and Mishra (19). Since h_y^{*2} is necessarily greater than h_y^2, the efficiency is always greater than that for Osborne's index when genetic correlations are zero. The limiting value of I_{1k}, as n and m approach infinity, is given by

$$(21.5) \qquad \lim_{\substack{n\to\infty \\ m\to\infty}} I_{1k} = \left[\frac{\{1-h_y^{*2}(\underset{\sim}{C}-\underset{\sim}{r}_o)'\underset{\sim}{R}^{-1}(\underset{\sim}{C}-\underset{\sim}{R}_o)\}}{h_y^2\{(2-h_y^2)-h_y^{*2}(\underset{\sim}{C}-\underset{\sim}{R}_o)'\underset{\sim}{R}^{-1}(\underset{\sim}{C}-\underset{\sim}{R}_o)\}}\right]^{\frac{1}{2}}$$

With either $\underline{R}_o = \underline{r}_o = \underline{o}$ or $\underline{C} = \underline{R}_o$, we get the limiting value of the Osborne's index.

In a similar manner we get E_{2k}, which is similar in form, to E_{1k} with α replaced β given by

$$(21.6) \qquad \beta = h_y^2\left[\frac{4M+N-MNh_y^2+4(1-Mh_y^2)}{16-4Mh_y^2-Nh_y^2-4h_y^2}\right]$$

The results for the particular cases are also similar in form to those for E_{1k} except that when we consider the limiting case of I_{2k}, we find it to be exactly the same as Lim.I_{1k}.

The efficiencies of the indices for two uncorrelated auxiliary traits I_{12} and I_{22} for variations in the values of h_y^2 for various combinations of R_o and r_o assumed to be the same for the auxiliary traits, when h_x^2, assumed to be the same for both the auxiliary traits, equals 0.5 and m=3, n=20, $\sigma_y = \sigma_x$, are shown in Figure 21.1. The efficiencies of the corresponding Osborne's indices (I_{10} and I_{20}) are also plotted for necessary comparisons. It is evident that the inclusion of two additional traits, on an individual basis, increases the efficiency of the new indices at all values of h_y^2. The maximum gain however, occurs when R_o and r_o are of opposite signs. A negative value of R_o has a greater influence than a negative value of r_o. When the genetic correlation is zero,

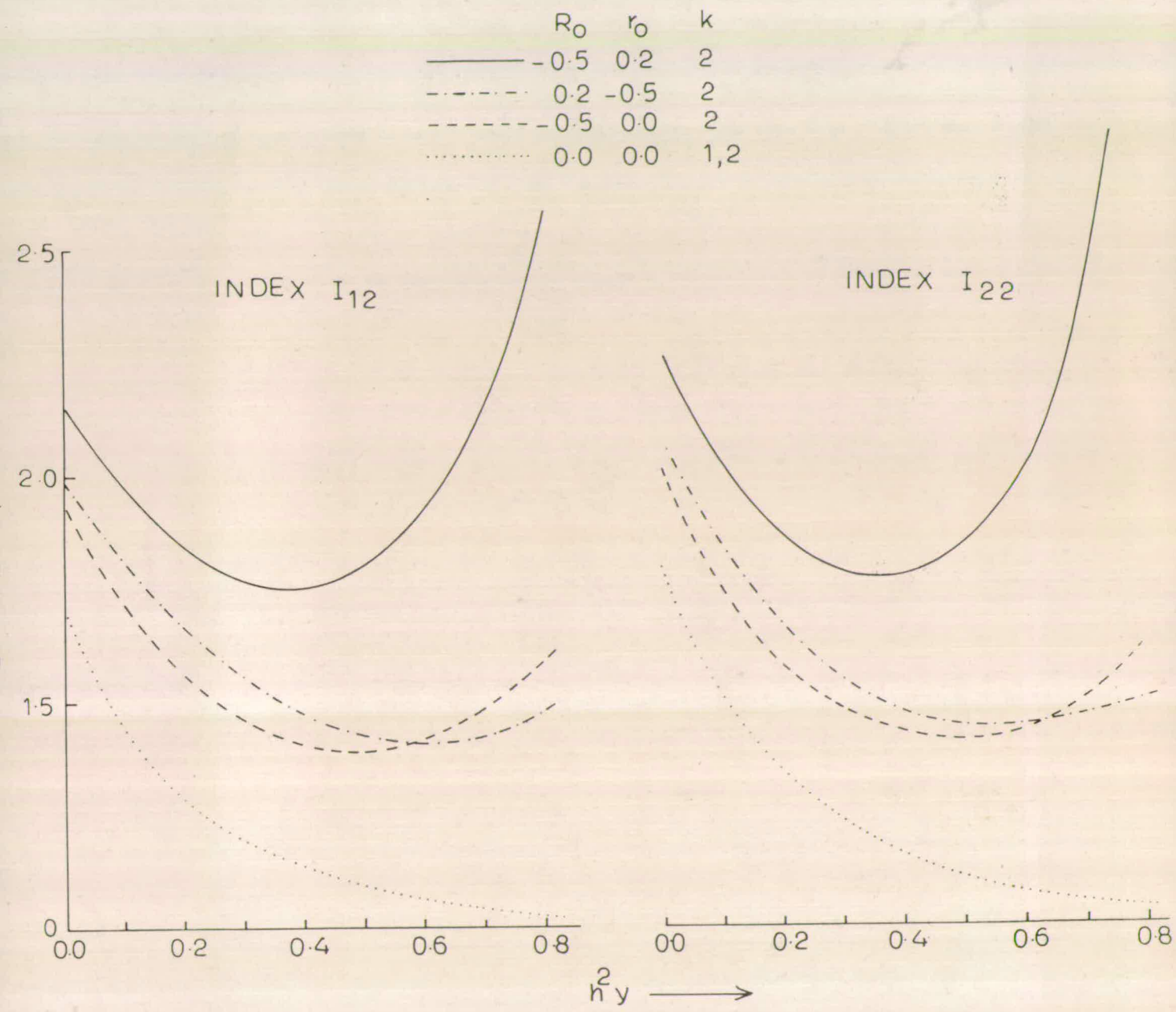

Fig. 21.1 Efficiency of selection indices for different values of h^2_y with $h^2_x = 0.5$, $m = 3$, $n = 20$ and $\sigma_y = \sigma_x$.

there is gain in the efficiency provided the phenotypic correlation is non-zero. An interesting result emerges that while Osborne's indices I_{10} and I_{20} are useful only at lower values of h_y^2, the new indices could be useful even at higher values of h_y^2.

In order to compare the effect of including two auxiliary traits as compared to one auxiliary trait, the efficiencies for variations in h_y^2 for the combination $R_o = -0.5$, $r_o = 0.2$ for k=1 and 2 are shown in Figure 21.2 along with those of Osborne's indices when $h_x^2 = 0.5$, m=3, n=20 and $\sigma_y = \sigma_x$. It is apparent that the inclusion of one more auxiliary trait improves the efficiency with higher gains at higher values of h_y^2.

The effect of a positive or a negative correlation (ρ) between the two auxiliary traits on the efficiency has been shown in Figure 21.3 taking the combination $R_o = 0.2$, $r_o = -0.5$ and the same values of other parameters. It is quite clear that a positive correlation between the auxiliary traits increases the efficiency whereas a negative correlation decreases it compared to the uncorrelated case for all values of h_y^2 and for either of the two indices. It is important to note that although the efficiency gets reduced for a negative correlation between the auxiliary traits, it is still higher than those of the Osborne's indices.

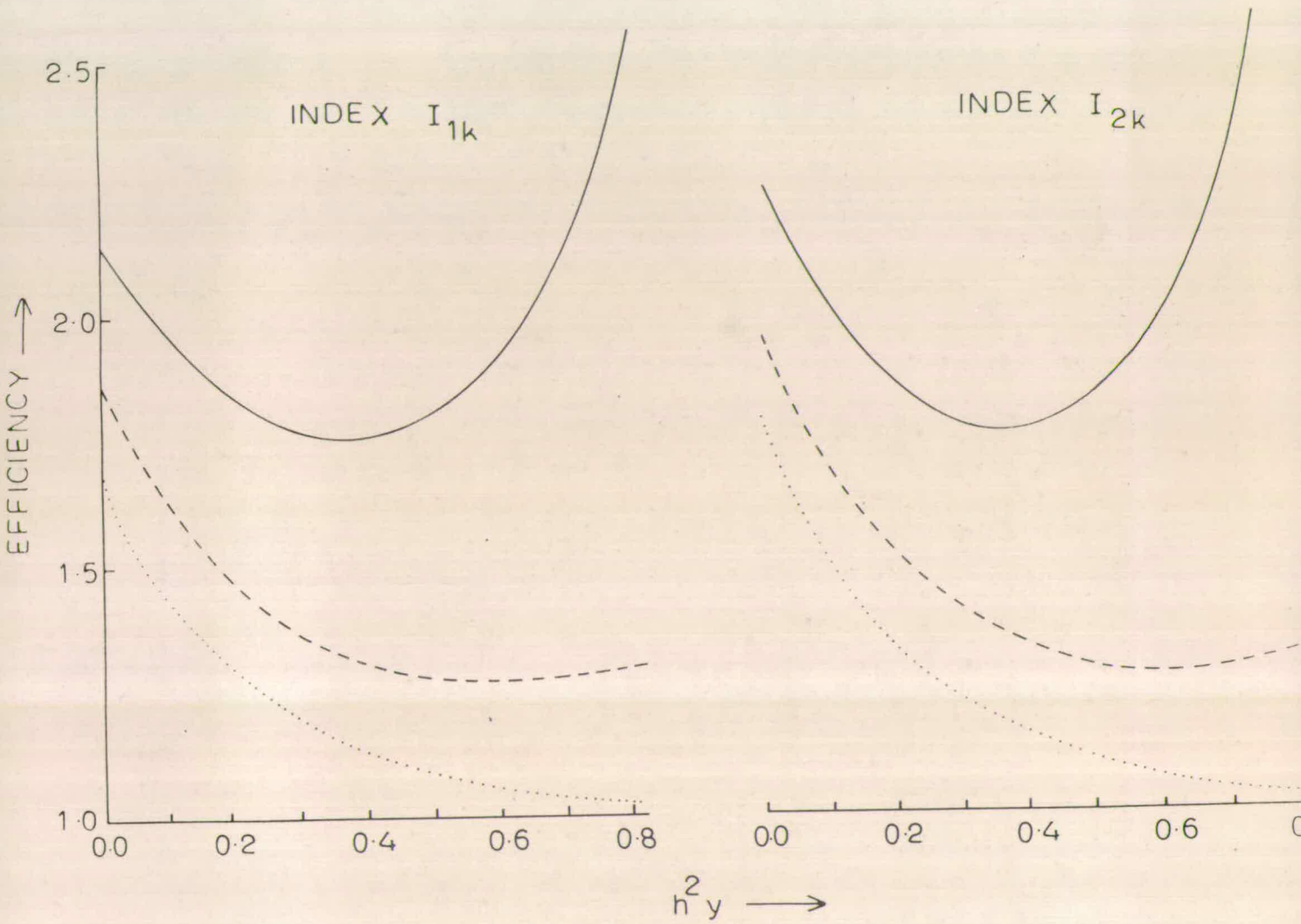

Fig. 21.2 Effect of the number of auxiliary traits on the efficiency of selection indices with $R_o = -0.5$, $r_o = 0.2$, $h^2_x = 0.5$, $m = 3$, $n = 20$ and $\sigma_y = \sigma_x$.

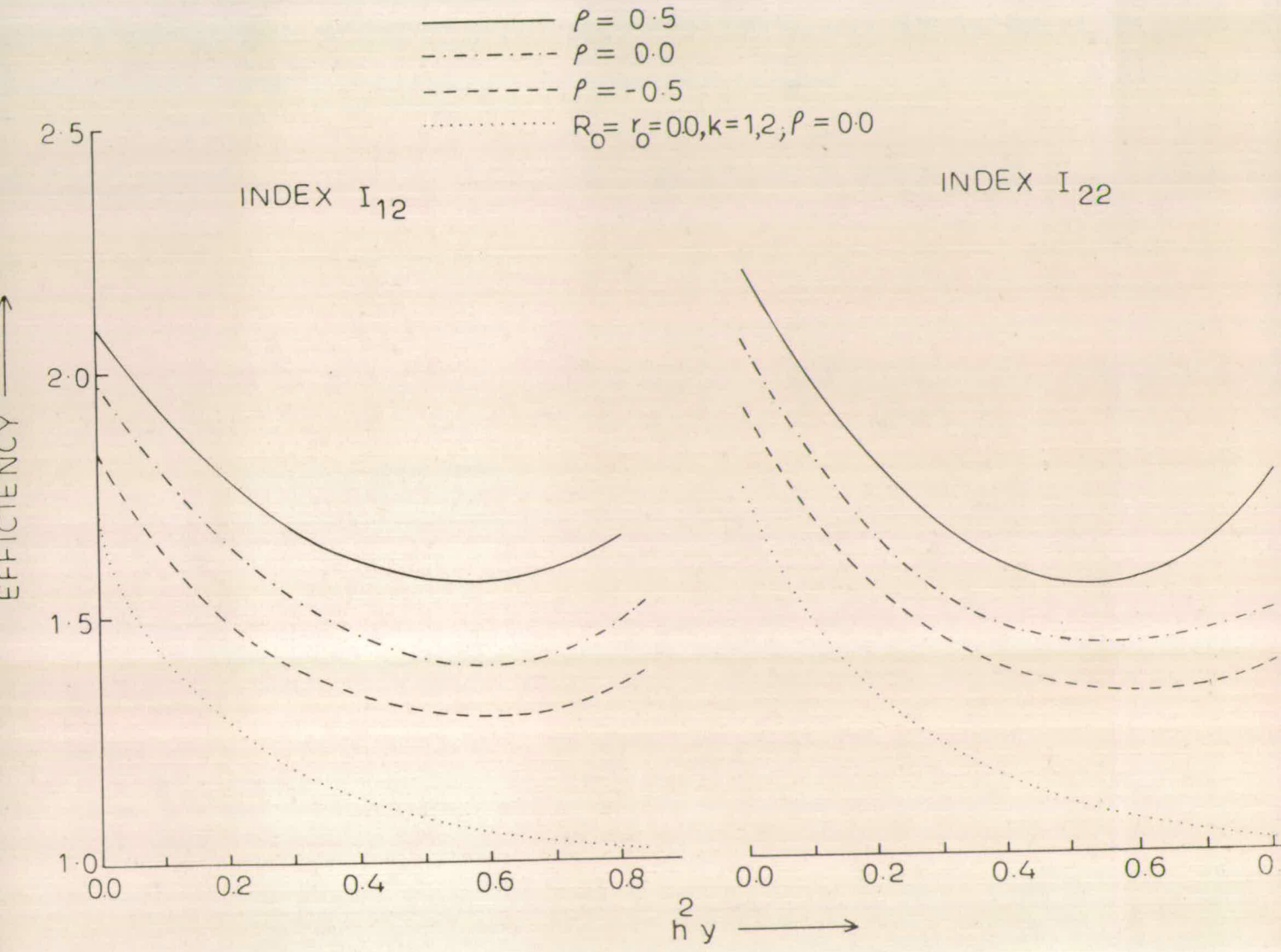

Fig. 21.3 Effect of the correlation between the auxiliary traits on the efficiency of the selection indices with $R_o = 0.2$, $r_o = -0.5$, $h_x^2 = 0.5$, $m = 3$, $n = 20$ and $\sigma_y = \sigma_x$.

22. Narain, P., Rao, C. Subba and Nigam, A.K. (1974). Partial diallel crosses based on extended triangular association scheme. Indian Journal of Genetics and Plant Breeding, 34: 309-317.

Let the number of lines n be of the form $p(p-1)(p-2)/6$ where p is a positive integer greater than 3. Now identify a line by a triplot abc where $1 \leq a < b < c \leq p$. All the lines can then be numbered off into (p-2) different triangles $T_1, T_2, \ldots, T_{(p-4)}$, $T_{(p-3)}$ and $T_{(p-2)}$ of orders $(p-2)\times(p-2)$, $(p-3)\times(p-3)$, ...2x2 and 1x1 respectively. The number of triplets in the i-th triangle is $(p-i)(p-i-1)/2$ for $i=1,2,\ldots,(p-2)$ which add to total number of given lines. There are then three possible designs for PDC. In design (1), all crosses of the type abc x def are sampled where a, b, c, d, e and f are all distinct. This gives $s_1=(p-3)(p-4)(p-5)/6$ (with $p>5$) which is the number of times any line is involved in crossing with other lines, resulting in a total number of $ns_1/2$ crosses to be raised. Design (2) samples all the crosses of the type abc x def where one of the letters (a, b, c, d, e and f) is common. This gives $s_2 = 3(p-3)(p-4)/2$ with total number of $ns_2/2$ crosses. Design (3), being a complimentary to the other two, picks up all the remaining crosses not sampled by the former two designs. Alternatively, it samples all the crosses of the type abc x def where two of

the letters a, b, c, d, e and f are in common. This gives $s_3 = 3(p-3)$ with total number of crosses as $ns_3/2$. It can be seen that these three possible PDCs correspond to picking up third, second and first associate of each treatment of the extended triangular association scheme and pairing the treatment with each member of the corresponding associate class. As such, this type of design of PDC is called Extended Triangular (ET) design.

The analysis of PDC constructed above follows the pattern of the analysis of three-associate PBIB design. The mean yield of the cross between i-th and j-th lines is expressed as

$$(22.1) \qquad \bar{Y}_{ij} = \mu + t_i + t_j + s_{ij} + \bar{e}_{ij}$$

where μ is the effect due to overall mean, t_i and t_j are the g.c.a. effects due to i-th and j-th lines respectively, s_{ij} is the s.c.a. effect due to the cross i x j and $\bar{e}_{ij}$ is the random error. We assume $\sum_{i=1}^{n} t_i = o$, $\sum_{j=1}^{n} s_{ij} = o$ for each i and that t_i, s_{ij} and $\bar{e}_{ij}$ are each independently normally distributed with zero means and variances σ_g^2, σ_s^2 and σ_e^2/r if there are r replications of the experiment. For design (1), the third associates are sampled giving the usual normal equations which involve the usual secondary parameters of the design, which are

solely functions of p. Solving the normal equations, we get estimates of t_i subject to $\sum_{i=1}^{n} \hat{t}_i = o$. This, in turn, gives the sum of squares due to g.c.a. effects as $\sum_{i=1}^{n} \hat{t}_i T_i$, where T_i is the total yield of all the crosses involving the i-th parent. The sum of squares due to s.c.a. effects is obtained by subtracting this sum of squares from the sum of squares due to crosses. The usual Analysis of Variance and F-tests then follow.

The variance of the difference between g.c.a. effects of a pair of lines is of three types; $V_1 = Var(\hat{t}_i - \hat{t}_j)$ where j-th line is crossed with i-th line, $V_2 = Var(\hat{t}_i - \hat{t}_k)$ where k-th line is not crossed with i-th line but is crossed with a line with which i-th line is crossed; and $V_3 = Var(\hat{t}_i - \hat{t}_l)$ where l-th line is neither crossed with i-th line nor with a line with which i-th line is crossed. The average variance (D_1) of the difference between the g.c.a. effects of any two lines is then

$$(22.2) \qquad D_1 = (s_1V_1 + s_2V_2 + s_3V_3) \Big/ (s_1 + s_2 + s_3)$$

In a similar manner, we can get D_2 and D_3 corresponding to design (2) and design (3) respectively. These are ultimately found to be functions of 'p' only. As such by giving different values to 'p', we can study the behaviour of average variances

in the three cases.

The efficiency of ETD compared to circulant design (CD) of Kempthorne and Curnow (1961) for PDC can be obtained by comparing the average variances in the two cases for the same number of crosses sampled. This has been presented in Table 22.1 for $n = 20$, $p = 6$ and $n = 35$, $p = 7$.

TABLE 22.1 Efficiency (E) of ETD relative to CD in estimating g.c.a. effects.

n	s		CD	ETD	Efficiency (E)
20	s_1	1	-	-	-
	s_2	9	0.522	0.263	1.986
	s_3	9	0.522	0.277	1.886
35	s_1	4	2.101	0.725	2.897
	s_2	18	0.121	0.116	1.046
	s_3	12	0.210	0.185	1.136

It is evident from this Table that the efficiency of ETD is always greater than one and design (1) of ETD is definitely much more efficient than the other two designs of ETD. For $n = 35$, for instance, ETD corresponding to design (3) is 14 per cent more efficient but that corresponding to design (2) is only 5 per cent more efficient than CD.

23. Narain, P. and Arya, A.S. (1981). Truncated triangular association scheme and related partial diallel crosses. Sankhya: The Indian Journal of Statistics, 43(Series B); 93-103.

Let n, the number of lines randomly arranged, be of the form p(p-2)/2 where p is an even positive integer greater than 6. An association scheme, termed as Truncated Triangular (TT) with five associate classes and p(p-2)/2 symbols arranged in a square array of p rows and p columns, satisfies the following properties:

(i) The positions in the principal diagonal (running from the upper left hand to the lower right hand corner) as well as in the other diagonal (running from the upper right hand to the lower left hand corner) are left blank.

(ii) The p(p-2)/2 positions above the principal diagonal are filled by the numbers 1,2,...,p(p-2)/2, corresponding to the symbols.

(iii) The p(p-2)/2 positions below the principal diagonal are filled so that the array is symmetrical about the principal diagonal.

(iv) For any symbol i, the first associates are exactly those that occur in the same row or in the same column as i, except those two symbols i' and i'' which occupy the same position as i with respect to the other diagonal (when positions above this diagonal are filled with the symbols and positions below this diagonal are filled so that the array is symmetrical about this diagonal).

(v) The symbols i' and i'' are the second associates of i.

(vi) The symbols occurring in the same column as i' or in the same column as i'' except the common symbol i''' between the two cases are exactly the third associates of i.

(vii) The symbol i''' is the fourth associate of i.

(viii) The remaining symbols are the fifth associates of i.

This gives the usual parameters of the association scheme as $v = p(p-2)/2$, $n_1 = 2(p-4)$, $n_2 = 2$, $n_3 = 2(p-4)$, $n_4 = 1$, $n_5 = (p-4)(p-6)/2$.

Three designs of PDC can be constructed with the help of this scheme. In design (1), we can grow all the crosses in which a line is crossed with all other lines falling in its first associate class, giving $s_1 = 2(p-4)$ and total number of crosses as $ns_1/2 = p(p-2)(p-4)/2$. In design (2), we can sample the crosses of the type $i \times j$ where i and j are third associates. Here again $s_2 = 2(p-4)$ with $ns_2/2 = p(p-2)(p-4)/2$. In

design (3), we can pick up the crosses of the type involving lines which are fifth associates, giving $s_3=(p-4)(p-6)/2$ where p must exceed 8. The remaining two designs with s=2 and s=1 corresponding to second and fourth associates respectively are not suitable as they lead to singular least squares equations.

For the analysis of PDC based on any one of the designs enumerated above, we have the usual normal equations in matrix notations

$$\underset{\sim}{A}\hat{\underline{g}} = \underline{Q} \tag{23.1}$$

where $\underset{\sim}{A}$ is a n x n matrix having diagonal elements a_{ii} all equal to s and $a_{ij}=a_{ji}=1$ if the cross (ixj) is sampled and o otherwise, $\hat{\underline{g}}'=(\hat{g}_1,\hat{g}_2,\ldots,\hat{g}_n)$, vector of g.c.a. effects of lines, $\underline{Q}'=(Q_1,Q_2,\ldots,Q_n)$ where $Q_i=\sum_{j(i)}\bar{y}_{ij}-(2G/n)$ is the right hand side of the i-th normal equation with $\sum_{j(i)}$ referring to the summation over lines j crossed with line i in the design and G is the total of cross mean yields. If $\underset{\sim}{A}$ is non-singular, the estimates of g.c.a. effects and sum of squares due to the estimates are

$$\hat{\underline{g}} = \underset{\sim}{A}^{-1}\underline{Q} \tag{23.2}$$

$$\text{S.S. due to } \hat{\underline{g}} = \underline{Q}'\underset{\sim}{A}^{-1}\underline{Q} \tag{23.3}$$

and dispersion matrix of the estimates is $\underset{\sim}{A}^{-1}\sigma^{-2}$ where σ^{-2} is the variance of $\bar{Y}_{ij}$ being $(\sigma_s^{-2} + \sigma_e^{-2}/r)$, r being the number of replications in the experiment. The ANOVA can be set up and F-tests applied as usual.

Since the PDC's under discussion are associated with 2-plot block designs with TT association scheme having 5 associate classes where $\alpha_1, \alpha_2, \ldots, \alpha_5$ take values either o or 1, $\underset{\sim}{A} = \underset{\sim}{N}\underset{\sim}{N}'$, where $\underset{\sim}{N}$ is the incidence matrix of the corresponding PBIB design. As such the number of distinct elements in $\underset{\sim}{A}^{-1}$ or in the associated idempotent matrices $\underset{\sim}{L}_i$'s cannot exceed 6. Following Bose and Mesner (1959), the latent roots of $\underset{\sim}{A}$, distinct elements of $\underset{\sim}{L}_i$'s and hence the elements of $\underset{\sim}{A}^{-1}$ using spectral decomposition $\underset{\sim}{A}^{-1} = \sum_{i=0}^{5} \phi_i^{-1} \underset{\sim}{L}_i$, with ϕ_i as the i-th latent root, were worked out by a cumbersome procedure given in the paper. Finally g_i's are obtained from these elements.

For the estimated difference $(\hat{g}_i - \hat{g}_j)$ there will, in general, be 5 different variances according as the lines i and j are first or second.....or fifth associates. For this purpose, one has to write down the association relationship among the n symbols representing the n lines. Let us denote the appropriate variance for the difference $(\hat{g}_i - \hat{g}_j)$ by V_k if the two lines correspond to k-th associates. Then

(23.4) $V_k = 2\sigma^{-2}(a^o - a^k)$, k=1,2,3,4,5

where a^{ij}, the ij-th element of $\underset{\sim}{A}^{-1}$, is a^k if the lines i and j are k-th associate (k=o,1,2,3,4,5), a^o being in the diagonal position. The average variance ($\bar{V}$), for comparing the efficiency of one design relative to the other, for the same values of n and s, is

(23.5) $$\bar{V} = \sum_{k=1}^{5} n_k V_k \Big/ \sum_{k=1}^{5} n_k$$

For the three designs, $\bar{V}/\sigma^{-2}$ is found to be entirely a function of 'p' only and therefore if p is known corresponding to a given n, we know the efficiency of the design. It may be noted that while design (3) is found to be a three-variance sample, designs (1) and (2) are five-variance samples. The average variances per unit σ^{-2} for the three designs for admissible values of n between 24 and 180 are presented in Table 23.1.

TABLE 23.1 Average variance of the difference between the g.c.a. effects per unit σ^{-2} for the three PDC's.

No. of lines		Design (1)		Design (2)		Design (3)	
n	p	s_1	$\bar{V}/\sigma^{-2}$	s_2	$\bar{V}/\sigma^{-2}$	s_3	$\bar{V}/\sigma^{-2}$
24	8	8	0.286	8	0.298	4	*
40	10	12	0.181	12	0.187	12	0.195
60	12	16	0.132	16	0.136	24	0.089
84	14	20	0.104	20	0.107	40	0.052
112	16	24	0.086	24	0.088	60	0.034
144	18	28	0.074	28	0.075	84	0.024
180	20	32	0.064	32	0.065	112	0.018

* Singular least squares equations.

It is clear from the Table that the average variance decreases with increase in the value of s, as expected. It does not change much, if s does not change irrespective of the design and value of n. For instance, with s=24, the average variances are found to be $0.089\sigma^{-2}$, $0.086\sigma^{-2}$ and $0.088\sigma^{-2}$ respectively for n equal to 60 for design (3), 112 for design (1) and 112 for design (2). A comparison of average variances indicates that design (1) is consistently more efficient than design (2) since for the same value of n and s, the former always has a lower average variance than does the latter, though as n increases, the differences in average variance tend to narrow down considerably. With n=40, s=12, design (1) has a lower average variance than design (2) as well as design (3).

REFERENCES

BOSE, R.C. and MESNER, D.M. (1959). On linear associative algebras corresponding to association schemes of partially balanced designs. Ann. Math. Statist., 30: 21-38.

CURNOW, R.N. (1963). Sampling the diallel cross. Biometrics, 19: 287-306.

FELLER, W. (1951). Diffusion processes in genetics. Proc. 2nd Berkeley Symp. Math., Statist. and Prob., 227-246, University of California Press, Berkeley, CA.

FINNEY, D.J. (1952). The equilibrium of a self-incompatible polymorphic species. Genetics, 26: 33-64.

FINNEY, D.J. (1962). Genetic gains under three methods of selection. Genet. Res. Camb., 3: 417-423.

FISHER, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edin., 52: 399-433.

FISHER, R.A. (1922). On the dominance ratio. Proc. Roy. Soc. Edin., 42: 321-341.

FISHER, R.A. (1930). The Genetical Theory of Natural Selection. Dover Publications, Inc., New York, 1958: 291 pp.

GILBERT, N. (1958). Diallel crosses in plant breeding, Heredity, 12: 477-492.

HAZEL, L.N. (1943). The genetic basis for constructing selection indexes. Genetics, 28: 476-490.

HALDANE, J.B.S. (1924). A mathematical theory of natural and artificial selection. Part I. Trans. Camb. Phil. Soc., 23: 19-41.

KARLIN, S. and LEVIKSON, B. (1974). Temporal fluctuations in selection intensities: Case of small population size. Theor. Popl. Biol., 6: 383-412.

KEMPTHORNE, O. and CURNOW, R.N. (1961). The partial diallel crosses. Biometrics, 17: 229-250.

KIMURA, M. (1957). Some problems of stochastic processes in genetics. Ann. Math. Statist., 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. Methuen's Review Series in Applied Probability, Vol.2: 56 pp.

KIMURA, M. (1968). Evolutionary rate at the molecular level. Nature, 217(5129): 624-626.

KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics, 61: 893-903.

KIMURA, M. (1970). The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet. Res., Camb. 15: 131-134.

KIMURA, M. and MARUYAMA, T. (1969). The substitutional load in a finite population. Heredity, 24: 101-114.

KIMURA, M. and OHTA, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. Genetics, 61: 763-771.

KIMURA, M. and OHTA, T. (1973). The age of a neutral mutant persisting in a finite population. Genetics, 75: 199-212.

MALÉCOT, G. (1948). Les Mathematiques de l'Hérédité. Masson et Cie, Paris, 63 pp.

MARUYAMA, T. (1974). The age of an allele in a finite population. Genet. Res., Camb. 23: 137-143.

MATHER, K. (1949). Biometrical Genetics. The study of continuous variation. Metheun and Co. Ltd., London, 162 pp.

NAGASAWA, M. and MARUYAMA, T. (1979). An application of time reversal of Markov processes to a problem of population genetics. Adv. Appl. Prob. 11: 457-478.

OSBORNE, R. (1957). Correction for regression on a secondary trait as a method of increasing the efficiency of selective breeding. Australian J. Biol. Sci., 10: 365-366.

PURSER, A.F. (1960). The use of correction for regression on a second character to increase the efficiency of selection. In Biometrical Genetics. Ed. O. Kempthorne, Pergamon Press, London, 234 pp.

RENDEL, J.M. (1954). The use of regression to increase heritability. Australian J. Biol. Sci., 7: 368-378.

ROBERTSON, A. (1957). Optimum group size in progeny testing and family selection. Biometrics, 13: 442-450.

ROBERTSON, A. (1960). A theory of limits in artificial selection. Proc. Roy. Soc., London. B 153: 234-249.

SCHNELL, F.W. (1961). Some general formulations of linkage effects in inbreeding. Genetics, 46: 947-957.

SCUDO, F.M. (1964). Sex population genetics. Ric. Sci., 34 (II-B): 93-146.

SEARLE, S.R. (1965). The value of indirect selection. I. Mass Selection. Biometrics, 21: 682-707.

YOUNG, S.S.Y. and WEILER, H. (1960). Selection for two correlated traits by independent culling levels. J. Genetics, 57: 329-338.

WRIGHT, S. (1921). Systems of mating I-V. Genetics, 6: 111-178.

WRIGHT, S. (1931). Evolution in Mendelian populations. Genetics, 16: 97-159.

S E C T I O N T W O

STUDIES
IN
STATISTICAL GENETICS
- UNPUBLISHED WORK

C O N T E N T S

I	GENETIC DIFFERENTIATION OF QUANTITATIVE CHARACTERS BETWEEN POPULATIONS	159
II	PROGENY TESTING WITH AUXILIARY TRAITS	179
	REFERENCES	192

I. GENETIC DIFFERENTIATION OF QUANTITATIVE CHARACTERS BETWEEN POPULATIONS.

Several workers, in the past, studied the evolutionary changes of quantitative characters by considering two alleles at each of the several loci controlling the character and assuming an optimum model for selection acting at the phenotypic level (Fisher, 1922; Wright, 1931, 1937; Robertson, 1956). For the first time, it was Kimura (1965) who considered a more meaningful model involving mutation as producing multiple alleles with varying phenotypic effects at each of the several loci affecting the trait. He showed, for infinitely large populations, that the distribution of allelic effects tends to be normal at equilibrium between selection and mutational forces and that the mean and variance of the equilibrium distribution are determined by the amounts of increase in mean and variance of the genotypic value per gene per generation as well as by the intensity of fitness function. Chakraborty and Nei (1982) have recently considered a discrete allelic state model of mutation instead of continuous allelic state model of Kimura (1965) but they ignored selection and considered only mutation and random drift as the forces affecting the changes in the underlying gene frequencies of the loci controlling the character. Such an approach, however, needs modification as most quantitative characters are subject to adaptive genetic changes and hence

the role of selection cannot be ruled out.

Although adaptive genetic changes are often the outcome of environmental changes in somewhat complex manner, there are quantitative characters wherein the effect of environment seems fairly simple. Skin pigmentation in man is one case which is, to a great extent, related to adaptation to sunlight. By knowing the sunlight intensity for a given population, one may be able to model the evolutionary change of pigmentation. It is believed, for instance, that a small group of Caucasian race with fair skin moved out of Central Asia around 30,000 years ago and settled in southern parts of America with plenty of sunlight. Their skin pigmentation changed, over time, due to a shift in the optimum phenotype for pigmentation. Apparently, selection of an optimal type must have been at work so that individuals further away from the optimum for skin pigmentation tended to have lower fitness. If we consider selection along with discrete allelic model of mutation and study the variation within and between populations when the populations have been reproductively isolated for a long time, as in the above example, we may be able to determine the selective stress due to adaptation for the changed environment during the period of human migration. With this aim in view, we have developed in this study a model for genetic differentiation of quantitative characters between populations or species.

The evolution of quantitative characters can be studied in either of two ways. We may define and study the underlying model entirely at the level of phenotype avoiding any reference to gene frequencies. Lande (1976) approached the problem from this angle. The other way would be to define and study a genetic model assuming that a complete genetic analysis of the trait is possible. In this case, we have to start from the simplest situation of a single locus with two alleles and build over it the more complex systems of multiple alleles, several linked loci etc. The results obtained from the simpler situation of one or two loci give an insight to the problem even though the quantitative trait is governed by several loci, possibly linked. We follow this approach in the present study.

Consider a quantitative character controlled by k loci with an infinite number of possible allelic types at each locus. For a given locus, let A_i represent an allele occupying state i (any integer number from $-\infty$ to ∞) and having an allelic effect of i a. We assume that all allelic effects are additive with no dominance and no epistasis and that once A_i mutates, it changes to allelic state $(i + r)$ with probability

$$(1.1) \qquad a_r = a_{-r} = \binom{2m}{m-r}(1/2)^{2m} \qquad \text{for } o \leq r \leq m$$

$$a_r = o, \quad \text{otherwise}$$

where m is the number of discrete steps in which mutations can occur. We may note that this probability law is a shifted binomial distribution. If v denotes the mutation rate, the absolute probability of such a mutation would be va_r. If an allele mutates but has the same allelic effect i, the probability would be va_o. In the conventional sense therefore, the real mutation rate is $v' = (1-a_o)v$.

The selection operates on the total phenotypic value which is assumed to follow a normal distribution with mean $\bar{Y}$ and variance σ_P^2 with probability density function

$$(1.2) \qquad f(y) = (1/\sigma_P\sqrt{2\pi})\exp[-(y-\bar{y})^2/2\sigma_P^2]$$

The fitness function for the character value y is assumed to be of the form

$$(1.3) \qquad w(y) = w_{max}\exp[-(y-y_{opt})^2/2\sigma_w^2]$$

where the quantitative character is assumed to be optimum for fitness at $y = y_{opt}$ with maximum fitness of w_{max} and σ_w is the width of the function indicating the rate at which fitness declines with deviation of y from the optimum value y_{opt}. Taking $w_{max} = 1$, the mean fitness of the population is then found to be

$$(1.4) \qquad \bar{w} = \sigma_w \sqrt{2s}\, \exp[-s(\bar{y}-y_{opt})^2]$$

with $s = 1/2(\sigma_w^2 + \sigma_p^2)$ which indicates the strength of the selection at the group level. Weak selection of optimum type is indicated by a large σ_w for a fixed phenotypic variability. For the given locus, we assume that its contribution to σ_P^2 is small so that the sub-population of the values of the character for those individuals with genotype A_iA_j is normally distributed with mean $a(i+j)$ and variance σ_P^2 due to environment and effects at other loci. Then the fitness function w_{ij} for A_iA_j would be proportional to $\exp[-s\{a(i+j) - y_{opt}\}^2]$.

We first take the case when a population evolves from monomorphism with an optimum phenotype which can be taken at the origin of the phenotypic scale with $y_{opt} = o$. Let $x_i(t)$ denote the frequency of allele A_i in generation t with allelic effect a i. Then genotype A_iA_j would have a mean fitness $w_{ij} = \exp[-sa^2(i+j)^2]$. This gives the change in the gene frequency of A_i from generation (t-1) to t as

$$\begin{aligned}(1.5) \qquad \bar{w}_A(t)x_i(t) = {} & (1-v')\sum_j x_i(t-1)x_j(t-1)\exp[-sa^2(i+j)^2] \\ & + v\sum_{r=1}^{m} a_r \sum_j x_j(t-1)[x_{i+r}(t-1)\exp(-sa^2(i+j+r)^2) \\ & + x_{i-r}(t-1)\exp(-sa^2(i+j-r)^2)]\end{aligned}$$

where $\bar{w}_A(t)$ is the mean fitness of individuals at the locus in the generation t so adjusted as to make $\sum_j x_j(t) = 1$.

In general, this recurrence relation does not yield any explicit solution. As such, it was used on the computer to generate the distribution of allelic effects for studying the statistical properties both within and between populations. However, for $m = 1$, it is possible to arrive at an analytical solution by approximating the recurrence relation, neglecting powers and products of v and s. The mean fitness is then approximated as $\bar{w}_A(t) \approx 1 - s\sigma^2_{g_A}(t-1)$ where $\sigma^2_{g_A}(t-1)$ is the total genotypic variance contributed by the locus in the previous generation (t-1). From the approximated recurrence relation an expression for the change in the gene frequency at i-th locus per generation is given by

$$(1.6) \qquad \Delta x_i(t) = -v\left[x_i(t) - \frac{x_{i+1}(t) + x_{i-1}(t)}{2}\right] + s\left[\frac{\sigma^2_{g_A}(t)}{2} - a^2 i^2\right]$$

When the population reaches equilibrium under the opposing forces of mutation and selection, $\Delta x_i = 0$, giving

$$(1.7) \qquad \hat{x}_1 = (1 - SG)\hat{x}_0$$

$$\hat{x}_{i+1} - 2[1 - S(G - i^2)]\hat{x}_i + \hat{x}_{i-1} = 0 \qquad \text{for} \quad i \geq 1$$

with $\hat{x}_i = \hat{x}_{-i}$ for all non-zero integers i, where $S = (sa^2/v)$ and

$G = \frac{1}{2}\sum_i \sum_j \hat{x}_i \hat{x}_j (i+j)^2 = (\hat{\sigma}^2_{g_A}/2a^2)$. The recurrence relation for $i \geq 1$ is a second order linear difference equation with variable coefficients. For solving it, we transform it to

$$(1.8) \qquad \Delta^2 \hat{x}_i + 2S[G-(i+1)^2]\Delta \hat{x}_i + 2S[G-(i+1)^2]\hat{x}_i = 0$$

Substituting $\hat{x}_i = \sum_{n=0}^{\infty} C_n i^{(n)}$ in this expression where $i^{(n)}$ is factorial function defined by $i^{(n)} = i(i-1)(i-2).....(i-n+1)$, we get a polynomial equation in i, which in turn is an identity so that each coefficient of i^n becomes zero. This ultimately gives two arbitrary solutions $\hat{x}_{i1}$ and $\hat{x}_{i2}$ of $\hat{x}_i$ giving the general solution $\hat{x}_i = \alpha_1 \hat{x}_{i1} + \alpha_2 \hat{x}_{i2}$, α_1 and α_2 being determined from initial conditions. We finally get

$$(1.9) \qquad \hat{x}_i = \hat{x}_o[1-SGi^{(1)} + S(1-G)(1-SG)i^{(2)} + \frac{S}{3}\{3-SG(6-G) + 2S(1-G)(4-G)(1-SG)\} i^{(3)} + ...]$$

It may be noted, from solution for $\hat{x}_1$, that since SG has to be necessarily less than unity, $\hat{\sigma}^2_{g_A} < (2v/s)$, giving an upper bound to the equilibrium genetic variance in terms of v and s. This was found to be true in the results obtained on computer using the exact recurrence relation.

Apart from the above, the moments of the allelic effects as well as those of the genotypic effects were studied analytically for a general m. These moments of the r-th order in the t-th generation are defined by

(1.10) allelic effects : $M_r(t) = \sum_{i=\infty}^{\infty} a^r i^r x_i(t)$

genotypic effects : $\mu_r(t) = \sum_i \sum_j a^r (i+j)^r x_i(t) x_j(t)$

Noting that for all i, $x_i(t) = x_{-i}(t)$ at each generation since optimum genotype is at the origin and $x_i(o) = x_{-i}(o)$ initially, the recurrence relationship between even order moments for the allelic effects was obtained. This gave a relationship between the change of variance of allelic effects at generation t, in terms of second and fourth moments as

(1.11) $\Delta M_2(t) = \frac{mva^2}{2} + s[M_2^2(t) - M_4(t)]$

At equilibrium, therefore,

(1.12) $\hat{M}_4 = \hat{M}_2^2 + \frac{mva^2}{2s}$

Further, since $\mu_2(t) = 2M_2(t)$ and $\mu_4(t) = 2M_4(t) + 6M_2^2(t)$, the corresponding relationship at genotypic level is found to be

$$(1.13) \qquad \Delta\mu_2(t) = mva^2 - s[\mu_4(t) - 2\mu_2^2(t)]$$

If we assume normality of genotypic values at equilibrium, $\hat{\mu}_4 = 3\hat{\mu}_2^2$ and we get

$$(1.14) \qquad \hat{\sigma}^2_{g_A} = \hat{\mu}_2 = \sqrt{\frac{mva^2}{s}} = \sqrt{\frac{\Delta\sigma_m^2}{s}}$$

where $\Delta\sigma_m^2 = mva^2$ is the effect of mutational change on the genotypic value at a locus. This is exactly what Kimura (1965) obtained for his model based on continuous distribution of allelic effects.

Now we consider the case when an equilibrium population shifts to a new environment where the optimum phenotype is shifted by $d\sigma_p$ units away from the origin. This would disturb the equilibrium status of the genotypic distribution which would gradually shift to a new optimum. The nature and change of genetic variability under such circumstances were analysed analytically by considering again the recurrence relationships of gene frequency changes and changes of moments, remembering that in this new environment, the fitness w_{ij} of A_iA_j would be proportional to $\exp[-s(a(i+j) - d\sigma_p)^2]$. This gives finally the recurrence relationship for the mean genotypic value in the population in two successive generations as

$$(1.15) \qquad \mu_1'(t) = \mu_1'(t-1) - s[\mu_3(t-1) + 2\mu_2(t-1)\{\mu_1'(t-1) - d\sigma_P\}]$$

At equilibrium, therefore, $\mu_1' = d\sigma_P$, the optimum genotypic value. The equilibrium distribution is symmetric around this optimum while in the transient states, substantial skewness might exist. For the genotypic variance, we finally obtain

$$(1.16) \qquad \Delta\sigma_{g_A}^2(t) = mva^2 + s[2\sigma_{g_A}^4(t) - \mu_4(t) + \mu_3(t)\{d\sigma_P - 2\mu_1'(t)\}]$$

At equilibrium, $\hat{\mu}_3 = 0$ due to the symmetry of the distribution and with the assumption of normality, we again have the same steady state genotypic variance given by (1.14) as we got when the optimum phenotype is taken at the origin. Thus, even if the optimum is shifted by a certain s.d. units away from the original mean, as long as the strength of selection (s) remains the same, the genotypic variability eventually returns to the original level although the genotypic distribution becomes now centered around the new optimum. At the transitory stage, however, it is difficult to assert analytically how the variance is altered.

We now consider two linked loci A and B with recombination value c at each of which an infinite number of possible allelic types can exist. We assume that allelic effects over loci are additive and therefore the contribution of the gametic effect

of gamete A_iB_j would be $a(i+j)$. The mutational scheme for m=1 is now a two-dimensional extension of the scheme given in Chakraborty and Nei (1982) and is shown in Figure 1.1. The gamete A_iB_j can mutate to $A_{i+1}B_j$, $A_{i-1}B_j$, A_iB_{j+1} or A_iB_{j-1} each with probability (v/2) whereas it does not mutate, having the same gametic effect with probability (1-2v). We effectively assume that mutations of type A_iB_j to $A_{i\pm1}B_{j\pm1}$ are negligible. The mean fitness of w_{ijkl} individuals with genotype $A_iA_jB_kB_l$ is proportional to $\exp[-sa^2(i+j+k+l)^2]$. The optimal value of the character is taken at zero with optimal genotype $A_oA_oB_oB_o$ having relative fitness as unity. Assuming the selection, recombination and mutation events to occur in the stated order, the recurrence relationship between the gametic frequencies in two successive generation is written down. Neglecting powers of s, v and sv, the mean fitness $\bar{w}_{AB}(t+1)$ at generation (t+1) is approximated as $\bar{w}_{AB}(t+1) \approx 1 - s\sigma^2_{g_{AB}}(t)$ where $\sigma^2_{g_{AB}}(t)$ is the genotypic variance at the t-th generation. The approximate recurrence relationship is found to be

$$
\begin{aligned}
(1.17)\qquad \bar{w}_{AB}(t+1)X_{ij}(t+1) = \sum_k\sum_l (1-2v-sa^2z^2)[&(1-c)X_{ij}(t)X_{kl}(t) \\
&+ c\,X_{il}(t)X_{kj}(t)] + \frac{v}{2}[(1-c)\{X_{i+1,j}(t)+X_{i-1,j}(t) \\
&+ X_{i,j+1}(t)+X_{i,j-1}(t)\} + c\{x_{i+1}(t)y_j(t) \\
&+ x_i(t)y_{j+1}(t)+x_{i-1}(t)y_j(t)+x_i(t)y_{j-1}(t)\}]
\end{aligned}
$$

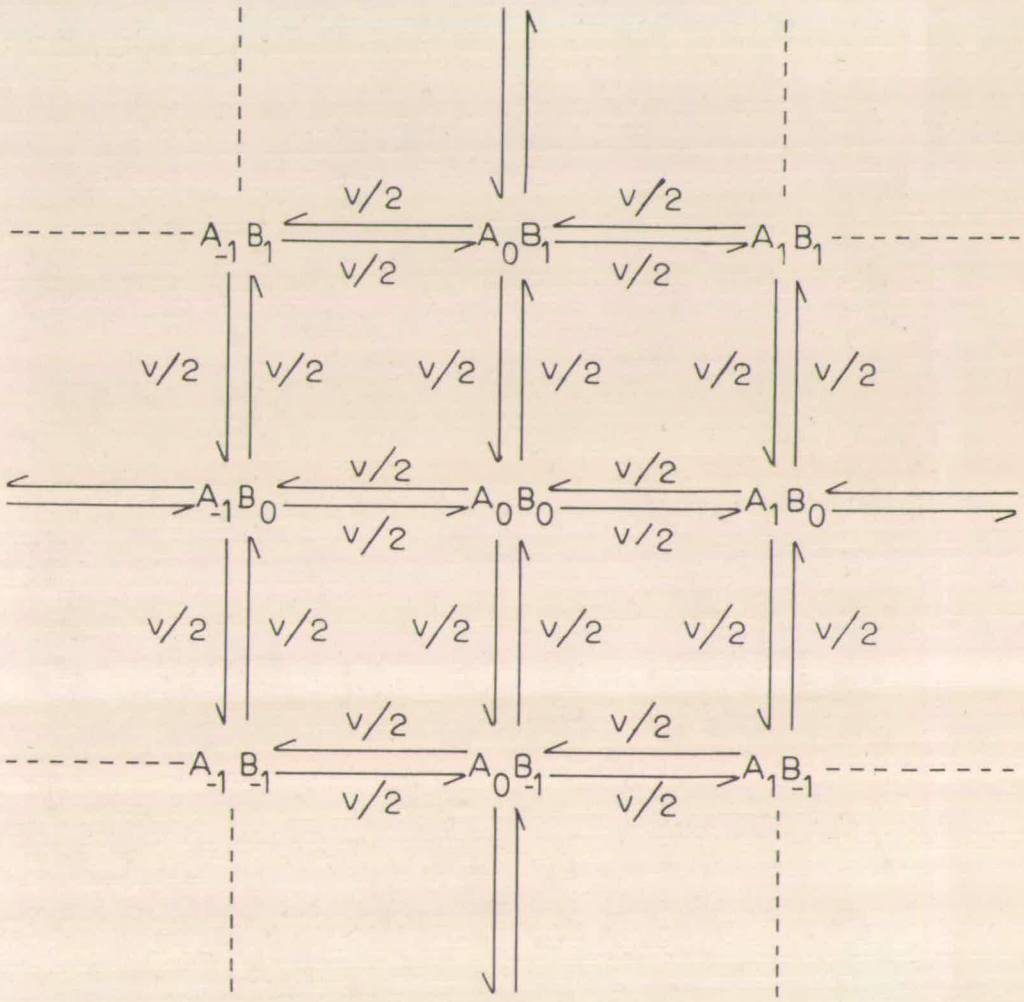

Fig. 1.1 Two loci allelic state model involving mutations from $A_i B_j$ to $A_i B_{j\pm1}$ or $A_{i\pm1} B_j$ with probability $v/2$ and from $A_i \overline{B}_j$ to $A_{i\pm1} B_{j\pm1}$ with negligible probability.

where $x_i(t) = \sum_j X_{ij}(t)$, $y_j(t) = \sum_i X_{ij}(t)$ are the gene frequencies of alleles A_i and A_j at the A and B locus respectively.

In general, analytic expressions for the recurrence relationship for the moments of the genotypic distribution are tedious to obtain. However, it is easy to see that if the population is at linkage equilibrium i.e. $D_{ij}(t) = X_{ij}(t) - x_i(t)y_j(t)$ for all i and j, the above recurrence relationship would yield the corresponding one locus recurrence relations already discussed. We therefore performed numerical calculations iteratively on the computer. First gametic frequencies in a particular generation were obtained using the recurrence relation for given values of s and v. We then obtained the distribution of genotypic values and its various statistical properties like mean, variance, skewness and Kurtosis by forming genotypes by random union of gametes. The distribution of allele frequencies at A and B locus in each generation were studied by a statistic of the form

$$(1.18) \qquad D^2(t) = \sum_{ij}\sum D_{ij}^2(t)/H_A(t)H_B(t)$$

where $H_A(t) = 1 - \sum_i x_i^2(t)$, $H_B(t) = 1 - \sum_j y_j^2(t)$ are the respective heterozygosities at A and B locus respectively.

Results and Discussion

When the initial population is considered as monomorphic with optimum at the origin, numerical results on the mean and within population variance indicate that the population mean remains at zero all through but the variance increases slowly from zero and attains, at equilibrium, a value determined solely by (v/s). This behaviour of within population variance over time is exhibited in Table 1.1.

TABLE 1.1 Within population variance as a function of time for different values of s and m with v = 0.001.

t	m = 1			m = 5		
	s			s		
	0.001	0.002	0.008	0.001	0.002	0.008
100	0.095	0.090	0.068	0.370	0.296	0.141
500	0.376	0.300	0.112	0.900	0.577	0.187
1,000	0.568	0.393	0.122	1.104	0.662	0.188
5,000	0.772	0.440	0.122	1.307	0.708	0.188
10,000	0.773	0.440	0.122	1.310	0.708	0.188
∞	0.774	0.440	0.122	1.311	0.708	0.188

It is apparent from this Table that for more intense selection, the variance during the transient stage as well as at equilibrium are lower as otherwise expected. Mutation creates

variability while selection eliminates it so that for intense selection its role is dominant. Also, the approach to equilibrium is found to be quicker for more intense selection. Compared to one-step mutation, the 5-step mutation case gives higher values of within population variance both for weak as well as for intense selection.

When we shift the optimum to 6 s.d. units away from the mean on the right, the transient behaviour of the process as it approaches the same equilibrium presents some interesting features shown in Table 1.2.

TABLE 1.2 Statistical properties of distribution of genotypic values as a function of time with new optimum at 6 s.d. units away from original optimum ($v = 0.001$, $s = 2v$, $m = 5$).

t	Mean	Variance	Skewness	Kurtosis
0	0.000	0.456	0.000	2.741
50	0.378	0.728	0.997	1.886
100	0.930	1.296	0.684	0.433
500	3.448	0.815	0.119	0.764
1,000	3.811	0.618	-0.001	1.563
∞	4.000	0.456	0.000	2.741

It is seen from Table 1.2 that the mean increases from zero to

four at equilibrium. The variance on the other hand increases, attains a maximum and decreases back to the original value. The most interesting feature is, however, noticed in the skewness. Initially, the distribution is symmetrical but as we advance in time, its symmetry is disturbed. It gets skewed initially and then slowly the skewness decreases, changes sign and finally the distribution becomes again symmetrical at equilibrium. The Kurtosis of the distribution also behaves in a similar fashion. Starting from a value very near to three initially, it declines to a value less than half but increases thereafter and restores the initial value at the equilibrium.

When we consider two populations, in one of which the same optimum holds but in the other it is at 6 s.d. units away from the optimum in the first population, the genetic differentiation between populations built up over a period of time is studied in terms of between population variability measured by B_t as against V_t denoting the within population variance at generation t. The behaviour of B_t, V_t and the ratio (B_t/V_t) as a function of time of divergence of the two populations, for $v = 0.001$, $s = 2v$ and $m = 5$ is shown in Figure 1.2. As against the behaviour of V_t already discussed, B_t increases slowly initially and then almost linearly until it approaches a plateau at equilibrium. The ratio (B_t/V_t) almost mimics the behaviour of B_t at least in the initial stages but it attains

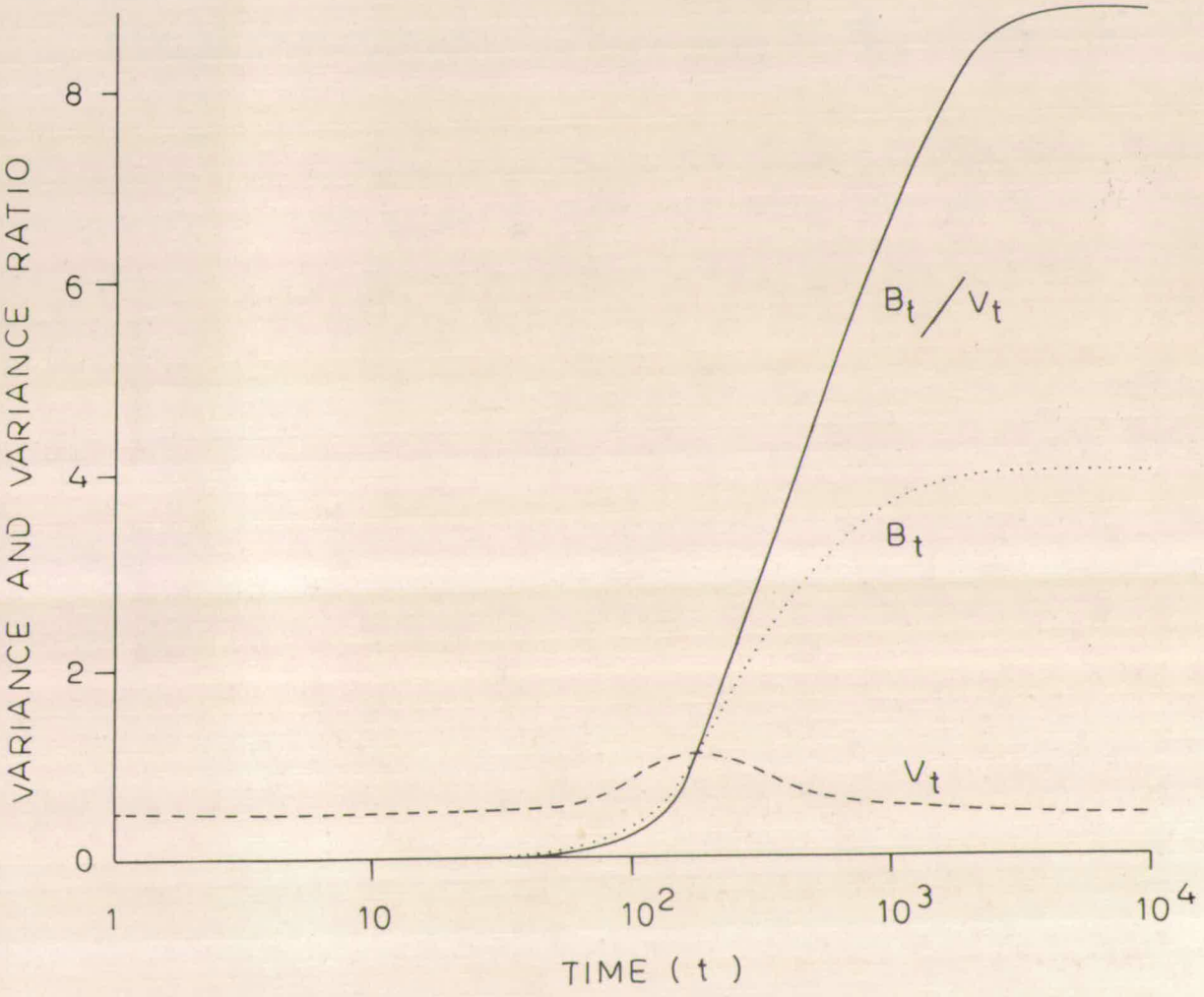

Fig. 1.2 Intra- (V_t) and inter- (B_t) population variance components along with the ratio B_t/V_t for different values of time (t) of divergence between two populations with $v = 0.001$, $s = 2v$ and $m = 5$.

a much higher value at equilibrium. This is because while V_t decreases, B_t increases as equilibrium is reached. In the initial stages, however, B_t and the ratio are almost the same because V_t has been increasing and reaching a maximum. After this stage, at which maximum V_t occurs, the quantities B_t and (B_t/V_t) diverge apart, increasing with time by different magnitudes.

Considering the joint effects of step-wise mutation and random drift, Chakraborty and Nei (1982) found that the ratio (B_t/V_t) increases linearly with time. In this paper, on the other hand, where step-wise mutation and centripetal selection are taken into account, this behaviour changes characteristically in that the ratio is no longer a monotone function of time. As such, this ratio can be helpful in testing the hypothesis of neutrality. Further, the ratio is also affected by the number of mutational steps. The more the number of mutational steps, the less is the ratio at equilibrium. This is because B_t reaches the same values at equilibrium irrespective of the value of m but V_t attains a higher value at equilibrium for a larger value of m. Also, a more intense selection leads to a higher value of the ratio all through the transient stage to equilibrium. The ratio (B_t/V_t) can thus be regarded as an index for determining the evolutionary forces under which a quantitative character changes over time.

Considering two loci with gametic selection, the results on V_t were obtained for the case when the initial population is monomorphic and the optimal genotype is at the origin. The results are presented in Table 1.3.

TABLE 1.3 Within population variance (V_t) as a function of time when the trait is controlled by two loci for different values of recombination fraction (c) with v=0.001, s=5v, m=1.

t	c		
	0.0	0.25	0.50
10	0.039	0.039	0.039
50	0.175	0.175	0.176
100	0.303	0.308	0.309
500	0.618	0.672	0.674
1,000	0.649	0.719	0.721
∞	0.662	0.722	0.724

It is apparent that the effect of linkage is not very pronounced in the initial stages but at equilibrium it is found to reduce the intra-population variance.

N.B. The work reported above pertains to a research carried out jointly with Dr. Ranajit Chakraborty at the Centre for Demographic and Population Genetics, University of Texas Health Science Centre, Houston, U.S.A. The following research paper incorporating these results

has been accepted for publication and is shortly to appear in print:

Narain, P. and R. Chakraborty (1984). Genetic differentiation of quantitative characters between populations or species II: Optimal selection in infinite populations. Theoretical Population Biology, 25 (In press).

II. PROGENY TESTING WITH AUXILIARY TRAITS

As already discussed in Narain (1979) sire index for a trait like milk production could be constructed by using the technique of phenotypic index introduced in Narain and Mishra (1975) wherein observations on milk could be expressed as deviation from their expected values predicted with the help of an auxiliary trait such as fat percentage. However, it may be more profitable if instead of using the auxiliary trait for necessary correction, we may use it in combination with the main trait for progeny testing the sire. Searle (1978) considered this aspect briefly while dealing with progeny testing using indirect selection. But he restricted himself to one auxiliary trait. In this study, we develop a general theory of progeny testing with several auxiliary traits.

Consider k auxiliary traits $x_i (i=1,2,\ldots,k)$ related to the main trait y and use the parameters of Narain and Mishra (1975), expressed in terms of vectors and matrices as

$\underset{\sim}{r}_o^T = (R^p_{o1}, R^p_{o2}, \ldots, R^p_{ok})$, $\underset{\sim}{C}^T = (C_1, C_2, \ldots, C_k)$, $\underset{\sim}{R} = ((R^p_{ij}))$,

$\underset{\sim}{H} = ((R^g_{ij} h_i h_j / h_o^2))$, $a_o = (4-h_o^2)/h_o^2$, $C_i = R^g_{oi}(h_i/h_o)$.

Let the average of the phenotypic values of n progenies of the sire for traits x_i and y be denoted by $\bar{D}(x_i)$ and $\bar{D}(y)$ respectively for $i=1,2,\ldots,k$. The statistical model in such a case

can be set up as follows:

$$(2.1) \qquad E[A(y)] = b_o\bar{D}(y) + \sum_{i=1}^{k} b_i\bar{D}(x_i)$$

where E[A(y)] stands for the expectation of the breeding value of sire for y and $b_o, b_1, b_2, \ldots, b_k$ are coefficients to be determined such that the multiple correlation coefficient between A(y) and E[A(y)] is maximised. This strategy leads to the solutions

$$(2.2) \qquad b_o = 2n(1-\underline{\beta}^T \underset{\sim}{Q}^{-1} \underline{C})/(\alpha - \underline{\beta}^T \underset{\sim}{Q}^{-1} \underline{\beta})$$

$$(2.3) \qquad \underline{b} = 2n\underset{\sim}{Q}^{-1}[(\alpha\underline{C}-\underline{\beta}) - \underline{C}\underline{\beta}^T \underset{\sim}{Q}^{-1} \underline{\beta} + \underline{\beta}\underline{\beta}^T \underset{\sim}{Q}^{-1} \underline{C}]/(\alpha - \underline{\beta}^T \underset{\sim}{Q}^{-1} \underline{\beta})$$

where $\underline{\beta} = [(1+a_o)\underline{r}_o + (n-1)\underline{C}]$, $\underset{\sim}{Q} = [(1+a_o)\underset{\sim}{R} + (n-1)\underset{\sim}{H}] = (n\underset{\sim}{I} + \underset{\sim}{A})\underset{\sim}{H}$, $\underset{\sim}{A} = (4\underset{\sim}{R}\underset{\sim}{H}^{-1} - h_o^2\underset{\sim}{I})/h_o^2$ and $\alpha = (n+a_o)$.

The accuracy of the progeny test measured in terms of the maximised multiple correlation coefficient (h^*_{pr}) is then given by

$$(2.4) \qquad h^{*2}_{pr} = h^2_{pr}\left[\frac{1-(n+a_o)d_1 + 2(1+a_o)d_3 - (1+a_o)^2(d_1d_2 - d_3^2)}{1-(n+a_o)d_1 + 2(1+a_o)d_3 - (1+a_o)^2 d_2/(n+a_o)}\right]$$

where $h^2_{pr} = n/(n+a_o)$, $d_1 = \underline{C}^T \underset{\sim}{Q}^{-1} \underline{C}$, $d_2 = (\underline{C}-\underline{r}_o)^T \underset{\sim}{Q}^{-1} (\underline{C}-\underline{r}_o)$ and

$d_3 = \underset{\sim}{C}^T \underset{\sim}{Q}^{-1} (\underset{\sim}{C} - \underset{\sim}{r}_o)$.

It is interesting to note that when $\underset{\sim}{C} = \underset{\sim}{r}_o$ i.e. $R^g_{oi}(h_i/h_o) = R^p_{oi}$ for $i=1,2,\ldots,k$, d_2 and d_3 reduce to zeros and $h^*_{pr} = h_{pr}$ indicating that the accuracy of the progeny test is the same as if there were no auxiliary traits. For the limiting case when n tends to be infinitely large, we see that $\underset{\sim}{Q}^{-1} = \underset{\sim}{H}^{-1}(n\underset{\sim}{I} + \underset{\sim}{A})^{-1} = \frac{\underset{\sim}{H}^{-1}}{n}[\underset{\sim}{I} - \frac{(\underset{\sim}{I}/n) + \underset{\sim}{A}^{-1}}{n}]$ tends to zero as $n \to \infty$ so that d_1, d_2 and d_3 also tend to zero and since h^2_{pr} tends to 1, so does h^{*2}_{pr}. In fact, with very large number of progeny available for a sire, we expect to know with complete accuracy, the breeding values of sire for each of the main and auxiliary traits. The index will then help very little as we would be virtually selecting on the basis of the main trait only. This was also noted by Robertson (1961) while discussing selection for several traits.

When the genetic correlation coefficients R^g_{oi}'s are all zero, $\underset{\sim}{C} = \underset{\sim}{o}$, $\underset{\sim}{H} = \underset{\sim}{o}$, $\underset{\sim}{Q} = (1+a_o)\underset{\sim}{R}$ giving $d_1 = d_3 = o$ and $d_2 = \underset{\sim}{r}_o^T \underset{\sim}{R}^{-1} \underset{\sim}{r}_o/(1+a_o)$. h^{*2}_{pr} then reduces to

$$(2.5) \qquad h^{*2}_{pr} = n/(n+a^*_o)$$

where $a^*_o = (4 - h^{*2}_o)/h^{*2}_o$ and $h^{*2}_o = h^2_o/(1 - \underset{\sim}{r}_o^T \underset{\sim}{R}^{-1} \underset{\sim}{r}_o)$ is the

heritability of the main trait y corrected for auxiliary traits $x_1, x_2, \ldots, x_k$ which are correlated with the main trait at the environmental level only, as given in Narain and Mishra (1975) and used in Narain (1979) while proposing a new sire index for milk production. This means that when all the auxiliary traits are related to the main trait only at the environmental level, the accuracy of the progeny test is of the same form as that without any auxiliary trait but the main trait y is replaced by the phenotypic index.

A generalised index for the breeding value of the sire is obtained by substituting the optimum values of $b_o, b_1, \ldots, b_k$ in E[A(y)], giving

$$(2.6) \qquad S_I = 2nW_k[\bar{D}(y)-(\frac{1+a_o}{1-\underline{\beta}^T \underset{\sim}{Q}^{-1}\underline{C}})\{(\underline{r}_o-\underline{C})^T-\underline{\beta}^T \underset{\sim}{Q}^{-1}(\underline{C}\underline{r}_o^T-\underline{r}_o\underline{C}^T)\}\underset{\sim}{Q}^{-1}\underline{\bar{D}}(x)]$$

where $W_k = (1-\underline{\beta}^T \underset{\sim}{Q}^{-1}\underline{C})/(\alpha-\underline{\beta}^T \underset{\sim}{Q}^{-1}\underline{\beta})$ and $\underline{\bar{D}}(x^T) = [\bar{D}(x_1), \bar{D}(x_2), \ldots, \bar{D}(x_k)]$. As expected, when either $C_i = R^p_{oi}$ or $R^p_{oi} = R^g_{oi} = o$ for each $i=1,2,\ldots,k$, the sire index reduces to $[2n/(n+a_o)]\bar{D}(y)$, the usual sire index based on daughter's average making allowance for finite number of daughters. But when only the genetic correlation coefficients R^g_{oi}'s and R^g_{ij}'s are all zero, $\underline{C} = \underline{o}$, $\underset{\sim}{H} = \underset{\sim}{o}$ so that $\underline{\beta} = (1+a_o)\underline{r}_o$, $\underset{\sim}{Q} = (1+a_o)\underset{\sim}{R}$, giving the index as

$$(2.7) \qquad S_I = \frac{2n}{(n+a_o^*)}[\bar{D}(y) - \underline{r}_o^T \underline{R}^{-1} \underline{\bar{D}}(x)]$$

For k=1, this is the same index as given in Narain (1979) where the auxiliary trait is correlated with the main trait only at the environmental level. When $n \to \infty$, $\underset{\sim}{Q}^{-1} \to \underset{\sim}{Q}$ and the sire index tends to twice the simple daughter average for trait y as it should.

For studying the behaviour of the accuracy of the progeny test with auxiliary traits due to variations in the various parameters we restrict ourselves to the situation when k=1. In this case, the vectors $\underline{r}_o$ and $\underline{C}$ reduce to scalars R_{o1}^p and $C_1 = R_{o1}^g(h_1/h_o)$ respectively whereas matrices $\underset{\sim}{R}$ and $\underset{\sim}{H}$ reduce to scalars 1 and (h_1^2/h_o^2) respectively. The matrix Q becomes the scalar $(h_1^2/h_o^2)(n+a_1)$ and the vector $\underline{\beta}$ reduces to $(1+a_o)R_{o1}^p + (n-1)C_1$. On simplification, we get an expression of accuracy as

$$(2.8) \qquad h_{pr(k=1)}^{*2} = h_{pr}^2 \left[\frac{K}{K-(C_1-R_{o1}^p)^2(1+a_o)/(n+a_o)}\right]$$

where $K = (C_1-R_{o1}^p)^2 + (1-R_{o1}^{p2}) + (n-1)(1-R_{o1}^{g2})/(1+a_1)$. It is obvious that $h_{pr(k=1)}^{*2}$ is always greater than or equal to h_{pr}^2. The accuracy of the progeny test is thus improved due to inclusion of the auxiliary trait. However, since it is now a function of five parameters h_o^2, h_1^2, R_{o1}^p, R_{o1}^g and n, we can only

numerically study its behaviour. In the first instance, we fix n=20 and $h_1^2 = 0.2$, the heritability of the auxiliary trait and plot the accuracy as a function of h_o^2 for different combinations of R_{o1}^p and R_{o1}^g. The results are shown graphically in Figures 2.1 and 2.2. The accuracy with auxiliary traits is always more than that without it for all values of h_o^2 and at every combination of R_{o1}^p and R_{o1}^g. When R_{o1}^p and R_{o1}^g are of opposite signs, the gain in accuracy is more than in the other cases and it is substantial at lower values of h_o^2. When the auxiliary trait is only genetically correlated with the main trait, the accuracy is more than when it is only phenotypically correlated for values of $h_o^2 < 0.2$. This phenomenon is reversed when $h_o^2 > 0.2$. The effect of h_1^2 is shown in Figure 2.3 where we fix n=20 and the combination of (R_{o1}^p, R_{o1}^g) at $(\mp 0.5, \pm 0.5)$. The curve for $h_1^2 = 0.5$ is higher than that for $h_1^2 = 0.01$ for values of $h_o^2 \leq 0.5$, the gain being considerable at lower values of h_o^2. Beyond $h_o^2 = 0.5$, the two curves seem to coincide all along. The most important finding about the accuracy of the progeny test is shown in Figure 2.4 wherein it is plotted against the group size (n) for $h_1^2 = 0.2$ and $h_o^2 = 0.2$. Clearly, the accuracy for all combinations of (R_{o1}^p, R_{o1}^g) increases sharply initially as we go from n=1 to n=30 but as n becomes larger and larger, all the curves tend to approach unity. It

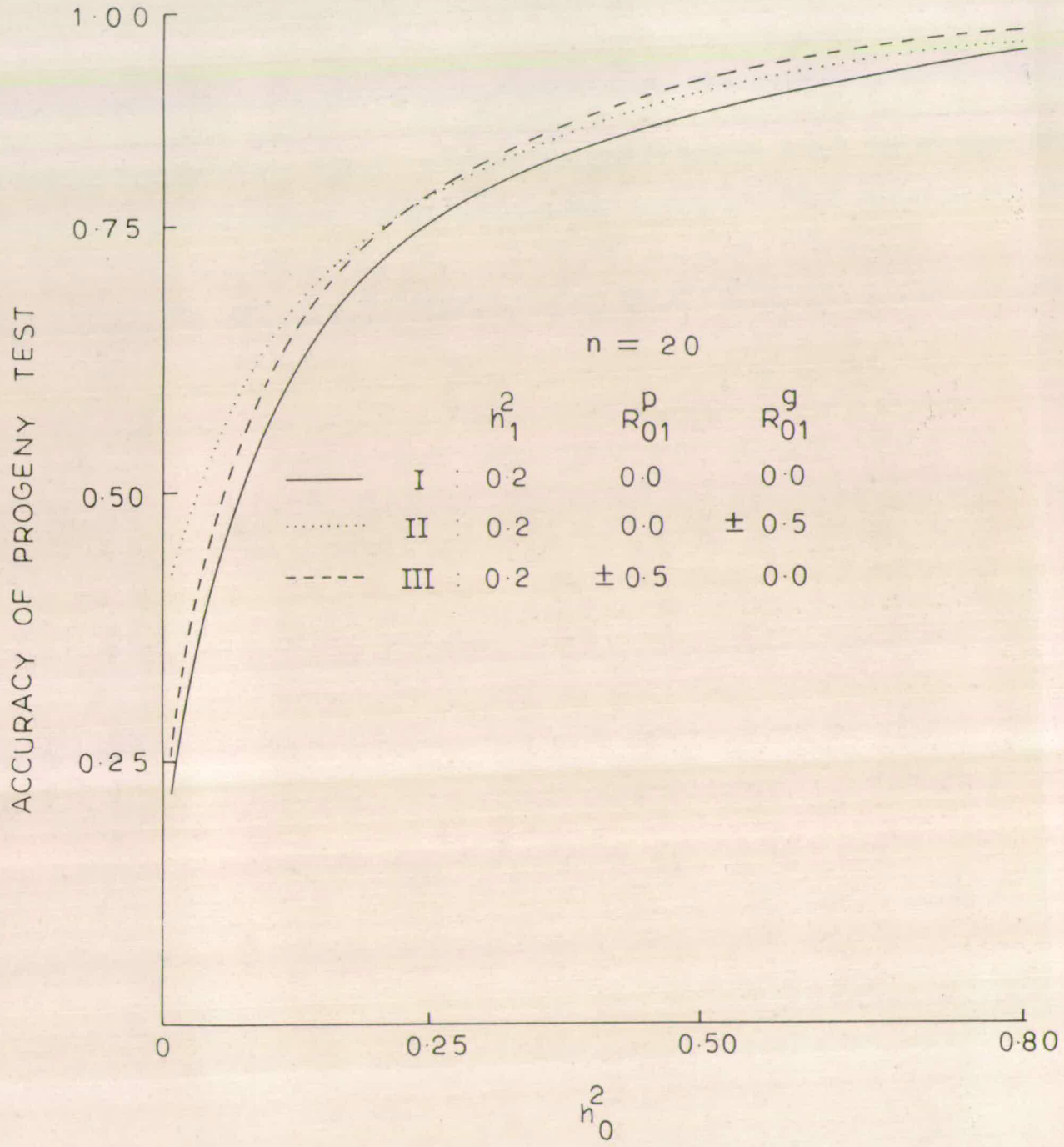

Fig. 2.1 Accuracy of the progeny test with one auxiliary trait for different values of h_o^2, (R_{o1}^p, R_{o1}^g) = (0.0, ± 0.5) and (± 0.5, 0.0) with n = 20 and $h_1^2 = 0.2$.

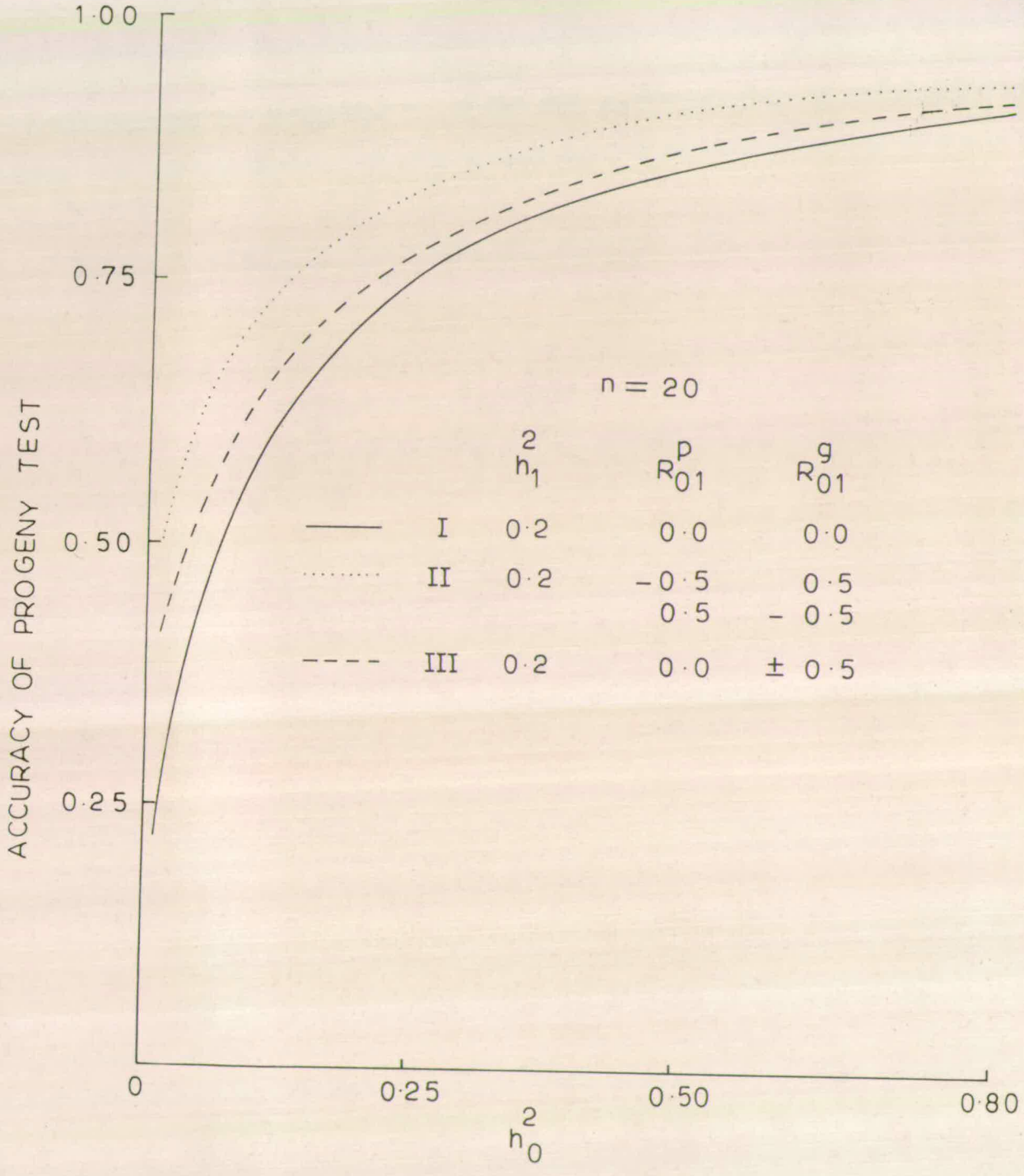

Fig. 2.2 Accuracy of the progeny test with one auxiliary trait for different values of h_o^2, $(R_{o1}^p, R_{o1}^g) = (\mp 0.5, \pm 0.5)$ and $(0.0, \pm 0.5)$ with $n = 20$ and $h_1^2 = 0.2$.

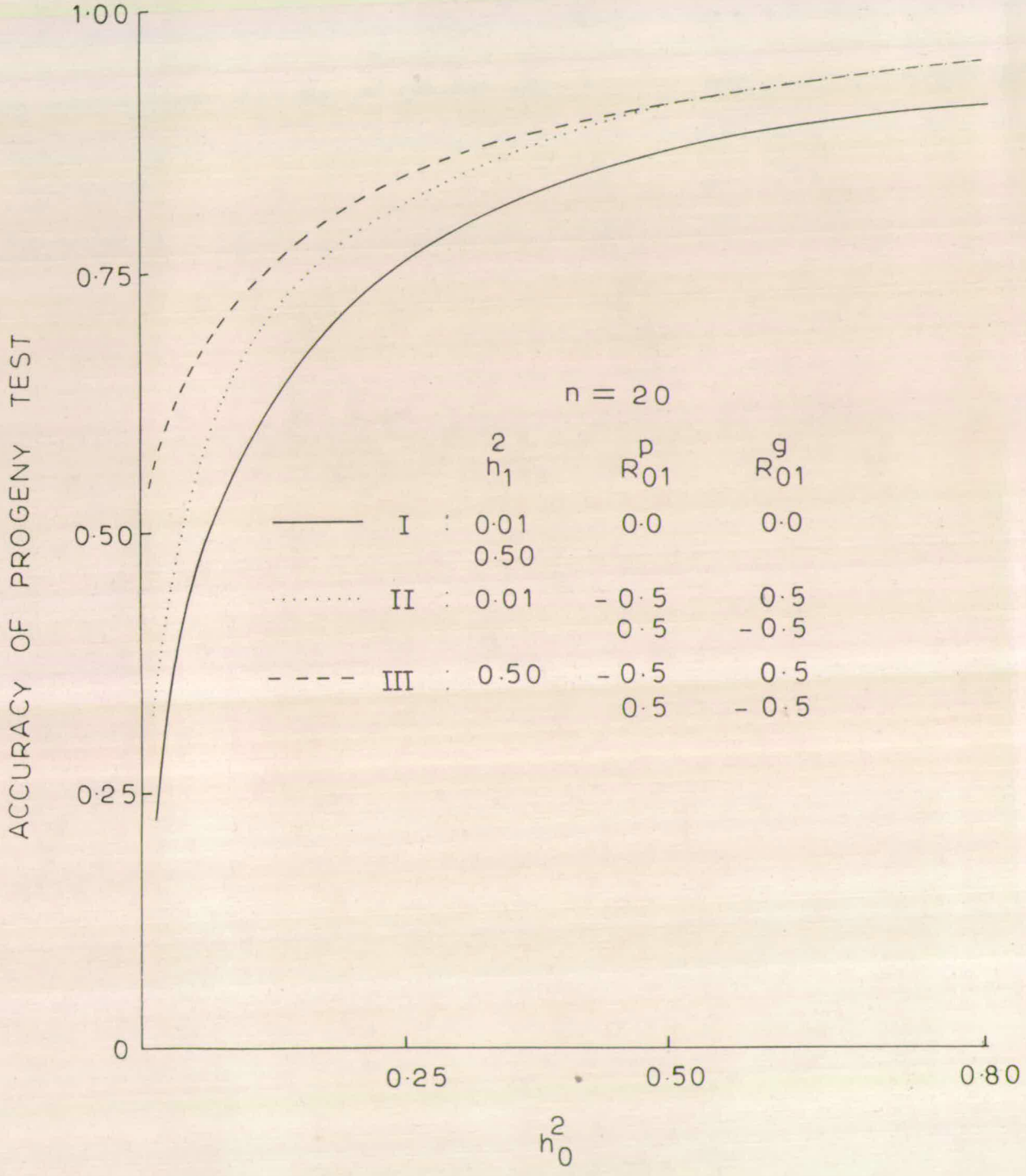

Fig. 2.3 Accuracy of the progeny test with one auxiliary trait for different values of h_o^2 and h_1^2 with $n = 20$ and $(R_{o1}^p, R_{o1}^g) = (\mp 0.5, \pm 0.5)$.

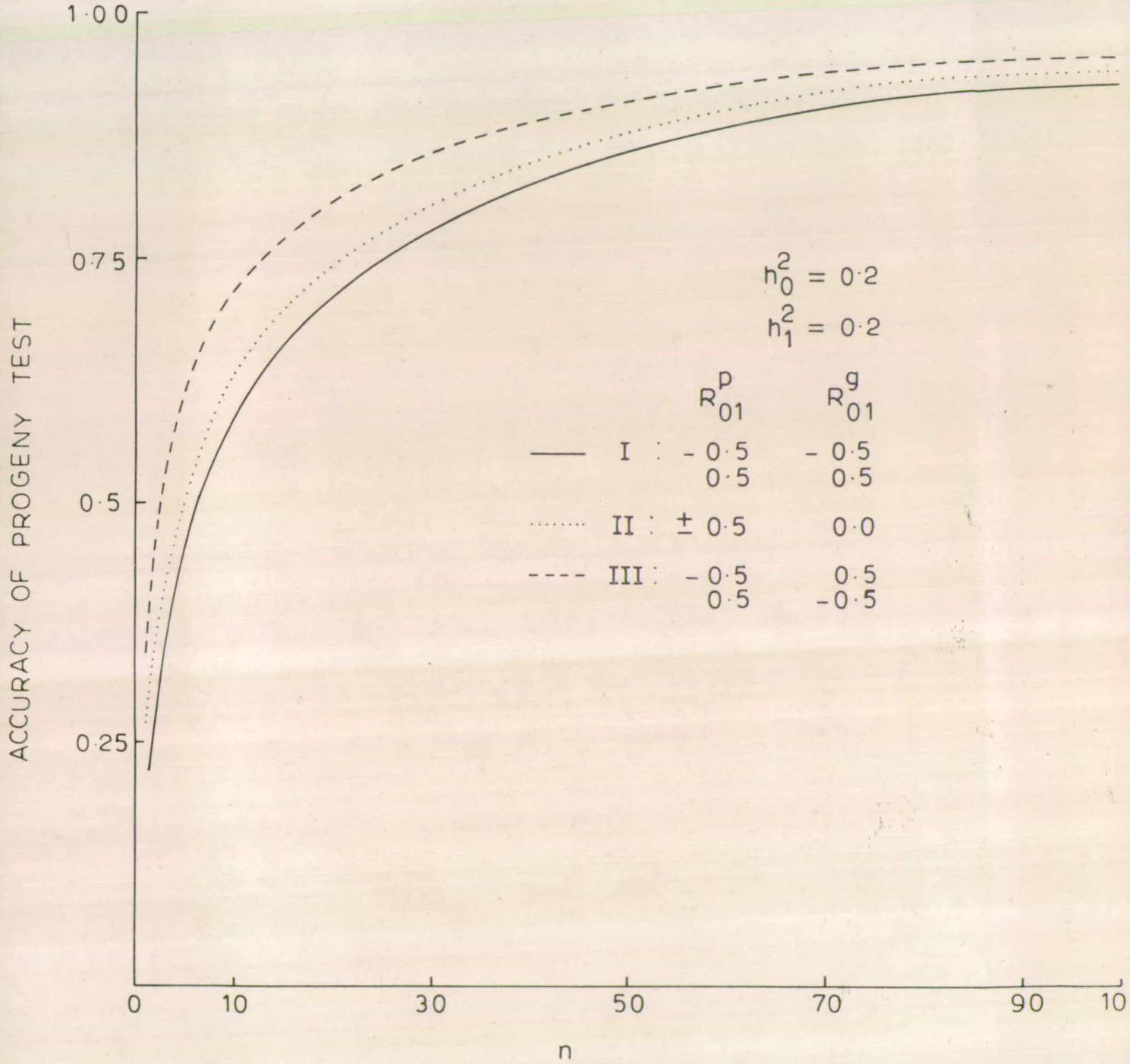

Fig. 2.4 Accuracy of the progeny test with one auxiliary trait for different values of group size (n), $(R_{01}^p, R_{01}^g) = (\mp 0.5, \mp 0.5)$, $(\pm 0.5, 0.0)$ and $(\mp 0.5, \pm 0.5)$ with $h_o^2 = h_1^2 = 0.2$.

is important to find that when R^{p}_{o1} and R^{g}_{o1} are of opposite signs, the curve is the highest.

The above results on the behaviour of the accuracy with variations in progeny group size indicate that if we fix a desired level of accuracy say 0.70 or more, we would need a smaller number of progeny for the test if a suitable auxiliary trait is taken into account compared to the case without any auxiliary trait. This can be numerically studied by solving the expression for h^{*}_{pr} for n. This gives a quadratic in n with two roots of which the positive root is found to depend on $h^{2}_{o}, h^{2}_{1}, R^{p}_{o1}, R^{g}_{o1}$ and of course h^{*}_{pr}. Numerically, we obtain the results as illustrated in Tables 2.1 and 2.2.

TABLE 2.1 Number of progeny required for a pre-assigned value of accuracy when $h^{2}_{1} = 0.2$.

		$h^{2}_{o} = 0.1$				$h^{2}_{o} = 0.3$			
	R^{p}_{o1}	0.0	$\pm$0.5	0.0	$\mp$0.5	0.0	$\pm$0.5	0.0	$\mp$0.5
	R^{g}_{o1}	0.0	0.0	$\pm$0.5	$\pm$0.5	0.0	0.0	$\pm$0.5	$\pm$0.5
	0.50	13	11	9	5	4	3	3	2
	0.55	17	14	12	6	5	4	5	2
	0.60	22	19	16	9	7	5	6	3
	0.65	28	25	22	12	9	7	8	4
	0.70	37	34	30	18	12	10	10	6
h^{*}_{pr}	0.75	50	46	42	26	16	13	14	8
	0.80	69	65	60	41	22	19	20	12
	0.85	101	97	91	70	32	28	29	19
	0.90	166	161	154	131	52	48	49	36
	0.95	361	356	348	323	114	109	110	94

Table 2.1 shows the effect of correlations. It is found that in all the cases, the use of auxiliary traits reduces the number of progeny required for a pre-assigned value of the accuracy of progeny test below the one required if no auxiliary trait is used. Also, this gain of reduced number is substantial when R^{p}_{o1} and R^{g}_{o1} are of opposite signs and around 0.5 in magnitude. For combinations $(\pm 0.5, 0)$ and $(0, \pm 0.5)$, it is found that when $h^2_o = 0.1$, smaller n is obtained when the correlation is only at genetic level than when it is only at the phenotypic level. This trend is however reversed when $h^2_o = 0.3$ but now the differences are only marginal.

TABLE 2.2 Number of progeny required for a pre-assigned value of accuracy when $R^{p}_{o1} = \mp 0.5$, $R^{g}_{o1} = \pm 0.5$.

	$h^2_o = 0.1$			$h^2_o = 0.3$		
	h^2_1			h^2_1		
h^*_{pr}	0.01	0.2	0.5	0.01	0.2	0.5
0.50	8	5	3	3	2	1
0.55	11	6	5	4	2	2
0.60	14	9	7	5	3	3
0.65	18	12	11	6	4	3
0.70	24	18	17	8	6	5
0.75	32	26	27	11	8	7
0.80	45	41	44	15	12	12
0.85	67	70	75	21	19	20
0.90	113	131	139	35	36	39
0.95	265	323	333	80	94	100

Table 2.2 shows the effect of the heritability of auxiliary trait. It is apparent that when $h_o^2 = 0.1$ (0.3) and the accuracy aimed at is above 80 (90) per cent, a lower value of $h_1^2 = 0.01$ is to be preferred. There also appears to be a range of the values of accuracy, depending on h_o^2, in which n does not vary much due to a change in the value of h_1^2, the heritability of the auxiliary trait.

N.B. The work reported above was presented in the Session on 'Animal Genetics and Breeding' during the XV International Congress of Genetics held at New Delhi, India from December 12 to 21, 1983. The following research paper incorporating these results has since been submitted for publication:

Narain, P. (1984). Progeny testing with auxiliary traits. Biometrics (Submitted).

REFERENCES

CHAKRABORTY, R. and NEI, M. (1982). Genetic differentiation of quantitative characters between population or species. I. Mutation and random genetic drift. Genet. Res., Camb. 39: 303-314.

KIMURA, M. (1965). A stochastic model concerning the maintenance of genetic variability in quantitative characters. Proc. Natl. Acad. Sci. U.S., 54: 731-736.

LANDE, R. (1976). Natural selection and random genetic drift in phenotypic evolution. Evolution, 30: 314-334.

ROBERTSON, A. (1956). The effect of selection against extreme deviants based on deviation or on homozygosis. J. Genet., 54: 236-248.

ROBERTSON, A. (1961). Selection for several characters. Lectures of II International Summer School on 'Scientific Problems of Breeding Systems and Breeding Plans of Domestic Animals'. Schriftenreihe des Max-Planck Instituts für Tierzucht und Tierernäh-rung, Mariensee. Special Vol., 213-224.

SEARLE, S.R. (1978). The value of indirect selection. II. Progeny testing. Theor. Appl. Genet. 51: 289-296.

WRIGHT, S. (1937). The distribution of gene frequencies in populations. Proc. Natl. Acad. Sci. U.S., 23: 307-320.

A P P E N D I X

TWENTY - THREE REPRINTS

Reprinted from the JOURNAL OF GENETICS Vol. 59, No. 3, June 1966.
Pages 254—266

1

HOMOZYGOSITY IN A SELFED POPULATION WITH AN ARBITRARY NUMBER OF LINKED LOCI

BY PREM NARAIN

Institute of Agricultural Research Statistics, New Delhi

INTRODUCTION

Quantitative measures of the intensity of inbreeding and degrees of relationship under various systems of mating were first given by Wright (1921) with the aid of path coefficients. The work of Malécot (1948) resulted in essentially the same formulae as that of Wright but his approach was to make use of the probabilities of genes being identical by descent at a locus, the coefficient of inbreeding F of an individual being defined as the probability that the two genes possessed by that individual at a locus are identical by descent. This could be extended directly to cover cases involving more than one locus provided the loci in question are not linked. The exact effect of linkage on the rate of inbreeding, however, did not receive attention until Rajagopalan (1958) studied its effect on the homozygosity of a selfed population using the generation matrix method adopted by Fisher (1949). His study was, however, not general in that it considered only two linked loci. Schnell (1961) considered for the first time, the probabilities of genes being identical by descent with respect to a given set of linked loci and gave generalised concepts of coefficient of inbreeding and panmictic index as inbreeding function (ϕ) and panmictic function (π). But he did not discuss any recurrence relation for a system of mating which would indicate how the homozygosity in a population increases with generation when linkage is operating. In this paper, this has been investigated when the population is inbred by selfing. The study takes into account an arbitrary number of linked loci to give a generalised treatment of the problem.

GENERALISATION OF "COEFFICIENT OF RELATIONSHIP"

Consider two individuals X and Y having genotypes $\frac{a_1 a_2 \ldots\ldots a_r}{b_1 b_2 \ldots\ldots b_r}$ and $\frac{c_1 c_2 \ldots\ldots c_r}{d_1 d_2 \ldots\ldots d_r}$ respectively where r is the number of loci and the horizontal line indicates that the genes above it lie on one chromosome and those below it lie on the other homologous chromosome. Taking only the ith locus into consideration, the coefficient of relationship between X and Y can be defined as

$$\rho_i^{XY}=\tfrac{1}{4}[P(a_i=c_i)+P(a_i=d_i)+P(b_i=c_i)+P(b_i=d_i)], \qquad (1)$$

where $P(a_i=c_i)$ denotes the probability that a random gene a_i from X is identical by descent with a random gene c_i from Y at the ith locus.

Taking two loci i and j with a recombination value p_{ij} between them, this coefficient of relationship can be defined as

$$\begin{aligned}\rho_{ij}^{XY} = \frac{1-p_{ij}}{4}\Big[&P(a_i=c_i;\ a_j=c_j) + P(b_i=c_i;\ b_j=c_j)\\ &+P(a_i=d_i;\ a_j=d_j) + P(b_i=d_i;\ b_j=d_j)\Big]\\ + \frac{p_{ij}}{4}\Big[&P(a_i=c_i;\ b_j=d_j) + P(b_i=c_i;\ a_j=d_j)\\ &+P(a_i=d_i;\ b_j=c_j) + P(b_i=d_i;\ a_j=c_j)\Big]\\ +\frac{p_{ij}(1-p_{ij})}{4}\Big[&P(a_i=c_i;\ b_j=c_j) + P(b_i=c_i;\ a_j=c_j)\\ &+P(a_i=c_i;\ a_j=d_j) + P(b_i=c_i;\ b_j=d_j)\\ &+P(a_i=d_i;\ a_j=c_j) + P(b_i=d_i;\ b_j=c_j)\\ &+P(a_i=d_i;\ b_j=d_j) + P(b_i=d_i;\ a_j=d_j)\Big],\end{aligned} \tag{2}$$

where $P(a_i=c_i;\ a_j=c_j)$ denotes the probability that a random gene a_i from X is identical by descent with a random gene c_i from Y at the ith locus and a random gene a_j from X is identical by descent with a random gene c_j from Y at the jth locus. Similarly ρ_{ijk}^{XY} in terms of probabilities $P(a_i=c_i;\ a_j=c_j;\ a_k=c_k)$ can be defined and finally $\rho_{123\ldots r}^{XY}$ can be defined in terms of probabilities $P\ (a_1=c_1;\ a_2=c_2;\ \ldots\ldots a_r=c_r)$.

As noted by Schnell (1961), new recombination values are, however, to be introduced when $r>3$ to enable the gametic frequencies to be expressed as linear functions of recombination values. With four loci, 1, 2, 3 and 4 for instance, there are six recombination values, $p_{12}, p_{23}, p_{34}, p_{13}, p_{14}$ and p_{24}. A new recombination value p_{1234} is required to be introduced measuring recombination between the segments corresponding to 1st and 2nd loci and 3rd and 4th loci. The following relations exist between these recombination parameters as given by Geiringer (1944):-

$$\left.\begin{aligned} p_{13} &= p_{12}+p_{23}-2c\,p_{12}\cdot p_{23},\\ p_{24} &= p_{23}+p_{34}-2c\,p_{23}\cdot p_{34},\\ p_{14} &= p_{12}+p_{23}+p_{34}-2c\,(p_{12}\cdot p_{23}+p_{23}\cdot p_{34}+p_{12}\cdot p_{34})+4c\,p_{12}\cdot p_{23}\cdot p_{34},\\ p_{1234} &= p_{12}+p_{34}-2c\,p_{12}\cdot p_{34},\end{aligned}\right\} \tag{3}$$

where c is the coefficient of coincidence.

When c is assumed to be unity, the expression for ρ_{1234}^{XY} involves p_{12}, p_{23} and p_{34} and the various compound probabilities.

INBREEDING FUNCTION AND PANMICTIC FUNCTION

As introduced by Schnell (1961), the inbreeding function ϕ of an individual is defined as the probability that the two gametes that produced the individual contain genes which are identical by descent regarding a given set of loci. Thus, with three loci, for instance,

$$\left.\begin{aligned} \phi^{X}_{ijl} &= P\begin{pmatrix} a_i=b_i \\ a_j=b_j \\ a_l=b_l \end{pmatrix} \\ \phi^{Y}_{ijl} &= P\begin{pmatrix} c_i=d_i \\ c_j=d_j \\ c_l=d_l \end{pmatrix}. \end{aligned}\right\} \tag{4}$$

Also, the function of inbreeding $\phi^{X\times Y}$ of an offspring resulting from the mating of two individuals X and Y is given by, for a given set of loci,

$$\phi^{X\times Y}_{ijl} = \rho^{XY}_{ijl}. \tag{5}$$

The panmictic function π of an individual is defined as the probability that the two gametes producing the individual contain genes which are unlike by descent for a given set of loci. The relations between the two set of functions are, for a given set, say three loci, given by

$$\left.\begin{aligned} \pi_{ijl} &= 1-\phi_i-\phi_j-\phi_l+\phi_{ij}+\phi_{jl}+\phi_{il}-\phi_{ijl} \\ \phi_{ijl} &= 1-\pi_i-\pi_j-\pi_l+\pi_{ij}+\pi_{jl}+\pi_{il}-\pi_{ijl}. \end{aligned}\right\} \tag{6}$$

As pointed out by Schnell (1961), ϕ and π are particular cases of a more general quantity ξ, the probability that the genes are identical by descent with respect to a given set of loci and also unlike by descent with respect to the remainder of the loci. Thus with three loci, under consideration, $\xi_{ijl\,(jl)}$ means the probability that the genes are identical by descent at the jth and lth loci but unlike by descent at the ith locus. Obviously then, we have

$$\pi_{ijl}+\xi_{ijl(i)}+\xi_{ijl(j)}+\xi_{ijl(l)}+\xi_{ijl(ij)}+\xi_{ijl(jl)}+\xi_{ijl(il)}+\phi_{ijl}=1. \tag{7}$$

We have also, the relations given by

$$\left.\begin{aligned} \pi_{ij} &= \xi_{ijl(l)}+\pi_{ijl} \\ \pi_{jl} &= \xi_{ijl(i)}+\pi_{ijl} \\ \pi_{il} &= \xi_{ijl(j)}+\pi_{ijl} \\ \pi_i &= \xi_{ijl(jl)}+\pi_{ij}+\pi_{il}-\pi_{ijl} \\ \pi_j &= \xi_{ijl(il)}+\pi_{jl}+\pi_{ij}-\pi_{ijl} \\ \pi_l &= \xi_{ijl(ij)}+\pi_{il}+\pi_{jl}-\pi_{ijl}. \end{aligned}\right\} \tag{8}$$

Similar relations hold with ϕ also.

Recurrence Relations

(a) *Two loci*

Considering only the ith locus, we have already the recurrence relations given in Kempthorne (1957),

$$\left.\begin{aligned} \phi_i^{(n+1)} &= \tfrac{1}{2}\left(1+\phi_i^{(n)}\right) \\ \pi_i^{(n+1)} &= \tfrac{1}{2}\pi_i^{(n)}, \end{aligned}\right\} \tag{9}$$

where $\pi_i = 1-\phi_i$ and $\pi_i^{(n)}$ etc. denotes the value of the function in the nth generation.

Comparing X with itself, taking into consideration the two loci, we have from (5) and (2)

$$\phi_{ij}^{X\times X} = \rho_{ij}^{XX}$$

$$= \frac{p_{ij}^2+(1-p_{ij})^2}{2}\left[1+\phi_{ij}^{X}\right] + p_{ij}(1-p_{ij})\left[\phi_i^X+\phi_j^X\right].$$

Replacing p_{ij} by $\lambda_{ij} = 1-2p_{ij}$, the linkage value introduced by Schnell (1961), we get

$$\phi_{ij}^{(n+1)} = \tfrac{1}{4}\left(1+\lambda_{ij}^2\right)\left(1+\phi_{ij}^{(n)}\right) + \tfrac{1}{4}\left(1-\lambda_{ij}^2\right)\left(\phi_i^{(n)}+\phi_j^{(n)}\right). \tag{10}$$

With the help of the relations

$$\left.\begin{aligned}\phi_{ij} &= 1-\pi_i-\pi_j+\pi_{ij}\\ \phi_i &= 1-\pi_i\\ \phi_j &= 1-\pi_j\end{aligned}\right\} \tag{11}$$

and (9), this can be expressed as

$$\pi_{ij}^{(n+1)} = \left(\frac{k_{ij}}{2}\right)\pi_{ij}^{(n)}, \tag{12}$$

where

$$k_{ij} = \left(1+\lambda_{ij}^2\right)/2.$$

Since π_{ij} can be expressed as $(\pi_j-\xi_{ij(i)})$, the recurrence relation (12) can be written as

$$\pi_j^{(n+1)} - \xi_{ij(i)}^{(n+1)} = \frac{k_{ij}}{2}\left(\pi_j^{(n)} - \xi_{ij\,(i)}^{(n)}\right),$$

which on using (9), reduces to

$$\xi_{ij(i)}^{(n+1)} - \frac{k_{ij}}{2}\xi_{ij(i)}^{(n)} = \frac{1-k_{ij}}{2}\pi_{ij}^{(n)}. \tag{13}$$

Replacing n by $(n+1)$ in (13) and then substituting $(1/2)\,\pi_j^{(n)}$ for $\pi_j^{(n+1)}$, we get

$$\xi_{ij(i)}^{(n+2)} - \frac{k_{ij}}{2}\xi_{ij(i)}^{(n+1)} = \frac{1-k_{ij}}{4}\pi_j^{(n)}. \tag{14}$$

Eliminating $\pi_j^{(n)}$ between (13) and (14), we get the recurrence relation for $\xi_{ij(i)}$ as

$$\xi_{ij(i)}^{(n+2)} - \frac{1+k_{ij}}{2}\xi_{ij(i)}^{(n+1)} + \frac{k_{ij}}{4}\xi_{ij(i)}^{(n)} = 0. \tag{15}$$

The recurrence relation for $\xi_{ij\,(j)}$ is the same as (15), i being replaced by j.

(b) *Three loci*

With three linked loci $i, j,$ and l, there are three linkage values $\lambda_{ij}, \lambda_{jl},$ and λ_{il}

respectively between ith and jth, jth and lth, and ith and lth loci connected by

$$\lambda_{il} = (1-c)\ (\lambda_{ij}+\lambda_{jl}-1) + c\ \lambda_{ij}\ \lambda_{jl}\ , \tag{16}$$

where c is the coefficient of coincidence.

When one or two loci are considered, we have the recurrence relations for π - functions as

$$\left.\begin{aligned} \pi_{i}^{(n+1)} &= \tfrac{1}{2}\,\pi_{i}^{(n)} \\ \pi_{ij}^{(n+1)} &= \frac{k_{ij}}{2}\,\pi_{ij}^{(n)}\ . \end{aligned}\right\} \tag{17}$$

When all three loci are taken into account and X is compared with itself, using (5), we get

$$\begin{aligned} \phi_{ijl}^{(n+1)} = {} & \tfrac{1}{8}\left[1+\lambda_{ij}^{2}+\lambda_{jl}^{2}+\lambda_{il}^{2}\right]\left[1+\phi_{ijl}^{(n)}\right] \\ & +\tfrac{1}{8}\left[1+\lambda_{ij}^{2}-\lambda_{jl}^{2}-\lambda_{il}^{2}\right]\left[\phi_{l}^{(n)}+\phi_{ij}^{(n)}\right] \\ & +\tfrac{1}{8}\left[1-\lambda_{ij}^{2}+\lambda_{jl}^{2}-\lambda_{il}^{2}\right]\left[\phi_{i}^{(n)}+\phi_{jl}^{(n)}\right] \\ & +\tfrac{1}{8}\left[1-\lambda_{ij}^{2}-\lambda_{jl}^{2}+\lambda_{il}^{2}\right]\left[\phi_{j}^{(n)}+\phi_{il}^{(n)}\right]. \end{aligned} \tag{18}$$

Making use of (6), (9), and (10) this can be expressed as

$$\pi_{ijl}^{(n+1)} = \tfrac{1}{8}\left[1+\lambda_{ij}^{2}+\lambda_{jl}^{2}+\lambda_{il}^{2}\right]\pi_{ijl}^{(n)}. \tag{19}$$

With no interference, $c = 1$ and (16) is simplified. The recurrence relation (19) can then be expressed simply as

$$\pi_{ijl}^{(n+1)} = \left(\frac{k_{ij}\ k_{jl}}{2}\right)\pi_{ijl}^{(n)}, \tag{20}$$

where

$$k_{ij} = \left(1+\lambda_{ij}^{2}\right)/2, \text{ and } k_{jl} = \left(1+\lambda_{jl}^{2}\right)/2.$$

The relations (8) enable π_{ijl} to be expressed in terms of the ξ-function and the π-function of order lower than three and since the recurrence relations for these π-functions, are given by (17), the recurrence relation (20) can ultimately lead to the following recurrence relations for ξ-functions.

$$\xi_{ijl(i)}^{(n+2)} - \frac{k_{jl}+k_{ij}\ k_{jl}}{2}\ \xi_{ijl(i)}^{(n+1)} + \frac{k_{ij}\,k_{jl}^{2}}{4}\ \xi_{ijl(i)}^{(n)} = 0; \tag{21}$$

$$\xi_{ijl(j)}^{(n+2)} - \frac{k_{il}+k_{ij}k_{jl}}{2}\ \xi_{ijl(j)}^{(n+1)} + \frac{k_{ij}k_{jl}k_{il}}{4}\ \xi_{ijl(j)}^{(n)} = 0; \tag{22}$$

$$\xi_{ijl(l)}^{(n+2)} - \frac{k_{ij}+k_{ij}k_{jl}}{2}\ \xi_{ijl(l)}^{(n+1)} + \frac{k_{ij}^{2}k_{jl}}{4}\ \xi_{ijl(l)}^{(n)} = 0; \tag{23}$$

$$\xi^{(n+4)}_{ijl(ij)} - \frac{1+k_{jl}+k_{il}+k_{ij}k_{jl}}{2}\,\xi^{(n+3)}_{ijl(ij)} + \frac{k_{jl}+k_{il}+k_{ij}k_{jl}+k_{jl}k_{il}+k_{ij}k_{jl}^2+k_{ij}k_{jl}k_{il}}{4}$$

$$\xi^{(n+2)}_{ijl(ij)} - \frac{k_{jl}k_{il}+k_{ij}k_{jl}^2+k_{ij}k_{jl}k_{il}+k_{ij}k_{jl}^2k_{il}}{8}\,\xi^{(n+1)}_{ijl(ij)} + \frac{k_{ij}k_{jl}^2k_{il}}{16}\,\xi^{(n)}_{ijl(ij)} = 0; \qquad (24)$$

$$\xi^{(n+4)}_{ijl(jl)} - \frac{1+k_{il}+k_{ij}+k_{ij}k_{jl}}{2}\,\xi^{(n+3)}_{ijl(jl)} + \frac{k_{il}+k_{ij}+k_{ij}k_{jl}+k_{ij}k_{il}+k_{ij}k_{jl}k_{il}+k_{ij}^2k_{jl}}{4}\,\xi^{(n+2)}_{ijl(jl)}$$

$$- \frac{k_{ij}k_{il}+k_{ij}k_{jl}k_{il}+k_{ij}^2k_{jl}+k_{ij}^2k_{jl}k_{il}}{8}\,\xi^{(n+1)}_{ijl(jl)} + \frac{k_{ij}^2k_{jl}k_{il}}{16}\,\xi^{(n)}_{ijl(jl)} = 0; \qquad (25)$$

$$\xi^{(n+4)}_{ijl(il)} - \frac{1+k_{ij}+k_{jl}+k_{ij}k_{jl}}{2}\,\xi^{(n+3)}_{ijl(il)} + \frac{k_{ij}+k_{jl}+2k_{ij}k_{jl}+k_{ij}^2k_{jl}+k_{ij}k_{ij}^2}{4}\,\xi^{(n+2)}_{ijl(il)}$$

$$- \frac{k_{ij}k_{jl}+k_{ij}^2k_{jl}+k_{ij}k_{jl}^2+k_{ij}^2k_{jl}^2}{8}\,\xi^{(n+1)}_{ijl(il)} + \frac{k_{ij}^2k_{jl}^2}{16}\,\xi^{(n)}_{ijl(il)} = 0. \qquad (26)$$

(c) *More than three loci*

The recurrence relations for the panmictic function π for the two and three loci cases given respectively by (12) and (20) show a general trend. For instance, for the four loci, 1,2,3 and 4 it would be given by

$$\pi^{(n+1)}_{1234} = \tfrac{1}{16}\left(1+\lambda^2_{12}+\lambda^2_{23}+\lambda^2_{34}+\lambda^2_{13}+\lambda^2_{14}+\lambda^2_{24}+\lambda^2_{1234}\right)\pi^{(n)}_{1234}. \qquad (27)$$

With no interference, this would, in view of (3), reduce to

$$\pi^{(n+1)}_{1234} = \left(\frac{k_{12}k_{23}k_{34}}{2}\right)\pi^{(n)}_{1234}, \qquad (28)$$

where $k_{ij} = \left(1+\lambda^2_{ij}\right)/2, \quad (ij)=(12),(23),(34).$

Thus, provided there is no interference, the recurrence relation for $\pi_{123\ldots r}$ is given by

$$\pi^{(n+1)}_{123\ldots r} = \left(\frac{k_{12}k_{23}\ldots k_{(r-1)r}}{2}\right)\pi^{(n)}_{123\ldots r} \qquad (29)$$

where $k_{ij} = \dfrac{1+\lambda^2_{ij}}{2} = \dfrac{1+(1-2p_{ij})^2}{2}$, p_{ij} being the probability of recombination between ith and $j=(i+1)$ th loci.

SOLUTIONS FOR RECURRENCE RELATIONS

(a) *Two loci*

If we initially start with a double heterozygote,

$$\left.\begin{aligned} &\pi^{(0)}_{ij} = \pi^{(0)}_{i} = \pi^{(0)}_{j} = 1 \\ &\xi^{(0)}_{ij(i)} = \xi^{(0)}_{ij(j)} = 0. \\ &\phi^{(0)}_{ij} = 0. \end{aligned}\right\} \qquad (30)$$

Then the recurrence relations for π-functions given by (9) and (12) lead to the solutions

$$\left.\begin{aligned} \pi_i^{(n)} &= (\tfrac{1}{2})^n \\ \pi_j^{(n)} &= (\tfrac{1}{2})^n \\ \pi_{ij}^{(n)} &= \left(\frac{k_{ij}}{2}\right)^n. \end{aligned}\right\} \tag{31}$$

The inbreeding function $\phi_{ij}^{(n)}$ after n generations of selfing would then, in view of (11), be given by

$$\phi_{ij}^{(n)} = 1 - 2(\tfrac{1}{2})^n + \left(\frac{k_{ij}}{2}\right)^n. \tag{32}$$

This result agrees with that given by Rajagopalan (1958).

Since $\xi_{ij(i)}$ and $\xi_{ij(j)}$ can be expressed respectively as $(\pi_{ij} - \pi_j)$ and $(\pi_{ij} - \pi_i)$, the solutions for ξ-functions can also be obtained with the help of (31) as given below.

$$\left.\begin{aligned} \xi_{ij(i)}^{(n)} &= (\tfrac{1}{2})^n - \left(\frac{k_{ij}}{2}\right)^n \\ \xi_{ij(j)}^{(n)} &= (\tfrac{1}{2})^n - \left(\frac{k_{ij}}{2}\right)^n. \end{aligned}\right\} \tag{33}$$

(b) *Three loci*

Again, initially starting with an individual heterozygous for all the three loci, the values of all the π-functions are unity whereas those of ξ- and ϕ-functions are zero in the zero-th generation. The solutions for the recurrence relations for π-functions given by (17) and (20) are

$$\left.\begin{aligned} \pi_i^{(n)} &= (\tfrac{1}{2})^n \\ \pi_j^{(n)} &= (\tfrac{1}{2})^n \\ \pi_l^{(n)} &= (\tfrac{1}{2})^n \\ \pi_{ij}^{(n)} &= \left(\frac{k_{ij}}{2}\right)^n \\ \pi_{jl}^{(n)} &= \left(\frac{k_{jl}}{2}\right)^n \\ \pi_{il}^{(n)} &= \left(\frac{k_{il}}{2}\right)^n \\ \pi_{ijl}^{(n)} &= \left(\frac{k_{ij}\, k_{jl}}{2}\right)^n \end{aligned}\right\} \tag{34}$$

The inbreeding function $\phi_{ijl}^{(n)}$ after n generations of selfing would then become, in view of (6),

$$\phi^{(n)}_{ijl}=1-3\,(\tfrac{1}{2})^n+\left(\frac{k_{ij}}{2}\right)^n+\left(\frac{k_{jl}}{2}\right)^n+\left(\frac{k_{il}}{2}\right)^n-\left(\frac{k_{ij}.k_{jl}}{2}\right)^n \quad (35)$$

The solutions for ξ-functions can similarly be obtained with the help of (8) and (34).

(c) *Four loci*

Starting with an individual heterozygous for each of the four loci 1, 2, 3 and 4, the solutions for various π-functions would be given by

$$\left.\begin{aligned}
&\pi^{(n)}_{i}=(\tfrac{1}{2})^n,\ (i)=(1),(2),(3),(4)\\
&\pi^{(n)}_{ij}=\left(\frac{k_{ij}}{2}\right)^n,\qquad (ij)=(12),(23),(34),(13)\ (14),(24)\\
&\pi^{(n)}_{ijl}=\left(\frac{k_{ij}.k_{jl}}{2}\right)^n,\quad (ijl)=(123),(124),(134),(234)\\
&\pi^{(n)}_{1234}=\left(\frac{k_{12}.k_{23}.k_{34}}{2}\right)^n.
\end{aligned}\right\} \quad (36)$$

The inbreeding function $\phi^{(n)}_{1234}$ after n generations of selfing would then become

$$\begin{aligned}\phi^{(n)}_{1231}=1-4(\tfrac{1}{2})^n&+\left(\frac{k_{12}}{2}\right)^n+\left(\frac{k_{23}}{2}\right)^n+\left(\frac{k_{34}}{2}\right)^n+\left(\frac{k_{13}}{2}\right)^n+\left(\frac{k_{14}}{2}\right)^n+\left(\frac{k_{24}}{2}\right)^n\\&-\left(\frac{k_{12}k_{23}}{2}\right)^n-\left(\frac{k_{23}k_{34}}{2}\right)^n-\left(\frac{k_{13}k_{34}}{2}\right)^n-\left(\frac{k_{12}k_{24}}{2}\right)^n+\left(\frac{k_{12}k_{23}k_{34}}{2}\right)^n \quad (37)\end{aligned}$$

(d) *More than four loci*

The results given by (32), (35) and (37) show a general trend. Thus, provided there is no interference, the inbreeding function ϕ, after n generations of selfing, having started with an individual heterozygous at each of the r loci involved, would be given by

$$\begin{aligned}\phi^{(n)}_{123\ldots r}=1-r(\tfrac{1}{2})^n&+\Sigma_1\left(\frac{k_{ij}}{2}\right)^n-\Sigma_2\left(\frac{k_{ij}k_{jl}}{2}\right)^n\\&+\Sigma_3\left(\frac{k_{ij}k_{jl}k_{lm}}{2}\right)^n+\ldots.\\&+(-)^r\left(\frac{k_{12}k_{23}\cdots k_{(r-1)r}}{2}\right)^n,\end{aligned} \quad (38)$$

where Σ_1 is summation over ${}^rc_2$ values of k given by $k_{12}, k_{23}, \ldots . k_{13}, k_{14}, \ldots . k_{(r-1)r}$. Similarly Σ_2 is summation over ${}^rc_3$ pairs of k values. The appropriate pair of k values involved would depend on the three loci selected out of r. For instance, if 1st, 4th and 9th loci are forming a trio, the two values will correspond to linkage between 1st and 4th, and 4th and 9th loci. Similar considerations are involved for other summations. The result given by (38) would reduce to

$$\phi^{(n)}_{123\ldots r}=1-(\tfrac{1}{2})^n \quad (39)$$

if all the loci are completely linked between themselves leading to all ks being 1, whereas it would reduce to

$$\phi^{(n)}_{123\ldots r}=[1-(\tfrac{1}{2})^n]^r \tag{40}$$

when there is no linkage between any pairs of loci i.e. all k values are each equal to 1/2.

Mean and Variance of the Number of Loci Homozygous by Descent

As can be seen from the above considerations, when there are more than one locus, there is a distribution of the number of loci homozygous by descent. The proportion of homozygosity by descent for 0, 1, 2,...loci depend on the k values, and the number of generations a population is inbred. The mean m and variance v of this distribution are given by

$$\left.\begin{aligned} m&=r[1-(\tfrac{1}{2})^n],\\ v&=r(\tfrac{1}{2})^n[1-r(\tfrac{1}{2})^n]+2\Sigma\left(\frac{k_{ij}}{2}\right)^n \end{aligned}\right\} r=2,\ 3,\ \text{etc.} \tag{41}$$

where Σ implies summation over ${}^rc_2$ values of i, i.e. summation is over all the pairs of loci which are distinct. Expressed in terms of ϕ_i and ϕ_{ij} these are given by

$$\left.\begin{aligned} m&=r\phi^{(n)}_i\\ v&=r\phi^{(n)}_i\left(1-r\phi^{(n)}_i\right)+2\Sigma\phi^{(n)}_{ij}. \end{aligned}\right\} \tag{42}$$

Rate of Inbreeding

Defining the rate of inbreeding in the nth generation as $\delta^{(n)}=\dfrac{\phi^{(n)}-\phi^{(n-1)}}{1-\phi^{(n-1)}}$, it may be seen that with one locus it is independent of the generation, whereas with more than one locus, it depends on the generation and the various recombination values. Thus

$$\left.\begin{aligned} \delta^{(n)}_1&=(\tfrac{1}{2})\\ \delta^{(n)}_{12}&\simeq(\tfrac{1}{2})\left[1-\tfrac{1}{2}\left\{k^{n-1}_{12}-k^n_{12}\right\}\right]\\ \delta^{(n)}_{123}&\simeq(\tfrac{1}{2})\left[1-\tfrac{1}{2}\left\{\left(k^{n-1}_{12}-k^n_{12}\right)+\left(k^{n-1}_{23}-k^n_{23}\right)+\left(k^{n-1}_{13}-k^n_{13}\right)\right\}\right]\\ &\cdots\cdots\cdots\\ \delta^{(n)}_{123\ldots r}&\simeq(\tfrac{1}{2})\left[1-\frac{1}{r}\Sigma_1\left(k^{n-1}_{ij}-k^n_{ij}\right)\right]. \end{aligned}\right\} \tag{43}$$

Numerical Results

Tables 1 and 2 respectively give the values of the ξ-function for the case of two and three loci, together with the means, variances and the rate of inbreeding over a period of five generations, assuming certain arbitrary recombination values.

Table 1. *Values of ξ-functions upto five generations for two loci with* $p_{12}=\cdot30$

ξ \ n	1	2	3	4	5
π_{12}	·2900	·0841	·0244	·0071	·0020
$\xi_{12(1)}$	·2100	·1659	·1006	·0554	·0292
$\xi_{12(2)}$	·2100	·1659	·1006	·0554	·0292
ϕ_{12}	·2900	·5841	·7744	·8821	·9396
m	1·0000	1·5000	1·7500	1·8750	1·9376
v	·5800	·4182	·2363	·1236	·0825
δ_{12}	·2900	·4142	·4576	·4774	·4877

Table 2. *Values of ξ-functions upto five generations for three loci with* $p_{12}=\cdot30$ *and* $p_{23}=\cdot20$

ξ \ n	1	2	3	4	5
π_{123}	·1972	·0389	·0077	·0015	·0003
$\xi_{123(1)}$	·1428	·0767	·0316	·0119	·0043
$\xi_{123(2)}$	·0628	·0287	·0099	·0031	·0009
$\xi_{123(3)}$	·0928	·0452	·0167	·0056	·0017
$\xi_{123(12)}$	·0972	·1057	·0758	·0460	·0257
$\xi_{123(23)}$	·1472	·1372	·0907	·0523	·0283
$\xi_{123(13)}$	·0672	·0892	·0690	·0435	·0249
ϕ_{123}	·1928	·4784	·6986	·8361	·9139
m	1·5000	2·2500	2·6250	2·8125	2·9064
v	1·0300	·7221	·3970	·2025	·1004
δ_{123}	·1928	·3538	·4222	·4562	·4747

It is observed from the above tables that the rate of inbreeding increases with further generations of selfing and that it is more for two loci than for three loci in every generation of selfing.

Tables 3 to 7 give the values of the inbreeding function for the case of three loci upto five generations of selfing with various combinations of values of p_{12} and p_{23}.

Table 3. *Values of the ϕ-function after one generation of selfing for three loci*

p_{23} \ p_{12}	0	·1	·3	·5
0	·5000	·4100	·2900	·2500
·1		·3338	·2322	·2100
·3			·1618	·1450
·5				·1250

Table 4. *Values of the ϕ-function after two generations of selfing for three loci*

p_{23} \ p_{12}	0	·1	·3	·5
0	·7500	·6681	·5841	·5625
·1		·5957	·5186	·5031
·3			·4524	·4381
·5				·4219

Table 5. *Values of the ϕ-function after three generations of selfing for three loci*

p_{23} \ p_{12}	0	·1	·3	·5
0	·8750	·8189	·7744	·7656
·1		·7677	·7246	·7171
·3			·6846	·6776
·5				·6699

Table 6. *Values of the ϕ-function after four generations of selfing for three loci*

p_{23} \ p_{12}	0	·1	·3	·5
0	·9375	·9033	·8821	·8789
·1		·8711	·8499	·8469
·3			·8298	·8270
·5				·8240

Table 7 *Values of the ϕ-function after five generations of selfing for three loci*

p_{23} \ p_{12}	0	·1	·3	·5
0	·9688	·9491	·9496	·9385
·1		·9305	·9206	·9197
·3			·9113	·9103
·5				·9093

It is apparent from the above tables that the range of effect of linkage on the homozygosity of a selfed population is maximum after one generation of selfing and goes on decreasing with further generations of selfing. This may be measured as the difference between the values of ϕ-function in the totally linked and unlinked cases. It is ·3700 after one generation, ·3281 after two generations, ·2051 after three generations, ·1135 after four generations and ·0595 after five generations of selfing for three loci. It is also seen that this range is more with three loci than with two loci in each generation of selfing. The above differences in the values of ϕ-function for two loci are ·2500, ·1875, ·1094, ·0586 and ·0303 respectively after the 1st, 2nd, 3rd, 4th and 5th generation of selfing.

It is also found in the case of three loci that after one generation of selfing the pairs of values for p_{12} and p_{23} can be ranked, in descending order of their effect on the homozygosity, as (0, 0), (·1, 0), (·1, ·1), (·3, 0), (·5, 0), (·3, ·1), (·5, ·1), (·3, ·3), (·5, ·3) and (·5, ·5). This ranking also holds true after two generations of selfing whereas after three generations, the effect of (·1, ·1) and (·3, 0) are almost the same. After four and five generations of selfing, the ranking is (0, 0), (·1, 0), (·3, 0), (·5, 0), (·1, ·1), (·3, ·1), (·5, ·1), (·3, ·3), (·5, ·3) and (·5, ·5). Thus, with two generations of selfing (·1, ·1) may be regarded as producing a tighter linkage than (·3, 0) or (·5, 0).

SUMMARY

1. A generalised 'coefficient of relationship' between two individuals X and Y has been defined with any number of linked loci.

2. Recurrence relations for ϕ-, π- and ξ-functions (Schnell, 1961) in the case of two and three loci have been obtained.

3. Solutions for recurrence relations have been given for ϕ-function upto the case of any number of linked loci.

4. It has been found that the effect of linkage on the homozygosity of a selfed population is more with a greater number of linked loci and is maximum after one generation of selfing. With three linked loci, the pairs of values of p_{12}, p_{23} when taken as (·1, ·1) exert a greater effect than (·3, 0) or (·5, 0) but this is true only upto two generations of selfing.

5. With more than one locus, the rate of inbreeding is not constant with further generations of selfing. It depends on the number of generations of selfing and the recombination values.

REFERENCES

FISHER, R. A. (1949). *The Theory of Inbreeding.* Oliver and Boyd: Edinburgh.

GEIRINGER, H. (1944). On the probability theory of linkage in Mendelian heredity. *Ann. Math. Statist.*, **15**, 25-57.

KEMPTHORNE, O. (1957). *An Introduction to Genetic Statistics.* John Wiley & Sons: New York.

MALÉCOT, G. (1948). *Les mathematiques de l'hérédité.* Masson et Cie: Paris.

RAJAGOPALAN, M. (1958). Effect of linkage on the homozygosity of selfed population. *J. Ind. Soc. Agri. Stat.*, **10**, 64-66.

SCHNELL, F. W. (1961). Some general formulations of linkage effects in inbreeding. *Genetics*, **46**, 947-957.

WRIGHT, S. (1921). Systems of mating I-V. *Genetics*, **6**, 111-178.

Reprint from GENETICS, Vol. 54, No. 1, Part 1, July, 1966. *Printed in U.S.A.*

2

EFFECT OF LINKAGE ON HOMOZYGOSITY OF A POPULATION UNDER MIXED SELFING AND RANDOM MATING

P. NARAIN

Institute of Agricultural Research Statistics, (I.C.A.R.), New Delhi, India

Received November 2, 1965

THE effect of a given system of consanguineous mating on the degree of homozygosity of resulting individuals was studied by WRIGHT (1921), MALÉCOT (1948) and FISHER (1949) using different approaches. KEMPTHORNE (1957) gives a comparative evaluation of these approaches. So far as the study of homozygosis under various inbreeding systems is concerned, the so-called probability method due to MALÉCOT which makes use of the probabilities of genes being identical by descent, appears to be simpler than the generation matrix method given by FISHER and more general than the method of path-coefficients developed by WRIGHT. KIMURA (1963) developed a method based on simple probability calculations which is similar to MALÉCOT's approach but not identical to it. HALDANE (1949) used MALÉCOT's treatment to study the association of phenotypes due to two linked loci under inbreeding and defined a coefficient of inbreeding for a linked gene pair. SCHNELL (1961), however, generalised the so-called probability approach of MALÉCOT for an arbitrary number of linked loci and gave some general formulations of the effects of linkage in inbreeding systems. NARAIN (1965) used this method to study the homozygosity of a selfed population with an arbitrary number of linked loci. Recently SHIKATA (1965) discussed a generalisation of the inbreeding coefficient in the form of a vector with components which are related to the inbreeding and panmictic functions defined by SCHNELL. In this paper the concept of identity by descent, has been adopted to study the homozygosity of a population under mixed selfing and random mating, for the case of an arbitrary number of linked loci. Such a system of mating was studied earlier by BENNETT and BINET (1956), BINET *et al.* (1959) and KIMURA (1963) for the case of two linked loci and by GHAI (1964) for the case of independently segregating loci.

One locus: Consider the ith locus with alleles A and a and let the original pool consist of the gametes A and a with frequencies u_i and v_i respectively. Then the probability that the two different genes in the pool are not identical in state is $2u_iv_i$. Let the population undergo inbreeding. Then an inbred individual can possess genes at the ith locus which are either identical by descent or not identical by descent. Let the probabilities of these two events be ϕ_i and π_i respectively. An individual can, therefore, be heterozygous if it carries the genes at the locus which are not identical by descent as well as which are not identical in state. The probability of heterozygosity is, therefore, $2u_iv_i\,\pi_i$ and since the gene frequencies u_i and v_i remain constant in the absence of selection, this probability is propor-

tional to π_i. This is WRIGHT's panmictic index which is used for deriving recurrence relations under various systems of inbreeding.

When the population is undergoing selfing, the panmictic index π_i with respect to the ith locus is known to have the following recurrence relation (KEMPTHORNE 1957)

$$\pi_i^{(n)} = \frac{1}{2}\pi_i^{(n-1)} \tag{1}$$

where the upper suffix refers to the generation number. Let there be a constant probability s of selfing and of $(1-s)$ of mating at random. Then an individual in the nth generation can possess the genes at the locus which are not identical by descent if either it is an offspring resulting from the randomly mating individuals or if it is an offspring of such a selfed individual in the $(n-1)$th generation which possesses genes at the locus which are also not identical by descent. In other words, the recurrence relation for π_i under the mixed selfing and random mating would be given by

$$\pi_i^{(n)} = s\left\{\frac{1}{2}\pi_i^{(n-1)}\right\} + (1-s) \tag{2}$$

The recurrence relation (2) gives the following solution:

$$\pi_i^{(n)} = \frac{2(1-s)}{2-s}\left\{1 - \left(\frac{s}{2}\right)^n\right\} + \left(\frac{s}{2}\right)^n \pi_i^{(0)} \tag{3}$$

The coefficient of inbreeding $\phi_i^{(n)}$ is then calculated by $\{1 - \pi_i^{(n)}\}$. As n tends to infinity, π_i tends to $2(1-s)/(2-s)$ and ϕ_i to $s/(2-s)$. Let the initial population be random mating, so that $\pi_i^{(0)}=1$. The rate of inbreeding $\delta_i^{(n)}$ in the nth generation can then be expressed by the following relation

$$\begin{aligned}\delta_i^{(n)} &= \frac{\phi_i^{(n)} - \phi_i^{(n-1)}}{\frac{s}{2-s} - \phi_i^{(n-1)}} \\ &= 1 - (s/2)\end{aligned} \tag{4}$$

Alternately, this rate can also be measured by

$$\delta_i^{(n)} \simeq -\log_e(s/2) \tag{5}$$

It is interesting to note that the rate of inbreeding with one locus is constant from generation to generation.

Two loci: Consider ith and jth loci with alleles A, a and B, b respectively. Let the original pool consist of the gametes AB, Ab, aB and ab with frequencies u_iu_j, u_iv_j, v_iu_j and v_iv_j respectively where u_i and v_i are the gene frequencies for alleles A, a and u_j and v_j are those for B, b. Such a choice of gametic frequencies assumes an initial random mating population which is in equilibrium. Then the probability that the two different gametes in the pool are not identical in state with respect to both the loci is $4u_iu_jv_iv_j$. On inbreeding, an individual can possess genes at the ith and jth loci such that either (i) genes at both the loci

are identical by descent, or (ii) genes at the ith locus are identical by descent but at the jth locus are not identical by descent and (iii) genes at the ith locus are not identical by descent but at the jth locus are identical by descent or (iv) genes at both the loci are not identical by descent. Let the probabilities of these four events be ϕ_{ij}, $\xi_{ij(i)}$, $\xi_{ij(j)}$ and π_{ij} respectively. An individual can, therefore, be heterozygous with respect to both the loci, if it carries the genes at both the loci which are not identical by descent as well as which are not identical in state. The probability of heterozygosity with respect to both the loci is, therefore, $4u_iu_jv_iv_j\,\pi_{ij}$ and since the gene frequencies u_i, v_i, u_j, v_j, remain constant in the absence of selection, this probability is proportional to π_{ij}. This is the generalized panmictic function for two loci due to SCHNELL, which can be used for deriving recurrence relations under various systems of inbreeding.

When the population is undergoing selfing, it was shown by NARAIN (1965) that the panmictic function π_{ij} has the following recurrence relation.

$$\pi_{ij}^{(n)} = \left(\frac{k_{ij}}{2}\right) \pi_{ij}^{(n-1)} \tag{6}$$

where $k_{ij} = p^2_{ij} + (1-p_{ij})^2$, p_{ij} being the probability of recombination between the ith and jth loci. When there is a mixture of selfing and random mating, it can be shown by the same arguments as given under one locus case, that the recurrrence relation for π_{ij} would be given by

$$\pi_{ij}^{(n)} = s\left\{\left(\frac{k_{ij}}{2}\right) \pi_{ij}^{(n-1)}\right\} + (1-s) \tag{7}$$

The recurrence relation (7) gives the following solution

$$\pi_{ij}^{(n)} = \frac{2(1-s)}{(2-sk_{ij})}\left\{1 - \left(\frac{sk_{ij}}{2}\right)^n\right\} + \left(\frac{sk_{ij}}{2}\right)^n \pi_{ij}^{(0)} \tag{8}$$

The function of inbreeding $\phi_{ij}^{(n)}$ is then calculated from the following relation

$$\phi_{ij}^{(n)} = 1 - \pi_i^{(n)} - \pi_j^{(n)} + \pi_{ij}^{(n)} \tag{9}$$

with the help of (3) and (8). As n tends to infinity, we get the following limiting values of π_{ij} and ϕ_{ij}

$$\begin{aligned} \operatorname*{Lt}_{n\to\infty} \pi_{ij}^{(n)} &= \frac{2(1-s)}{(2-sk_{ij})} \\ \operatorname*{Lt}_{n\to\infty} \phi_{ij}^{(n)} &= 1 - \frac{4(1-s)}{(2-s)} + \frac{2(1-s)}{(2-sk_{ij})} \end{aligned} \tag{10}$$

Considering initially a random mating population, $\pi_{ij}^{(0)} = 1$ and rate of inbreeding $\delta_{ij}^{(n)}$ is then given by

$$\begin{aligned} \delta_{ij}^{(n)} &= \frac{\phi_{ij}^{(n)} - \phi_{ij}^{(n-1)}}{\left\{\operatorname*{Lt}_{n\to\infty} \phi_{ij}^{(n)}\right\} - \phi_{ij}^{(n-1)}} \\ &= 1 - \left(\frac{s}{2}\right) d_{ij}^{(n)}/d_{ij}^{(n-1)} \end{aligned} \tag{11}$$

where

$$d_{ij}^{(n)} = \left(\frac{4}{2-s}\right) - 2\left(\frac{2-k_{ij}}{2-sk_{ij}}\right) k_{ij}^{(n-1)} .$$

Alternately the rate can be approximated as

$$\delta_{ij}^{(n)} \simeq -\log_e \left\{ \left(\frac{s}{2}\right) \frac{d_{ij}^{(n)}}{d_{ij}^{(n-1)}} \right\} \tag{12}$$

The rate of inbreeding, therefore, varies from generation to generation.

Three loci: The discussion given under two loci case can be generalised to three loci i, j and l. Provided the initial random mating population is in equilibrium so that the gametic frequencies are the product of the respective frequencies of genes, it can be seen that the probability of heterozygosity with respect to all the three loci is proportional to π_{ijl} which is the generalised panmictic function with respect to three loci. Thus the heterozygosis with respect to all the loci for a selfed population can be studied from the recurrence relation for π_{ijl} given by NARAIN (1965) as

$$\pi_{ijl}^{(n)} = \left(\frac{k_{ij}\, k_{jl}}{2}\right) \pi_{ijl}^{(n-1)} \tag{13}$$

where $k_{jl} = p^2_{jl} + (1-p_{jl})^2$, it being assumed that there is no interference and p_{jl} being the probability of recombination between jth and lth loci.

For the mixed selfing and random mating under consideration, the recurrence relation would be given by

$$\pi_{ijl}^{(n)} = s \left\{ \left(\frac{k_{ij}\, k_{jl}}{2}\right) \pi_{ijl}^{(n-1)} \right\} + (1-s) \tag{14}$$

The solution of this recurrence relation is given by

$$\pi_{ijl}^{(n)} = \frac{2(1-s)}{(2-sk_{ij}\, k_{jl})} \left\{ 1 - \left(\frac{sk_{ij}\, k_{jl}}{2}\right)^n \right\} + \left(\frac{sk_{ij}\, k_{jl}}{2}\right)^n \pi_{ijl}^{(0)} \tag{15}$$

The inbreeding function $\phi_{ijl}^{(n)}$ is then calculated from the relation

$$\phi_{ijl}^{(n)} = 1 - \pi_i^{(n)} - \pi_j^{(n)} - \pi_l^{(n)} + \pi_{ij}^{(n)} + \pi_{jl}^{(n)} + \pi_{il}^{(n)} - \pi_{ijl}^{(n)} \tag{16}$$

with the help of (3), (8) and (15). As n tends to infinity, we get the following limiting values

$$\begin{aligned} &\operatorname*{Lt}_{n\to\infty} \pi_{ijl}^{(n)} = 2(1-s)/(2-sk_{ij}k_{jl}) \\ &\operatorname*{Lt}_{n\to\infty} \phi_{ijl}^{(n)} = 1 - \frac{6(1-s)}{(2-s)} + 2(1-s)\left\{\frac{1}{(2-sk_{ij})} + \frac{1}{(2-sk_{jl})} + \frac{1}{(2-sk_{il})}\right\} \\ &\qquad - 2(1-s)/(2-sk_{ij}\, k_{jl}) \end{aligned} \tag{17}$$

where $k_{il} = (1 - k_{ij} - k_{jl} + 2k_{ij}k_{jl})$. Initially starting with a random mating population the rate of inbreeding $\delta_{ijl}^{(n)}$ is given by

$$\begin{aligned} \delta_{ijl}^{(n)} &= \frac{\phi_{ijl}^{(n)} - \phi^{(n-1)}_{ijl}}{\left\{ \operatorname*{Lt}_{n\to\infty} \phi_{ijl}^{(n)} \right\} - \phi^{(n-1)}_{ijl}} \\ &= 1 - (s/2)\ (d_{ijl}^{(n)}/d^{(n-1)}_{ijl}) \end{aligned} \tag{18}$$

where

$$d_{ijl}^{(n)} = \left(\frac{6}{2-s}\right) - 2(\overline{2-k_{ij}}/\overline{2-sk_{ij}})\ k_{ij}^{n-1} - 2(\overline{2-k_{jl}}/\overline{2-sk_{jl}})\ k_{jl}^{n-1} - 2(\overline{2-k_{il}}/\overline{2-sk_{il}})\ k_{il}^{n-1} + 2(\overline{2-k_{ij}k_{jl}}/\overline{2-sk_{ij}k_{jl}})\ k_{ij}^{n-1}k_{jl}^{n-1}$$

Alternately, it can be approximated by

$$\delta_{ijl}^{(n)} \simeq -\log_e\{(s/2)(d_{ijl}^{(n)}/d_{ijl}^{(n-1)})\} \tag{19}$$

Generalisation to an arbitrary number of loci: Consider r loci, 1,2,3---r. Under the assumptions stated in previous sections, the recurrence relations for π-function are given by

$$\begin{aligned}
&\pi_i^{(n)} = w\,\pi_i^{(n-1)} + (1-2w),\ i = 1,2,\text{------}r\\
&\pi_{ij}^{(n)} = w_{ij}\,\pi_{ij}^{(n-1)} + (1-2w),\ i<j,\ i = 1,2,\text{------}(r-1)\\
&\qquad\qquad j = 2,3,\text{------}(r)\\
&\pi_{ijl}^{(n)} = w_{ijl}\,\pi_{ijl}^{(n-1)} + (1-2w),\ i<j<l,\ i = 1,2,\text{------}(r-2)\\
&\qquad\qquad j = 2,3,\text{------}(r-1)\\
&\qquad\qquad l = 3,4,\text{------}(r)\\
&\text{---------------------------------------}\\
&\pi_{123\text{----}r}^{(n)} = w_{123\text{----}r}\,\pi_{123\text{----}r}^{(n-1)} + (1-2w)
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
w &= s/2\\
w_{ij} &= w\,k_{ij}\\
w_{ijl} &= w_{ij}\,k_{jl}\\
&\text{---}\quad\text{----}\\
&\text{---}\quad\text{----}\\
w_{123\text{----}(r-1)} &= w_{123\text{----}(r-2)}\,k_{(r-2),(r-1)}\\
w_{123\text{----}(r)} &= w_{123\text{----}(r-1)}\,k_{(r-1),(r)}
\end{aligned}$$

The solution of these recurrence relations is given by

$$\begin{aligned}
&\pi_i^{(n)} = \{(1-2w)(1-w^n)/(1-w)\} + w^n\,\pi_i^{(0)},\ i = 1,2,3,\text{.....}r\\
&\pi_{ij}^{(n)} = \{(1-2w)(1-w_{ij}^n)/(1-w_{ij})\} + w_{ij}^n\,\pi_{ij}^{(0)},\ i<j,\\
&\qquad\qquad i = 1,2,\text{.....}(r-1)\\
&\qquad\qquad j = 2,3,\text{.....}(r)\\
&\pi_{ijl}^{(n)} = \{(1-2w)(1-w_{ijl}^n)/(1-w_{ijl})\} + w_{ijl}^n\,\pi_{ijl}^{(0)},\ i<j<l,\\
&\qquad\qquad i = 1,2,\text{.....}(r-2)\\
&\qquad\qquad j = 2,3,\text{.....}(r-1)\\
&\qquad\qquad l = 3,4,\text{.....}(r)\\
&\text{---------------------------------------}\\
&\pi_{123\text{..}r}^{(n)} = \{(1-2w)(1-w_{123\text{.....}r}^n)/(1-w_{123\text{.....}r})\} + w_{123\text{.....}r}^n\,\pi_{123\text{.....}r}^{(0)}
\end{aligned} \tag{21}$$

The inbreeding function $\phi^{(n)}_{123\ldots r}$ is then calculated from the relation

$$\phi^{(n)}_{123\cdots r} = 1 - \sum_i \pi^{(n)}_i + \sum_{i<j} \pi^{(n)}_{ij} - \sum_{i<j<l} \pi^{(n)}_{ijl} + \text{-----} + (-)^r \pi^{(n)}_{123\cdots r} \tag{22}$$

As n tends to infinity, the inbreeding function $\phi^{(n)}_{123\cdots r}$ tends to the following limiting value

$$\underset{n\to\infty}{\text{Lt}}\, \phi^{(n)}_{123\cdots r} = 1 - (1-2w)\left\{\frac{r}{(1-w)} - \sum_{i<j}\frac{1}{(1-w_{ij})} + \sum_{i<j<l}\frac{1}{(1-w_{ijl})} - \text{----------} + (-)^r \frac{1}{(1-w_{123\cdots r})}\right\} \tag{23}$$

Initially starting with a random mating population, the rate of inbreeding with respect to r loci is given by

$$\delta^{(n)}_{123\cdots r} = \frac{\phi^{(n)}_{123\cdots r} - \phi^{(n-1)}_{123\cdots r}}{\left\{\underset{n\to\infty}{\text{Lt}}\, \phi^{(n)}_{123\cdots r}\right\} - \phi^{(n-1)}_{123\cdots r}} = 1 - w\frac{d^{(n)}_{123\cdots r}}{d^{(n-1)}_{123\cdots r}} \tag{24}$$

where

$$d^{(n)}_{123\cdots r} = \{r/(1-w)\} - w^n\left[\sum_{i<j}(2w-w_{ij})w^{n-1}_{ij}/(1-w_{ij}) - \sum_{i<j<l}(2w-w_{ijl})\,w^{n-1}_{ijl}/(1-w_{ijl}) + \text{- - - - - - - - - - - - - - - - - -} + (-)^r(2w-w_{123\cdots r})w^{n-1}_{123\cdots r}/(1-w_{123\cdots r})\right]$$

Alternately $\delta^{(n)}_{123\cdots r}$ can be approximated by

$$\delta^{(n)}_{123\cdots r} \simeq -\log_e\{w\, d^{(n)}_{123\cdots r}/d^{(n-1)}_{123\cdots r}\} \tag{25}$$

Numerical analysis: In order to study the effect of linkage the case of two loci ith and jth involving one function of recombination parameter $k = p^2 + (1-p)^2$ is taken up in detail. Assuming that initially the population is in equilibrium under random mating alone, the value of $\pi^{(0)}_{ij} = 1$, so that the function of inbreeding $\phi^{(n)}_{ij}$ with respect to two loci in the nth generation, as calculated from (9), is given by

$$\phi^{(n)}_{ij} = \phi(s,k,n) = s\{2k(1-s) + s(2-k)\}/(2-s)(2-sk) - \{2s/(2-s)\}(s/2)^n + \{s(2-k)/(2-sk)\}(sk/2)^n \tag{26}$$

whereas the rate of inbreeding $\delta^{(n)}_{ij}$ as calculated from (11) is given by

$$\delta^{(n)}_{ij} = \delta(s,k,n) = 1 - (s/2)\frac{[(4/\overline{2-s}) - 2(\overline{2-k}/\overline{2-sk})k^n]}{[(4/\overline{2-s}) - 2(\overline{2-k}/\overline{2-sk})k^{n-1}]} \tag{27}$$

These two functions, in variables s, k and n were studied numerically on the 1620 IBM Electronic Computor at the Institute of Agricultural Research Statistics. The extensive tables are, however, not reproduced here.

(a) *Inbreeding function–$\phi(s, k, n)$*. It is apparent from the numerical results obtained for this function that the inbreeding function ϕ increases for increasing amounts of self-fertilization for a fixed value of linkage after each generation of mixed selfing and random mating. Also, for a fixed level of selfing, it increases as recombination probability decreases from 0.5 to zero after each generation of mixed selfing and random mating. The value of k as .5 corresponds to no linkage, whereas the value of k as 1.0 corresponds to complete linkage. Therefore, the range of the effect of linkage can be studied by taking the difference in the values of inbreeding function for $k = 1.0$ and $k = 0.5$ at a fixed value of s after each generation of the mixed system of breeding. Using (26) this difference is given by

$$\begin{aligned}
\Delta\phi(s,n) &= \phi(s,k{=}1.0,n) - \phi(s,k{=}0.5,n) \\
&= \frac{s}{(2-s)(4-s)} \left\{2(1-s) + (4-s)\left(\frac{s}{2}\right)^n - 3(2-s)\left(\frac{s}{4}\right)^n\right\} \\
&= \frac{s}{2^{2n}(2-s)(4-s)} \{2^{2n+1} - 2^{2n+1}s + (2^{n+2}-6)s^n - (2^n-3)s^{n+1}\} \\
&= \frac{s}{2^{2n}} \{2^{2(n-1)} - 2^{2(n-2)}s + (1+2^2)2^{2(n-3)}s^2 \\
&\qquad - (1+2^2+2^3)2^{2(n-4)}s^3 - - - - - - \\
&\qquad - (1+2^2+ - - - - + 2^{n-1})s^{n-1}\} \\
&= -\sum_{r=-1}^{(n-2)} \left\{\frac{2^{(r+2)}-3}{2^{2(r+2)}}\right\} s^{r+2} \qquad (28)
\end{aligned}$$

Clearly this difference depends on n as well as s. At a particular integral value of n, it is a function of s and its various powers. The rate of change of $\Delta\phi$ with change in the value of s, at a fixed n, is given by

$$\frac{\partial\Delta\phi}{\partial s} = -\sum_{r=-1}^{(n-2)} \left\{\frac{(2^{(r+2)}-3)(r+2)}{2^{2(r+2)}}\right\} s^{r+1} \qquad (29)$$

Also

$$\begin{aligned}
\frac{\partial^2\Delta\phi}{\partial s^2} &= -\sum_{r=-1}^{(n-2)} \left[\frac{\{2^{(r+2)}-3\}(r+2)(r+1)}{2^{2(r+2)}}\right] s^r \\
&= -\sum_{r=0}^{(n-2)} \left[\frac{\{2^{(r+2)}-3\}(r+2)(r+1)}{2^{2(r+2)}}\right] s^r \qquad (30)
\end{aligned}$$

Since the expression under the summation sign in (30) is positive for $r \geqslant 0$, it follows that $\Delta\phi(s,n)$ has maxima for $n \geqslant 2$ at values of s given by the roots of the $(n-1)$th degree equation,

$$\sum_{r=0}^{(n-2)} \left[\frac{(2^{r+2}-3)(r+2)}{2^{2(r+2)}}\right] s^{r+1} = \tfrac{1}{4} \qquad (31)$$

For $n=2$ however the equation (31) gives a solution $s=2$ which lies outside the permissible range of s, namely $0 < s \leqslant 1$. Hence it follows that $\Delta\phi(s,n)$ has maxima for $n > 2$ at values of s lying between 0 and 1 obtained by solving (31). For $n=1$ and 2, the maximum values can be taken as those corresponding to $s=1$.

TABLE 1

Maximum values of $\Delta\phi$-function

n	$\Delta\phi(s)$	s	Maximum value of $\Delta\phi$
1	$(s/4)$	1.00	0.2500
2	$(s/4) - (s^2/16)$	1.00	0.1875
3	$(s/4) - (s^2/16) - (5s^3/64)$	0.80	0.1200
4	$(s/4) - (s^2/16) - (5s^3/64) - (13s^4/256)$	0.70	0.1054
5	$(s/4) - (s^2/16) - (5s^3/64) - (13s^4/256) - (29s^5/1024)$	0.60	0.1019

Maximum values of $\Delta\phi$ corresponding to values of s falling in the permissible range, for n=1 to 5 are shown in Table 1.

It is evident, therefore, that the range of the effect of linkage goes on increasing with the increase in the amount of self-fertilization for the first two generations of the mixed system of breeding. In the subsequent generations, the range of the effect of linkage increases up to a certain value of the proportion of self-fertilization and decreases thereafter.

(b) *Rate of inbreeding–$\delta(s,k,n)$*. From the numerical results obtained for this function it is seen that the rate of inbreeding δ decreases for increasing amounts of self-fertilization at a given value of linkage after each generation of mixed selfing and random mating. After the first generation, the rate of inbreeding increases for decreasing values of recombination probability from 0.5 to zero at all levels of selfing. After the second generation, however, this linear trend is true only for values of selfing up to 0.5. For values of s beyond 0.6 in second generation and for all values of s in subsequent generations, the trend is quadratic i.e. decreasing first, attaining a minimum and then increasing. The points of minima for the function δ can be obtained from the equation:

$$\frac{\partial\delta(s,k,n)}{\partial k} = 0, \quad (n > 1) \tag{32}$$

i.e. from the equation,

$$(2-s)(2-k)^2 k^n - 2snk^3 + 2\{2(n+1) + 3s(n-1)\}k^2 + 4\{(2-n)s - 3n\}k + 8(n-1) = 0, \qquad (n > 1) \tag{33}$$

It can also be noted that when the linkage is complete, i.e. $k = 1.0$, the rate of inbreeding is the same after each generation for a given proportion of selfing. The range of the effect of linkage can be measured by

$$\Delta\delta(s,n) = \delta(s, k = 1.0, n) - \delta(s, k = 0.5, n)$$
$$= \frac{3s(2-s)}{(4-s)}\left(\frac{1}{2}\right)^{n+2}\left\{1 - \frac{3(2-s)}{(4-s)}\left(\frac{1}{2}\right)^n\right\}^{-1} \tag{34}$$

This goes on decreasing with n increasing. In fact after four to five generations of the mixed system of breeding, presence or absence of linkage practically makes little difference in the rate of inbreeding.

DISCUSSION

It has been demonstrated by NARAIN (1965) and also in this paper how, with the use of the concept of identity by descent, the recurrence relations for the panmictic functions with an arbitrary number of linked loci can be obtained for the case of (a) selfing and (b) mixed selfing and random mating with considerable ease. In the case of (b) the initial population has been assumed to be in random mating equilibrium so that the gametic frequencies are the product of the gene frequencies and therefore do not undergo any change during the subsequent generations as a result of the given system of mating. In such a case, the probability of heterozygosity in any particular generation is simply proportional to the panmictic function, the constant of proportionality depending only on the gene frequencies and therefore the changes in the value of panmictic functions directly reveal the changes in the heterozygosity of the population. If the initial population is an arbitrary population not in equilibrium, the recurrence relations obtained in this paper for the panmictic functions would still hold true because their derivations do not employ the assumption of the equilibrium. But now the proportion of genotypes heterozygous with respect to a set of loci in a particular generation would not only depend on the corresponding panmictic function but also on some other factors related to the amount of disequilibrium except when $s = 1$. After practising the given system of mating for an infinite number of generations, however, the additional factors would disappear so that in the limit the heterozygosity would bear again the same simple relation with the panmictic function as if the initial population had been in equilibrium. For the case of two linked loci the relationship between the heterozygosity and the panmictic function in any particular generation when an initial arbitrary population is subject to the mixed selfing and random mating, can be easily worked out and is discussed below.

For the pair of linked loci ith and jth let $p_{ij}^{(n)}(AaBb)$ and $p_{ij}^{(0)}(AaBb)$ be the frequencies of double heterozygotes in the nth generation and the initial population respectively. Let $L_{ij}^{(r)}(AB)$, $(r = 1,2, - - - \overline{n-1})$ denote the intensity of linkage disequilibrium with respect to the set of two loci at the gametic stage following the rth generation, defined by BENNETT (1954) as

$$L_{ij}^{(r)}(AB) = p_{ij}^{(r)}(AB) - u_i u_j \qquad (35)$$

where $p_{ij}^{(r)}(AB)$ is the frequency of the gamete AB of the rth generation. The recurrence relation for $p_{ij}(AaBb)$ is then given by

$$p_{ij}^{(n)}(AaBb) = (sk_{ij}/2)\, p_{ij}^{(n-1)}(AaBb) + A_{ij}^{(n-1)}(AB) + 4(1-s)u_i u_j v_i v_j \qquad (36)$$

where

$$A_{ij}^{(n-1)}(AB) = 2(1-s)\{(u_i - v_i)(u_j - v_j) + 2\, L_{ij}^{(n-1)}(AB)\}\, L_{ij}^{(n-1)}(AB)$$

This relation gives the following solution

$$p_{ij}^{(n)}(AaBb) = (sk_{ij}/2)^n\, p_{ij}^{(0)}(AaBb) + \{8(1-s)u_i u_j v_i v_j/(2-sk_{ij})\}$$

$$\times \left\{1-\left(\frac{sk_{ij}}{2}\right)^n\right\} + (sk_{ij}/2)^{n-1} B_{ij}^{(n)}(AB) \quad (37)$$

where

$$B_{ij}^{(n)}(AB) = \sum_{r=1}^{n} (2/sk_{ij})^{r-1} A_{ij}^{(r-1)}(AB)$$

From (37) and (8) and taking $\pi_{ij}^{(0)} = 1$, since in the initial population the genes are not identical by descent at both the loci, we get the following relation between the frequency of double heterozygotes and the panmictic function

$$p_{ij}^{(n)}(AaBb) = 4u_iu_jv_iv_j\,\pi_{ij}^{(n)} + \left(\frac{sk_{ij}}{2}\right)^n \{p_{ij}^{(0)}(AaBb) - 4u_iu_jv_iv_j\}$$

$$+\left(\frac{sk_{ij}}{2}\right)^{n-1} B_{ij}^{(n)}(AB) \quad (38)$$

When the initial population is in linkage equilibrium, $p_{ij}^{(0)}(AaBb) = 4u_iu_jv_iv_j$ and

$$L_{ij}^{(r)}(AB) = 0, \ (r = 1, 2, ---- n-1)$$

so that the relationship reduces to

$$p_{ij}^{(n)}(AaBb) = 4u_iu_jv_iv_j\,\pi_{ij}^{(n)} \quad (39)$$

Also in the limit when $n \to \infty$, (38) reduces to

$$p_{ij}^{(\infty)}(AaBb) = 4u_iu_jv_iv_j\,\pi_{ij}^{(\infty)}$$

$$= \frac{8\,(1-s)u_iu_jv_iv_j}{(2-sk_{ij})} \quad (40)$$

in view of (10).

With complete selfing i.e. $s = 1$, (38) reduces to

$$p_{ij}^{(n)}(AaBb) = p_{ij}^{(0)}(AaBb)\,\pi_{ij}^{(n)} \quad (41)$$

This shows that only when $s = 1$ the panmictic function can also be regarded as "heterozygosity relative to the initial population" as in the one locus case.

The rate of decrease of double heterozygotes is given by

$$\frac{p_{ij}^{(n-1)}(AaBb) - p_{ij}^{(n)}(AaBb)}{p_{ij}^{(n-1)}(AaBb) - p_{ij}^{(\infty)}(AaBb)} = 1 - \left(\frac{sk_{ij}}{2}\right)$$

$$- \frac{A_{ij}^{(n-1)}(AB)}{\left(\frac{sk_{ij}}{2}\right)^{n-2}\left[\frac{sk_{ij}}{2}\{p_{ij}^{(0)}(AaBb) - p_{ij}^{(\infty)}(AaBb)\} + B_{ij}^{(n-1)}(AB)\right]} \quad (42)$$

whereas the rate of decrease of the corresponding panmictic function is given by

$$\frac{\pi_{ij}^{(n-1)} - \pi_{ij}^{(n)}}{\pi_{ij}^{(n-1)} - \pi_{ij}^{(\infty)}} = 1 - \left(\frac{sk_{ij}}{2}\right) \quad (43)$$

It is thus clear that so far as two linked loci are concerned the rate of decrease in

the panmictic function due to one generation of mixed selfing and random mating will not reflect the rate at which double heterozygotes decrease under the influence of this system unless the initial population is in linkage equilibrium. For the case of selfing only, i.e. $s = 1$, the rates would be identical as otherwise expected. When more than two loci are considered, the relationship between the heterozygosity and the panmictic function would involve linkage disequilibrium of higher orders as defined by BENNETT (1954).

The numerical results obtained for the case of two linked loci suggest that the effect of linkage on the homozygosity interacts with the proportion of self-fertilization prevailing in the population, in the absence of selection. It is obvious that two tightly linked loci would tend to become simultaneously homozygous more than two loosely linked loci. The degree of this effect of linkage, however, does not remain constant with changes in the proportion of selfing. During the first two generations subsequent to the operation of the mating system on an initial population in equilibrium, the degree of the effect of linkage as measured by $\Delta\phi(s,n)$ goes on increasing and attains its peak value for complete selfing. The peak value, however, is lower for the second generation as compared to that of the first. It, therefore, appears that the inbreeding process accelerates the effect of linkage on the homozygosity. But at the same time this accelerated effect tends to diminish in the next generation. It is interesting to note that in the third and the subsequent generations the inbreeding process accelerates the effect of linkage only up to a particular level of the proportion of selfing and retards thereafter. Such a particular level of selfing, however, becomes smaller and smaller in subsequent generations, whereas the peak value of $\Delta\phi(s,n)$ tends to settle down. It can, therefore, be argued that in the initial stages the inbreeding process and linkage both tend to act in the same direction to increase the simultaneous homozygosity at both the loci but afterwards the inbreeding process tends to dampen the effect of linkage.

The rate of inbreeding with respect to the two loci taken simultaneously is also found to depend on the recombination probability. Unlike the case with one locus where it depends only on the selfing proportion, the rate of inbreeding in this case varies also from generation to generation. The effect of linkage is to increase the rate of inbreeding but this increase is linear only after one generation of the mating system. In the later stages, this rate of inbreeding practically becomes constant at all values of recombination probability.

SUMMARY

General results for an arbitrary number of linked loci have been obtained for the inbreeding function and the rate of inbreeding for a population undergoing mixed selfing and random mating by the Malécot's approach. The effect of linkage on the homozygosity of the population has been studied numerically for the two loci case. It has been found that the linkage interacts with the proportion of self-fertilization prevailing in the population in the absence of selection. The relationship between the probability of heterozygosity and the panmictic function when the initial population is not in linkage equilibrium has also been discussed.

LITERATURE CITED

BENNETT, H. J., 1954 On the theory of random mating. Ann. Eugenics **18**: 311–317.

BENNETT, H. J., and F. E. BINET, 1956 Association between Mendelian factors with mixed selfing and random mating. Heredity **10**: 51–55.

BINET, F. E., A. M. CLARK, and H. T. CLIFFORD, 1959 Correlation due to linkage in certain wild plants. Genetics **44**: 5–13.

FISHER, R. A., 1949 *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

GHAI, G. L., 1964 The genotypic composition and variability in plant populations under mixed self-fertilization and random mating. J. Indian Soc. Agri. Stat. **16**: 94–125.

HALDANE, J. B. S., 1949 The association of characters as a result of inbreeding and linkage. Ann. Eugenics **15**: 15–23.

KEMPTHORNE, O., 1957 *An Introduction to Genetic Statistics*. Wiley, New York.

KIMURA, M., 1963 A probability method for treating inbreeding systems especially with linked genes. Biometrics **19**: 1–17.

MALÉCOT, G., 1948 *Les mathématiques de l'hérédité*. Masson, Paris.

NARAIN, P., 1965 Homozygosity in a selfed population with an arbitrary number of linked loci. J. Genet. **59**: 1–13.

SCHNELL, F. W., 1961 Some general formulations of linkage effects in inbreeding. Genetics **46**: 947–957.

SHIKATA, M., 1965 A generalisation of the inbreeding coefficient. Biometrics **21**: 665–681.

WRIGHT, S., 1921 Systems of mating. I–V. Genetics **6**: 111–178.

Reprinted from *Indian Journal of Heredity*, Vol. 4, No. 1, 1972, page 6-10.

A NOTE ON THE GENERALISATION OF HOMOZYGOTE × HETEROZYGOTE MATINGS

P. NARAIN AND K. M. REDDY*

Institute of Agricultural Research Statistics, New Delhi

A genetic incompatibility model is one in which only certain specific types of mating out of all the possible types of mating produce viable offspring. Self-sterility mechanism in *Nicotiana* (East and Mangelsdorf, 1925) and heterostyly in *Primula* (Bodmer, 1960) are some of the examples conforming to incompatibility models. Finney (1952) however, introduced certain incompatibility models in which the only possible type of mating is between homozygotes and heterozygotes with respect to a single locus and two alleles. This mating system can be looked upon either as a mating system in which only the matings between homozygotes and heterozygotes produce viable offspring or the homozygotes and heterozygotes belong to two different sets with matings possible only between individuals picked up at random from the two different sets. Scudo (1964) described this model as a basis for polygenic sex-determination. Karlin and Feldman (1968) studied the effect of the introduction of a third allele into this system and their maintenance. In this note these models have been generalized to situations in which several loci are segregating independently.

With two loci segregating independently in a population, there are 9 genotypes AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb and aabb. Suppose these 9 genotypes are divided into two groups, one group consisting of the four homozygotes and the double heterozygote and the other consisting of the four single heterozygotes. The first group is divided into two sets, one consisting of the four homozygotes and the other consisting of the only double heterozygote. The other group is also divided into two sets, one consisting of the two genotypes heterozygous at the first locus and the other consisting of the other two genotypes heterozygous at the other locus. The matings are now allowed only between the two sets within each of the two groups. This ensures a mechanism for matings only between homozygotes and heterozygotes at each of the two loci separately as well as simultaneously. The eight types of mating produced from the above arrangement would give rise to the 9 genotypes in the offspring generation with frequencies which can be related to their frequencies in the previous generation. Let the frequencies of the 9 genotypes be u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}, u_{31}, u_{32}, and u_{33} respectively with super-scripts n and (n-1) for the respective nth and (n-1) th generation and with their sums as one. Then we have:

$$T^{(n-1)}\, u_{11}^{(n)} = u_{11}^{(n-1)}\, u_{22}^{(n-1)} + u_{21}^{(n-1)}\, u_{12}^{(n-1)}$$

$$T^{(n-1)}\, u_{12}^{(n)} = u_{22}^{(n-1)} [u_{11}^{(n-1)} + u_{13}^{(n-1)}] + u_{12}^{(n-1)} [u_{21}^{(n-1)} + u_{23}^{(n-1)}]$$

$$T^{(n-1)}\, u_{13}^{(n)} = u_{13}^{(n-1)}\, u_{22}^{(n-1)} + u_{23}^{(n-1)}\, u_{12}^{(n-1)}$$

* Present Address: Directorate of Research Labs., West Block 8, New Delhi.

$$T^{(n-1)}\, u_{21}^{(n)} = u_{22}^{(n-1)}\,[u_{11}^{(n-1)} + u_{31}^{(n-1)}] + u_{21}^{(n-1)}\,[u_{12}^{(n-1)} + u_{32}^{(n-1)}]$$

$$T^{(n-1)}\, u_{22}^{(n)} = u_{22}^{(n-1)}\,[u_{11}^{(n-1)} + u_{13}^{(n-1)} + u_{31}^{(n-1)} + u_{33}^{(n-1)}] + [u_{21}^{(n-1)} + u_{23}^{(n-1)}]\,[u_{12}^{(n-1)} + u_{32}^{(n-1)}]$$

$$T^{(n-1)}\, u_{23}^{(n)} = u_{22}^{(n-1)}\,[u_{13}^{(n-1)} + u_{33}^{(n-1)}] + u_{23}^{(n-1)}\,[u_{12}^{(n-1)} + u_{32}^{(n-1)}]$$

$$T^{(n-1)}\, u_{31}^{(n)} = u_{31}^{(n-1)}\, u_{22}^{(n-1)} + u_{21}^{(n-1)}\, u_{32}^{(n-1)}$$

$$T^{(n-1)}\, u_{32}^{(n)} = u_{22}^{(n-1)}\,[u_{31}^{(n-1)} + u_{33}^{(n-1)}] + u_{32}^{(n-1)}\,[u_{21}^{(n-1)} + u_{23}^{(n-1)}]$$

$$T^{(n-1)}\, u_{33}^{(n)} = u_{22}^{(n-1)}\, u_{33}^{(n-1)} + u_{23}^{(n-1)}\, u_{32}^{(n-1)}$$

where

$$T^{(n-1)} = 4\, u_{22}^{(n-1)}\,[u_{11}^{(n-1)} + u_{13}^{(n-1)} + u_{31}^{(n-1)} + u_{33}^{(n-1)}] + 4\,[u_{21}^{(n-1)} + u_{23}^{(n-1)}]\,[u_{12}^{(n-1)} + u_{32}^{(n-1)}]$$

If $u_{1\cdot}$, $u_{2\cdot}$, $u_{3\cdot}$ represent the frequencies of genotypes AA, Aa and aa respectively with obvious relations like $u_{1\cdot} = u_{11} + u_{12} + u_{13}$ etc., it follows from the results of homozygote × heterozygote mating with single locus, that for $n \geq 1$,

$$u_{1\cdot}^{(n)} = \frac{p_1}{2(1+p_1)}$$

$$u_{2\cdot}^{(n)} = \tfrac{1}{2}$$

$$u_{3\cdot}^{(n)} = \frac{1}{2(1+p_1)}$$

where

$$p_1 = \frac{u_{12}^{(1)}}{u_{32}^{(1)}} = \frac{u_{1\cdot}^{(1)}}{u_{3\cdot}^{(1)}} = \quad \ldots\ldots\ldots \quad = \frac{u_{1\cdot}^{(n)}}{u_{3\cdot}^{(n)}}$$

Similarly if $u_{\cdot 1}$, $u_{\cdot 2}$ and $u_{\cdot 3}$ represent the frequencies of genotypes BB, Bb and bb respectively with relations like $u_{\cdot 1} = u_{11} + u_{21} + u_{31}$ etc,. then for $n \geq 1$

$$u_{\cdot 1}^{(n)} = \frac{p_2}{2(1+p_2)}$$

$$u_{\cdot 2}^{(n)} = \tfrac{1}{2}$$

$$u_{\cdot 3}^{(n)} = \frac{1}{2(1+p_2)}$$

where

$$p_2 = \frac{u_{21}^{(1)}}{u_{23}^{(1)}} = \frac{u_{\cdot 1}^{(1)}}{u_{\cdot 3}^{(1)}} = \quad \ldots\ldots\ldots \quad = \frac{u_{\cdot 1}^{(n)}}{u_{\cdot 3}^{(n)}}$$

It can be easily seen from the recurrence relations and results given above that for $n \geq 1$,

$$u_{12}^{(n)} = u_{11}^{(n)} + u_{13}^{(n)} = \frac{p_1}{4(1+p_1)}$$

Similarly, for $n \geq 1$,

$$u_{21}^{(n)} = u_{11}^{(n)} + u_{31}^{(n)} = \frac{p_2}{4(1+p_2)}$$

$$u_{22}^{(n)} = u_{21}^{(n)} + u_{23}^{(n)} = \frac{1}{4}$$

$$u_{23}^{(n)} = u_{13}^{(n)} + u_{33}^{(n)} = \frac{1}{4(1+p_2)}$$

$$u_{32}^{(n)} = u_{31}^{(n)} + u_{33}^{(n)} = \frac{1}{4(1+p_1)}$$

Now consider the ratios

$$x_1^{(n)} = u_{11}^{(n)}/u_{13}^{(n)}$$
$$x_2^{(n)} = u_{31}^{(n)}/u_{33}^{(n)}$$
$$x_3^{(n)} = u_{11}^{(n)}/u_{31}^{(n)}$$
$$x_4^{(n)} = u_{13}^{(n)}/u_{33}^{(n)}$$

It can be shown that as $n \to \infty$, $x_1^{(n)}$ and $x_2^{(n)}$ each tends to p_2 whereas $x_3^{(n)}$ and $x_4^{(n)}$ each tends to p_1. For instance, the recurrence relation for x_1 is given by

$$x_1^{(n)} = \frac{p_2 + (1+2p_2)x_1^{(n-1)}}{(2+p_2) + x_1^{(n-1)}}$$

from which the result follows. Hence after a sufficiently large number of generations, the frequency of four homozygotes AABB, AAbb, aaBB and aabb are respectively given by

$$u_{11}^{(\infty)} = \frac{p_1 p_2}{4(1+p_1)(1+p_2)}$$

$$u_{13}^{(\infty)} = \frac{p_1}{4(1+p_1)(1+p_2)}$$

$$u_{31}^{(\infty)} = \frac{p_2}{4(1+p_1)(1+p_2)}$$

$$u_{33}^{(\infty)} = \frac{1}{4(1+p_1)(1+p_2)}$$

Thus so far as the heterozygotes—single or double—are concerned the population attains equilibrium just after one generation but so far as the homozygotes are concerned, it takes an indefinitely large number of generations for the attainment of equilibrium. The equilibrium genotypic frequencies depend on the ratios p_1 and p_2 which depend on the initial genotypic frequencies.

The above results can be generalised to any arbitrary number of independently segregating loci. Thus for three such loci, there are 27 genotypes which are first divided into four groups. The first group consists of the triple heterozygote and all the eight homozygotes. The second group consists of two double heterozygotes with respect to the first two loci and four single heterozygotes with respect to the third loci. The third group consists of two double heterozygotes with respect to the first and third loci and four single heterozygotes with respect to the second loci. Similarly the last group consists of two double heterozygotes with respect to the second and third loci and four single heterozygotes with respect to the first loci. The first group is divided into two sets, one consisting of the triple heterozygote and the other consisting of the eight homogygotes. Each of the other three groups are divided into two sets, one consisting of the double heterozygotes and the other consisting of the single heterozygotes. The matings are now allowed only between the two sets in each of the four groups. This ensures the homozygote × heterozygote mating system. There would then be 32 types of mating determining the offspring generation. Proceeding in the same way as for the case of two loci, it can be shown that so far as the heterozygotes—single, double or triple—are concerned, the population attains equilibrium in just one generation. However, it takes an indefinitely large number of generations for the eight homozygotes to attain their equilibrium frequencies. The equilibrium genotypic frequencies now depend on the three ratios p_1, p_2, and p_3 which are expressed in terms of the initial genotypic frequencies with the help of the results of single locus for each of the three loci respectively. It is interesting to note that, as in the case of two loci, the equilibrium frequency of a genotype is the product of the equilibrium frequencies of the three genotypes at each of the three loci separately. It was further found with the help of electronic computer, that the attainment of the equilibrium for the homozygotes in the three loci case is achieved in a larger number of generations (about 35 generations) than for the homozygotes in the two loci case (about 20 generations).

SUMMARY

The homozygote × heterozygote mating system has been studied with respect to several independently segregating loci. With respect to all the genotypes, the population attains equilibrium after an infinitely large number of generations in contrast to the case of single locus where the population takes only one generation to attain equilibrium. However, in respect of genotypes which are heterozygous for at least one locus and homozygous for remaining loci it takes only one generation to attain equilibrium. Such genotypes as are homozygous at all the loci take an infinitely large number of generations to attain equilibrium. In this case it has been further found that as the number of loci increases, the number of generations required to attain the equilibrium also increases. The equilibrium genotypic frequencies are found to depend on certain ratios between the initial genotypic frequencies and can be easily obtained by multiplying the equilibrium genotypic frequencies expected at each of the loci separately.

REFERENCES

BODMER, W. F. 1960. Genetics of homostyly in populations of *Primula vulgaris*. Phil. Trans. *Roy. Soc. B.*, *212*, 517-549.

EAST, E. M. AND MANGELSDORF, A. J. 1925. A new interpretation of the hereditary behaviour of self-sterile plants. *Proc. Nat. Acad. Sci.*, **2**, 166-171.

FINNEY, D. J. 1952. The equilibrium of a self-incompatible polymorphic species. *Genetics* **26**, 33-64.

KARLIN, S. AND FELDMAN, M. W. 1968. Analysis of models with homozygote × heterozygote matings. *Genetics* **59**, 105-116.

SCUDO, F. M. 1964. Sex population genetics. *Ric. Sci.* **34**, (II-B), 93-146.

PRINTED IN INDIA AT THE WESLEY PRESS, MYSORE

Reprinted from "Curr. Sci.", November 20, 1976, Vol. 45, No. 22, 800-801

ON FISHER'S FUNDAMENTAL THEOREM OF NATURAL SELECTION WITH NON-OVERLAPPING GENERATIONS

THE Fundamental Theorem of Natural Selection first given by Fisher[1] broadly appears in two forms (Turner[2]). According to one form, the change in the average fitness of a population is equal to the genotypic variance in fitness. The other form, which includes the effect of a mating system, states that for random mating population, with two-allele system, the rate of increase in average fitness at any time is equal to its additive genetic variance at that time. Thus in the absence of dominance these two forms are identical. However, the interpretation of the dominance in a two-allele system is essentially that of an interaction between the two alleles, the three fitness values attached to the three genotypes forming an arithmetic series in the absence of dominance. Now it may happen that the three fitness values may form a geometric series so that on the logarithmic scale there would be no dominance although on the arithmetic scale some partial dominance will be exhibited resulting in two different forms of the theorem. The purpose of this communication is therefore to show that in situations where the dominance is due to scale effects, the two forms of the theorem would still be identical, although the variance due to dominance deviations on the arithmetic scale is not zero.

Let the relative fitness of the three genotypes *AA*, *Aa* and *aa* be respectively w_2, w_1 and w_0 in a random mating population with gene frequencies p for *A* and q for *a* with $p + q = 1$. The average fitness of such a population, denoted by $\bar{w}$ can be expressed as

$$\bar{w} = pw_2 + qw_0 - pq\,(w_2 - 2w_1 + w_0) \qquad (1)$$

where $(w_2 - 2w_1 + w_0)$ expresses the degree of dominance on the arithmetic scale. After the operation of natural selection, the increase in average fitness denoted by $\Delta \bar{w}$ is

$$\Delta \bar{w} = \sigma_w^2/\bar{w} \qquad (2)$$

where σ_w^2 is genotypic variance of the fitness values. Now the genotypic variance is the sum of additive and dominance variances which can be expressed respectively as

$$\sigma_A^2 = 2pq[\bar{w}(w_2 - 2w_1 + w_0) + (w_1^2 - w_2 w_0)] \qquad (3)$$

$$\sigma_D^2 = p^2q^2(w_2 - 2w_1 + w_0)^2. \qquad (4)$$

An alternative expression for the change in the average fitness can then be shown to be equal to

$$\Delta \bar{w} = \frac{\sigma_A^2}{\bar{w}}\left[1 + \frac{pq(w_2 - 2w_1 + w_0)}{2(\bar{w} + \theta)}\right] \qquad (5)$$

where

$$\theta = (w_1^2 - w_2 w_0)/(w_2 - 2w_1 + w_0). \qquad (6)$$

Taking into account the round of random mating in addition to the effect of natural selection, the change in the average fitness of the population, denoted by $\Delta \bar{w}^*$ is, according to the derivations of Li[3] and using (1), given by

$$\Delta \bar{w}^* = \frac{\sigma_A^2}{2\bar{w}}\left[1 + \left(\frac{pw_2 + qw_0}{\bar{w}}\right)\right]$$

$$= \frac{\sigma_A^2}{\bar{\omega}}\left[1 + \frac{pq(w_2 - 2w_1 + w_0)}{2\bar{\omega}}\right]. \qquad (7)$$

Comparing the two expressions (5) and (7), it is clear that even if $(w_2 - 2w_1 + w_0)$ is not zero, the two $\Delta \bar{w}$ and $\Delta \bar{w}^*$ are identical provided $w_1^2 = w_2 w_0$. This would mean that, on the geometric scale, there is no dominance.

It is apparent therefore that while the behaviour of the average fitness under natural selection without involving the effects of mating system is more rigorous, the change in average fitness, taking into account the random mating, is very much dependent on the scale on which the fitness values are measured. This also points out to the limitation of the concept of the additive genetic variance in that it involves the linearity assumption. On the other hand, the genotypic variance is a quantity free from any such assumption and therefore the Fundamental Theorem of Natural Selection employing it, *viz.*, in the form of (2) is much more general than its other form given by (7) which has been debated by several workers.

Institute of Agricultural Research Statistics,
Library Avenue,
New Delhi 110 012, *April* 23, 1976.

PREM NARAIN.

1. Fisher, R. A., *The Genetical Theory of Natural Selection*, Clarendon Press, Oxford, 1930.
2. Turner, J. R. G., *Mathematical Topics in Population Genetics*, Edited by K. Kojima, Springer-Verlag, Berlin-Heidelberg, New York, 1970.
3. Li, C. C., *Nature*, London, 1967, 214, 505.

The Indian Journal of Heredity, Vol. 1, Number 1, 1969

5

LIMITS AND DURATION OF RESPONSE TO SELECTION IN FINITE POPULATIONS : THE USE OF TRANSITION PROBABILITY MATRICES

P. NARAIN* AND ALAN ROBERTSON

Institute of Animal Genetics, West Mains Road, Edinburgh-9, U.K.

INTRODUCTION

The theory of limits of response to selection in finite populations has been developed primarily on the basis of the diffusion approximation to the distribution of gene frequencies (Wright 1945, Kimura 1957, Robertson 1960). The assumptions in this approach are that the population size, though finite, is very large so that the frequency of a gene in the population varies continuously between 0 and 1 and that the changes in gene frequency due to selection are small from generation to generation so that the time variable can also be taken as continuous. This amounts to regarding the selection process as a stochastic process continuous in time and space. The distribution of gene frequencies is obtained by solving the Kolmogorov forward equation whereas the probability of fixation is obtained by using the backward equation.

In the exact description of the process, however, the gene frequency is a discrete variate between 0 and 1, and the changes in the gene frequency from generation to generation are not necessarily small. The process is discrete in time and space. The distribution of gene frequencies and the probability of fixation can then be studied by using transition probability matrices. This approach has been used by Robertson (1960), Ewens (1963), Allan and Robertson (1964), Hill and Robertson (1966), Hill and Robertson (1968), Hill (1969) and others, usually with the help of a computer. An analytical treatment of the determination of the distribution of gene frequencies and the probability of fixation using this approach, does not appear to have been given previously.

In this paper a general theory of the transition matrix approach is developed which gives the chance of fixation as well as the expected change in the gene frequency by a given time in any genetic situation. For selection at a single locus with two alleles, the matrix formulae have been expanded as a power series in the selection coefficient s and compared with those from the diffusion approach.

The duration of response to selection in a finite population is often expressed in terms of half-life i.e. the time in which the expected gene frequency gets halfway to the limit (Robertson 1960, Hill 1969). Using the expression for the expected change in the gene frequency by a given generation, a more accurate

* Present address: Institute of Agricultural Research Statistics (I.C.A.R.), Library Avenue, New Delhi.

formula for half-life than that given by Robertson has been derived in this paper for additive gene action. The distribution of time to fixation of a particular allele, disregarding the cases in which it is lost, appears to give a better picture of the duration of response to selection in finite populations. The general theory of the transition matrix approach provides the matrix formulae for the determination of the mean and variance of time until fixation of a gene in any genetic situation. For a single locus with two alleles, comparable computer results have been obtained for the mean and the coefficient of variation of time until fixation. Kimura and Ohta (1968) have recently given formulae for the mean time until fixation and Narain (1969) has given the formulae for its variance from the diffusion approach.

The distribution of time until homozygosity i.e. either fixation or loss of the gene, can also be derived. When the chance of fixation of a gene is very near to unity, this will coincide with that of time to fixation. Ewens (1963) gave computer results based on transition matrices as well as the diffusion approximation for the mean time until homozygosity. In this paper a comparison of the mean and variance of time until fixation with those until homozygosity has also been given.

THEORY OF THE TRANSITION MATRIX APPROACH

Consider a finite population of gametes of constant size $n=2N$ in a population of diploid individuals of size N and a single locus with two alleles A_1 and A_2. Such a population can assume $(n+1)$ states $E_0, E_1, \ldots, E_{(n-1)}, E_n$, the i^{th} state E_i representing the state of i A_1 genes and $^{(n-i)}$ A_2 genes. The states E_0 and E_n represent respectively the states with A_2 or A_1 genes entirely and once the population assumes these states, it is fixed for either A_2 or A_1 respectively. In Markov chain terminology, these are absorbing states. Any other state E_i, $i = 1, 2, \ldots . (n\text{-}1)$, represents a mixture of A_1 and A_2 genes and any population in such a state has in the next generation a possibility of going to any of the other states. These are the transient states. Suppose P_{ij} represents the conditional probability that there are j A_1 genes out of n genes in a population, given that there were i A_1 genes in the previous generation. Since there are $(n+1)$ states we have $(n+1) \times (n+1)$ P_{ij}'s which can be conveniently represented by a matrix **P** as given below:

$$(1) \qquad \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ P_{10} & P_{11} & P_{12} & \cdots & P_{1,\,n-1} & P_{1,n} \\ P_{20} & P_{21} & P_{22} & \cdots & P_{2,\,n-1} & P_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{(n-1),0} & P_{(n-1),1} & \cdots & & P_{(n-1),(n-1)} & P_{(n-1),n} \\ 0 & 0 & \cdots & & & 1 \end{bmatrix}$$

If we consider transitions between the transient states only, then the matrix can be represented by **Q**.

$$(2)\quad \mathbf{Q} = \begin{bmatrix} P_{11} & \cdots & P_{1,(n-1)} \\ \cdots & \cdots & \cdots \\ P_{(n-1),1} & \cdots & P_{(n-1),(n-1)} \end{bmatrix}$$

We now consider the transition probabilities over an interval of t generations and denote them by $P_{ij}^{(t)}$ with the corresponding Q-matrix $\mathbf{Q}(t)$. We know from the theory of finite Markov chains (Kemeny and Snell 1960) that

$$(3)\quad \mathbf{Q}(t)=\mathbf{Q}^t$$

Now consider the matrix sum

$$(4)\quad \mathbf{T}(t)=\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\ldots+\mathbf{Q}^t$$

where $\mathbf{I}$ is a unit matrix with ones as the diagonal elements and zeros elsewhere. The j^{th} element in the i^{th} row of $\mathbf{T}^{(t)}$ give the expected total number of times the population spends in the state E_j upto the t^{th} generation, having started from the state E_i.

Let $p_i^{(t)}$ be the expected gene frequency of A_1 by t^{th} generation, given that the initial population has its frequency as $p_i^{(0)} = i/n$. The expected change in the gene frequency of A_1 at the t^{th} generation is then $r_i^{(t)}$ given by

$$(5)\quad r_i^{(t)} = p_i^{(t)} - p_i^{(0)}$$

If δp_i denotes the change in the gene frequency of A_1 in a single step when the initial frequency is $p_i^{(0)}$, we have

$$(6)\quad \delta p_i = p_i^{(1)} - p_i^{(0)}$$

The elements $r_i^{(t)}$ and δp_i for $i = 1, 2 \ldots (n\text{-}1)$ can be expressed respectively as column vectors $\mathbf{r}(t)$ and $\mathbf{p}$-s. In a similar manner $\mathbf{u}(t)$ denotes the column vector of the fixation probabilities by the t^{th} generation. $\mathbf{u}^{(1)}$ would then refer to the fixation probabilities in one step which are given by the last column of $\mathbf{Q}$. The fixation probability by the t^{th} generation is the sum of the expected total number of times the population spends in the different transient states by the $(t\text{-}1)^{th}$ generation multiplied by the corresponding probability of fixation in one step from that state. That is,

$$(7)\quad \mathbf{u}(t)=\mathbf{T}(t\text{-}1)\mathbf{u}(1)$$

Similar considerations show that

$$(8)\quad \mathbf{r}(t)=\mathbf{T}(t\text{-}1)\delta\mathbf{p}$$

Since the P-matrix is a Markov matrix with elements as probabilities it follows from matrix theory (Faddeeva 1958) that the roots of $\mathbf{Q}$ are all positive and less than unity and that the inverse of $(\mathbf{I}\text{-}\mathbf{Q})$ exists i.e. $[\mathbf{I}\text{-}\mathbf{Q}] \neq 0$, then

$$\mathbf{T}(t\text{-}1)=\mathbf{I}+\mathbf{Q}+\ldots+\mathbf{Q}^{(t-1)} \tag{9}$$
$$=(\mathbf{I}-\mathbf{Q}^{t})\ (\mathbf{I}-\mathbf{Q})^{-1}$$

Hence

$$\mathbf{u}(t)=(\mathbf{I}-\mathbf{Q}^{t})\ (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}(1) \tag{10}$$

$$\mathbf{r}(t)=(\mathbf{I}-\mathbf{Q}^{t})\ (\mathbf{I}-\mathbf{Q})^{-1}\delta\mathbf{p} \tag{11}$$

As **t** tends to infinity, $\mathbf{Q}^{t}$ tends to zero, so that if **u** and **r** denote the vectors of the eventual fixation probability and the expected change in the gene frequency of A_1 in the limit (i.e. the selection limit) respectively, these are given by

$$\mathbf{u}=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}(1) \tag{12}$$

$$\mathbf{r}=(\mathbf{I}-\mathbf{Q})^{-1}\delta\mathbf{p} \tag{13}$$

Alternatively,

$$\mathbf{u}(t)=(\mathbf{I}-\mathbf{Q}^{t})\,\mathbf{u} \tag{14}$$

$$\mathbf{r}(t)=(\mathbf{I}-\mathbf{Q}^{t})\,\mathbf{r} \tag{15}$$

Since

$$\text{Lim. } \mathbf{T}(t)=\mathbf{I}+\mathbf{Q}+\mathbf{Q}^{2}+\ldots \tag{16}$$
$$t\to\infty \quad =(\mathbf{I}-\mathbf{Q})^{-1}$$
$$=\mathbf{T}$$

it follows that the elements in the i^{th} row of **T** are the expected total number of times the population spends in the different transient states on the way to eventual loss or fixation of the gene from an initial state i. This is known as the fundamental matrix of the absorbing Markov chain. The row sums of this matrix then give the average time to absorption. In our genetic context this is the average time to homozygosity. If **m** denotes the column vector for the average time to homozygosity and **e** denotes a column vector with all the elements as unity, then

$$\mathbf{m}=\mathbf{T}\,\mathbf{e} \tag{17}$$

As the proportion of times that a population goes from a particular state to the fixation of the allele A_1 is given by the elements of the vector **u**, the vector **m** given by

$$\mathbf{m}=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u} \tag{18}$$
$$=(\mathbf{I}-\mathbf{Q})^{-2}\mathbf{u}(1)$$

gives the expected total number of steps in processes which end with the fixation of the allele A_1. The mean time until fixation of gene A_1 is therefore given by the ratio of the elements of vectors **m** and **u** respectively.

If we expand $(\mathbf{I}-\mathbf{Q})^{-2}$ in (18)

$$\mathbf{m}=(\mathbf{I}+2\mathbf{Q}+3\mathbf{Q}^{2}+\ldots)\mathbf{u}\,(1) \tag{19}$$

We can then obtain the moments of the distribution of time until fixation of the desirable allele by generalizing this formula. For instance, the second moment is given by the ratio of the elements of vectors **v** and **u** respectively, where **v** is given by

$$\text{(20)} \qquad \begin{aligned} \mathbf{v} &= (\mathbf{I}+2^2\mathbf{Q}+3^2\mathbf{Q}^2+\ldots)\mathbf{u}(1) \\ &= (2(\mathbf{I}-\mathbf{Q})^{-3}-(\mathbf{I}-\mathbf{Q})^{-2})\mathbf{u}(1) \end{aligned}$$

The variance can then be obtained by subtracting the square of the mean from the second moment. The matrix formulae for the moments of the distribution of time until loss of A_1, disregarding the cases in which it is fixed, can be obtained by using expressions for **u**, **m**, and but now **u**(1) refers to the one-step fixation probabilities given by the first column of **P** with first and last elements dropped.

THE EXPANSION OF THE TRANSITION PROBABILITY MATRIX

The theory developed in the previous section can be applied to specific genetic models. If no selective forces are operating and the transition probabilities are binomial, the situation is often known as Wright's model (Wright 1931). Here, starting with a given frequency p_i of allele A_1, in lines of constant breeding size of N individuals, we can consider the second generation as derived from the first by sampling of groups of $n=2N$ haploid sets, the gene frequency in the different groups being distributed binomially with mean np_i and index n. The next generation is then the repetition of this process, each line giving rise to a group of lines whose gene frequencies are binomially distributed about the mean of the parent line.

In this context we have to consider the processes of selection and sampling as occurring sequentially in that order. The selection is, however, at the gametic level. We can let A_1 alleles have a selective value $(1+s/2)$ compared to A_2 alleles with selective value unity, so that the relative number of offspring have expectations proportional to $(1+s/2)$ and 1 respectively where s is small. We, then, assume that a large number of offspring are produced but exactly N of them survive. Sampling of groups of $n=2N$ haploid sets is made to produce the next generation. Now the gene frequency in the different lines is distributed binomially with mean np_i' and index n, where p_i' is the gene frequency of A after selection and is given by

$$\text{(21)} \qquad \begin{aligned} p_i &= p_i + \delta p_i \\ &= p_i + s/2\ p_i\ (1-p_i)\ /\ (1 + p_i\ s/2) \end{aligned}$$

Thus with haploid selection and binomial sampling, we can regard the number of A_1 genes in any generation as a Markovian variate with transition probability P_{ij} given by

$$\text{(22)} \qquad p_{ij} = \binom{n}{j} (p_i')^j (1-p_i')^{n-j}$$

These probabilities give the P- and the Q-matrices introduced in the previous section. The matrix formulae for the limits and duration of response to selection developed in the previous section can now be used with the help of a high speed computer to give results numerically up to a given degree of accuracy. Since the capacity of a computer is limited, this can be done only for populations of reasonable size. On the other hand we can also obtain results for $\mathbf{r}$ and $\mathbf{r}(t)$, by expanding the transition probabilities in powers of the selection coefficient s, for a given N, utilizing the properties of the transition probability matrices with no selection, referred to hereafter by $\mathbf{P}_o$ and $\mathbf{Q}_o$. The properties of these matrices, in terms of the eigen-roots and vectors, are discussed in Feller (1951) and Robertson (1952).

Let Σp_i be approximated by $\frac{s}{2} p_i(1-p_i)(1-\frac{s}{2}p_i)$. Then P_{ij} can be expressed approximately as

$$P_{ij} = P_{ij}(o)\,[1 + sa_{ij} + s^2 b_{ij} + O(s^3)] \tag{23}$$

where

$$P_{ij}(o) = \binom{n}{j} p_i^{\,j} (1-p_i)^{\overline{n-j}} \tag{24}$$

is the transition probability with no selection and a_{ij} and b_{ij} are given by

$$a_{ij} = N(p_j - p_i) \tag{25}$$

$$b_{ij} = \frac{N^2}{2}\left[p_j^2 + \left(1 + \frac{1}{n}\right) p_i^2 - 2p_i p_j - \frac{1}{n} p_j\right] \tag{26}$$

p_j being j/n. In terms of Q- matrices, we can write

$$\mathbf{Q} = \mathbf{Q}_o + s\mathbf{Q}'_o + s^2\mathbf{Q}''_o + O(s^3) \tag{27}$$

where $\mathbf{Q}'_o$ and $\mathbf{Q}''_o$ are $(n-1)$ x $(n-1)$ matrices with the ij^{th} element as $a_{ij}\,P_{ij}(o)$ and $b_{ij}\,P_{ij}(o)$ respectively.

Now we have seen, in the previous section, that the quantities $\mathbf{r}$ and $\mathbf{r}(t)$ can be obtained by operating certain functions of $\mathbf{Q}$ on to the column vector of the change in the mean gene frequency in a single generation, which is expressible in terms of the right hand eigen-vectors of $\mathbf{Q}_o$. Hence the expanded form of $\mathbf{r}$ and $\mathbf{r}(t)$ can be determined if we know the operations of certain functions of $\mathbf{Q}_o$ on to these vectors. It has been found that we need nine operators and the first three vectors of $\mathbf{Q}_o$. The vectors are

$$x_1 = [\,p_i\ (1-p_i\ )] \tag{28}$$

$$x_2 = [\,p_i\ (1-p_i\ )\ (1-2p_i\ )\,] \tag{29}$$

$$x_3 = \left[\,p_i\ (1-p_i\ )\left(\frac{n-1}{5n-6} - p_i\ (1-p_i\ )\right)\right] \tag{30}$$

corresponding to the three eigen-roots of $\mathbf{Q}_o$ given by

$$\lambda_1 = \left(1 - \frac{1}{n}\right) \tag{31}$$

(32) $$\lambda_2 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)$$

(33) $$\lambda_3 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)$$

The results of performing these operations are given in Table 1 where only the 18 operations out of the 27 possible which are actually needed in the expansions of $\mathbf{r}$ and $\mathbf{r}(t)$ are given.

(*a*) *Expected Selection limit*

The selection limit vector $\mathbf{r}$ has been expressed as the operation of $\mathbf{T}=(\mathbf{I}-\mathbf{Q})^{-1}$ on to $\delta\mathbf{p}$, the vector of the expected change in the gene frequency in one step. When expressed in powers of s up to terms involving s^2, these are given by

(34) $$\mathbf{T} = \mathbf{T}_0 + s\mathbf{T}_0\mathbf{Q}'_0\mathbf{T}_0 + s^2[\mathbf{T}_0\mathbf{Q}''_0\mathbf{T}_0 + \mathbf{T}_0\mathbf{Q}_0\mathbf{T}_0\mathbf{Q}_0\mathbf{T}_0]$$

(35) $$\delta\mathbf{p} = \frac{s}{2}\left(1 - \frac{s}{4}\right)\mathbf{x}_1 + \frac{s^2}{8}\mathbf{x}_2$$

where

(36) $$\mathbf{T}_0 = (\mathbf{I}-\mathbf{Q}_0)^{-1}$$

The item by item operation of the terms in the expansion of the matrix on to those in the expansion of the vector has been performed with the help of Table 1. The result, expressed as the linear function of the three vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, is given by

(37) $$\mathbf{r} = \alpha_{10}\mathbf{x}_1 + \alpha_{20}\mathbf{x}_2 + \alpha_{30}\mathbf{x}_3$$

where α_{10}, α_{20} and α_{30} are given in Table 2.

Starting with an initial frequency p, the expected selection limit becomes

(38) $$u(p) = p + \alpha_{10}p(1-p) + \alpha_{20}p(1-p)(1-2p) + \alpha_{30}p(1-p)\left[\frac{n-1}{5n-6} - p(1-p)\right]$$

When N becomes very large and Ns is kept constant, α_{10}, α_{20}, α_{30} tend to their limiting values as given in Table 2. This gives

(39) $$\begin{aligned} u(p) &= p + Ns(1-N^2s^2/15)\,p(1-p) + \tfrac{1}{3}N^2s^2p(1-p)(1-2p) \\ &\quad + \frac{1}{3}N^3s^3p(1-p)\left[\frac{1}{5} - p(1-p)\right] + \ldots \\ &= p + Nsp(1-p) + \tfrac{1}{3}N^2s^2p(1-p)(1-2p) - \tfrac{1}{3}N^3s^3p^2(1-p)^2 + \ldots \end{aligned}$$

This expression is exactly the same, up to terms involving N^3s^3, as given in Kimura (1964).

A comparable expansion has been obtained by the junior author for the situation with k alleles with additive gene action. The fitness of the zygote formed

by the i^{th} and j^{th} allele is assumed to be $1+(s_i+s_j)/2$, with the conditions that all s's are small and that $\sum_{j=1}^{k} s_j p_j = o$, where p_j is the initial frequency of the j^{th} allele. By an expansion of the transition matrix as a power series in the coefficients, the chance of final fixation of the j^{th} allele can be shown to be

$$u_j = p_j + Ns_j p_j + \frac{N^2}{3} p_j (s_j 2 - 2V) - \frac{2N^3}{3} s_j p_j V + \ldots \tag{39a}$$

where V is the variance in fitness between individuals in the initial population. With only two alleles present, this reduces to the earlier expansion, bearing in mind that now $s_j = s(1-p)$ and $V = s^2 p(1-p)/2$. Details of the derivation will be given elsewhere.

(b) Expected change in the gene frequency by the t^{th} generation

The vector of the expected change in the gene frequency of A_1 by the t^{th} generation has been expressed as the operation of the matrix $(\mathbf{I}-\mathbf{Q}^t)$ on to $\mathbf{r}$, the vector of the expected selection limit. The former, when expressed in powers of s upto terms involving s^2, is given by

$$(\mathbf{I}-\mathbf{Q}^t) = (\mathbf{I}-\mathbf{Q}_o^t) - s(\mathbf{Q}_o^t)' - s^2(\mathbf{Q}_o^t)'' \tag{40}$$

and $\mathbf{r}$ has already been expressed as a linear function of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ by (37). The item by item operation of terms in $(\mathbf{I}-\mathbf{Q}^t)$ on to those in $\mathbf{r}$ has been performed with the help of Table 1. The result, expressed as a linear function of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ is given by

$$\begin{aligned} \mathbf{r}(t) = {} & (\alpha_{10} + \alpha_{11}\lambda_1^t + \alpha_{12}\lambda_2^t)\mathbf{x}_1 \\ & + (\alpha_{20} + \alpha_{21}\lambda_1^t + \alpha_{22}\lambda_2^t)\mathbf{x}_2 \\ & + (\alpha_{30} + \alpha_{31}\lambda_1^t + \alpha_{32}\lambda_2^t + \alpha_{33}\lambda_3^t)\mathbf{x}_3 \end{aligned} \tag{41}$$

where α_{10}, α_{20}, α_{30}, α_{11}, α_{12}, α_{21}, α_{22}, α_{31}, α_{32} and α_{33} are given in Table 2.

When N becomes very large and s becomes very small such that Ns is kept constant, the α's tend to their limiting values. This gives for the expected gene frequency at time t

$$\begin{aligned} E(p_t) = p + Ns \Big[& 1 - (1 - \frac{3}{100} N^2 s^2) e^{-t/2N} - \frac{N^2 s^2}{10} e^{-2t/2N} \\ & + \frac{N^2 s^2}{12} e^{-3t/2N} - \frac{N^2 s^2}{75} e^{-6t/2N} \Big] p(1-p) \\ & + \frac{1}{3} N^2 s^2 \Big[1 - \frac{3}{2} e^{-t/2N} + \frac{1}{2} e^{-3t/2N} \Big] p(1-p)(1-2p) \\ & - \frac{1}{3} N^3 s^3 \Big[1 - \frac{9}{5} e^{-t/2N} + e^{-3t/2N} - \frac{1}{5} e^{-6t/2N} \Big] p^2 (1-p)^2 \end{aligned} \tag{42}$$

(c) Half-life of the selection process

The half-life is the time t_h by which the expected change in the gene frequency is half of that in the limit. Hence we have to solve for t in the following equation

(43) $$E(p_t) - p = (\tfrac{1}{8})[u(p) - p]$$

after substituting $u(p)$ and $E(p_t)$ from (39) and (42) respectively. This gives an equation of the sixth order in $x = e^{-t/2N}$,

(44) $$Ax^6 + Bx^3 + Cx^2 + Dx + E = O$$

where

(45) $$A = -\frac{1}{15} N^2 s^2 \left[\frac{1}{5} - p(1-p)\right]$$

$$B = \frac{1}{6} Ns\,(1-2p) + \frac{1}{3} N^2 s^2 \left[\frac{1}{4} - p(1-p)\right]$$

$$C = -\frac{1}{10} N^2 s^2$$

$$D = -1 - \frac{1}{2} Ns\,(1-2p) + \frac{3}{5} N^2 s^2 \left[\frac{1}{20} + p(1-p)\right]$$

$$E = \frac{1}{2} + \frac{1}{6} Ns(1-2p) - \frac{1}{6} N^2 s^2 p\,(1-p)$$

Since $x_0 = \frac{1}{2}$ is a solution when Ns is small, an improved value of x can be obtained by applying the Newton-Raphson method to the equation. This is

(46) $$x = x_0 - \frac{Ax_0^6 + Bx_0^3 + Cx_0^2 + Dx_0 + E}{6Ax_0^5 + 3Bx_0^2 + 2Cx_0 + D}$$

$$= \frac{1}{2}\left[\frac{1 + \frac{1}{4} Ns\,(1-2p) + \frac{5}{96} N^2 s^2 \left(\frac{1}{5} - p(1-p)\right) - \frac{1}{8} N^2 s^2 p\,(1-p)}{1 + \frac{3}{8} Ns\,(1-2p) + \frac{1}{20} N^2 s^2 \left(\frac{1}{5} - p(1-p)\right) - \frac{5}{16} N^2 s^2 p(1-p)}\right]$$

Then, $t_h = -2N \log_e x$, is given by

(47) $$t_h = -2N \left[Log_e \frac{1}{2} + Log_e \left[1 + \frac{1}{4} Ns\,(1-2p) + \frac{5}{96} N^2 s^2 \left(\frac{1}{5} - p\,(1-p)\right) - \frac{1}{8} N^2 s^2 p\,(1-p)\right] - Log_e \left[1 + \frac{3}{8} Ns\,(1-2p) + \frac{1}{20} N^2 s^2 \left(\frac{1}{5} - p(1-p)\right) - \frac{5}{16} N^2 s^2 p(1-p)\right]\right]$$

As p tends to 0 or 1, the half-lives are found to be 1.58 N and 1.03 N respectively when $Ns = 1$. We can further approximate by expanding the terms in logarithm and retaining terms upto those involving $(Ns)^2$. This gives

(48) $$t_h = \left[1.4 + \tfrac{1}{4} Ns\,(1-2p) + 1.12 N^2 s^2 (0.16 - p\,(1-p))\right] N$$

For $Ns = 1$ and $p = \frac{1}{2}$, t_h is found to be 1.30 N. The half-lives thus obtained for $Ns = 1$ and $p = \frac{1}{2}$, tending to 0 and 1 are very close to those read from Fig. 11 of Hill and Robertson (1966), or from Fig. 1 of Hill (1969).

MEAN AND VARIABILITY OF THE TIME TO FIXATION OF A PARTICULAR ALLELE

We have not discussed the expansion of the mean and variance of the time to fixation of the desirable allele in the previous section since it involves, in addition, the use of the operation of the functions of $\mathbf{Q}_0$ on to a vector $\mathbf{x}_0=[p_i]$ which is not an eigen-vector of $\mathbf{Q}_0$ and hence not included in Table 1. However it is shown in Narain (1969), with the help of the probability generating function, that on expanding this function as a power series in the selection coefficient, the term in s vanishes. This shows that the moments of the distribution of time to fixation of the particular allele, disregarding the cases in which it is lost, is independent of terms in s. For small values of s, therefore, the diffusion approximations to the mean and variance of the time to fixation of a neutral mutant would be quite satisfactory. These are derived in Kimura and Ohta (1968) and Narain (1969). When the initial gene frequency tends to zero, the mean is found to be $4N_e$ and variance as $16N_e^2(\frac{\pi^2}{3}-3)$ with a coefficient of variation of about 54 per cent.

The evaluation of matrix formulae on the computer requires that the population size N and the selective coefficient of the gene s be specified. But it is known (Robertson 1960 and Hill and Robertson 1966) that, under the conditions in which the diffusion approximation holds, the time scale of the selection process is proportional to N and, if the time is measured in units of N, the pattern of the selection process is determined by the parameter Ns for a given initial gene frequency. The mean time has, therefore, been expressed in units of N. In order to see whether the mean time divided by N and the coefficient of variation for fixation of a gene, disregarding the cases in which it is lost, is a function of Ns only, a comparison of these for a few population sizes at a particular value of Ns are shown in Tables 3 and 4 respectively. There is found to be a fair degree of stability in these quantities with respect to variations in N at a fixed value of Ns. The dependence of the mean time and the coefficient of variation on Ns has, therefore, been shown graphically in Figs. 1 and 2 respectively for initial gene frequencies 0.0312, 0.5000 and 0.9687. Both the mean and the coefficient of variation decrease as Ns increases. For a fixed Ns, however, the mean time is highest at low initial gene frequency and lowest at high initial gene frequency whereas the coefficient of variation is highest at high gene frequency and lowest at low gene frequency. When $Ns=1$ and the initial gene frequency is 0.5, the mean time is about $2N$ with a coefficient of variation of about 70 per cent, but for a low initial gene frequency of 0.0312, the mean rises to about $3N$ with a coefficient of variation of about 50 per cent.

From computer runs covering $N=2$ to 16, it was found that with no selection the mean time to fixation divided by N and its coefficient of variation were independent of N though dependent on initial gene frequency, as illustrated in Figs. 3 and 5. The limiting value of the mean time to fixation and its coefficient of variation as the gene frequency approaches zero are $4N$ and 55 per cent respectively, close to the results from the diffusion approach.

COMPARISONS OF TIMES TO FIXATION, LOSS AND HOMOZYGOSITY

It is interesting to compare the mean and the variability of the time to fixation of a particular allele with those of the time to loss and to homozygosity. From the diffusion approach, the results of Ewens (1964) and Kimura and Ohta (1968) provide the following expressions for a neutral gene with initial frequency p.

Mean time until fixation $\quad -4N_e\left(\frac{1-p}{p}\right)Log_e(1-p)$

Mean time until loss $\quad -4N_e\left(\frac{p}{1-p}\right)Log_e\, p$

Mean time until homozygosity $\quad -4N_e\,[p\,Log_e\,p+(1-p)\,Log_e(1-p)]$

In these formulae N_e is the variance effective number (Crow and Morton 1955) which may differ from the actual population number N if the matings is not random or if the distribution of the number of offspring does not follow a Poisson distribution. When $p=1/(2N)$ and N is very large, the mean time until fixation is close to $4N_e$ while that until loss is $\frac{2N_e}{N}\,Log_e\,2N$ so that the mean time until homozygosity should approach the value $\frac{2N_e}{N}\,Log_e\,2N$ as N increases. These comparisons, on the basis of transition matrix methods, are presented in Figs. 3 and 4 for $Ns=0$ and 2 respectively. The dotted lines refer to time until fixation and until loss (marked respectively 1 and 2 in the figures) whereas solid line (marked 3 in the figures) refers to time until homozygosity. When $Ns=0$ and initial frequency of gene is 0.5, all the three curves give the same value of about $2.55N$ as against the diffusion approximation of $-4N_e\,Log_e\,\frac{1}{2}=2.8N_e$. The diffusion approximation, therefore, over-estimates the mean time. For initial gene frequencies greater than or less than 0.5, the three curves differ and, as expected, the curve 3 always lies between the other two. For $p<0.5$ the mean time until homozygosity is nearer to the mean time until loss than to that until fixation so that when $p=1/(2N)$, the mean time until loss is $0.79N$ as against that until homozygosity of $1.02N$. The difference is expected to decrease as the computer results from higher values of N are compared. In the limit, these should, therefore, be the same as expected from the diffusion approximation. For $p>0.5$, the mean time until homozygosity gets nearer to the mean time until fixation. When $Ns=2$, the mean time until homozygosity is found to be practically the same as the mean time until fixation unless the gene frequency is well below 0.5. It is interesting to observe that, while with no selection the maximum mean time until homozygosity occurs at $p=0.5$, with selection this maximum shifts and occurs at gene frequency less than 0.5. This shift has further been found to increase as selection becomes more intense.

The comparisons of the coefficients of variation are shown in Figs. 5 and 6 for $Ns=0$ and 2 respectively. As before, dotted lines refer to time until fixation (1) and loss (2) whereas solid line (3) refers to homozygosity. When $Ns=0$ and

initial gene frequency is 0.5, all the three curves give the same value of about 76 per cent. For $p<0.5$, the coefficient of variation for time until homozygosity is more near the Coeff of variation for time until loss than that until fixation. For $p>0.5$, it gets nearer to that until fixation. When $Ns=2$, the coefficients of variation for homozygosity and fixation are practically the same unless p is well below 0.5. It is interesting to note that the coefficient of variation of time until homozygosity is minimum at $p=0.5$ with no selection but this minimum shifts to $p<0.5$ with selection.

SUMMARY

Previous algebraic treatment of selection processes in finite populations involving single locus have assumed that time could be treated as a continuous variable. Answers are then obtained in the form of solutions to differential equations. In this paper, generations are assumed to be discrete and selection is described in terms of a matrix of transition probabilities. General equations are presented for the chance of fixation or loss of a gene and for the mean and variance of the time to fixation or loss. If each transition probability is expanded as a power series in the selective advantage s, the transition matrix can be considered as similarly expanded, the coefficients now being matrices. It is then possible to obtain expansions for the chance of fixation and the time to fixation as a power series in s. These are obtained and compared with the results from the continuous approach. Finally, using computers to carry out the matrix operations, values are obtained for the times to fixation or loss in specific instances for comparison with the series expansion.

TABLE 1. MATRIX OPERATORS AND VECTORS

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{Q}_o$	$\lambda_1 \mathbf{x}_1$	$\lambda_2 \mathbf{x}_2$	$\lambda_3 \mathbf{x}_3$
$\mathbf{T}_o$	$2N \mathbf{x}_1$	$\frac{2N^2}{3N-1} \mathbf{x}_2$	$\frac{4N^3}{12N^2-11N+3} \mathbf{x}_3$
$\mathbf{Q}^t_o$	$\lambda_1^t \mathbf{x}_1$	$\lambda_2^t \mathbf{x}_2$	$\lambda_3^t \mathbf{x}_3$
$\mathbf{Q}'_o$	$\frac{1}{2}\lambda_1 \mathbf{x}_2$	$\frac{1}{2}\lambda_2\left(6\mathbf{x}_3 - \frac{N}{5N-3}\mathbf{x}_1\right)$	...
$\mathbf{T}_o \mathbf{Q}'_o \mathbf{T}_o$	$\frac{2N^2}{3N-1}\lambda_1 \mathbf{x}_2$	$\frac{2N^2}{3N-1}\lambda_2\left(\frac{12N^2}{12N^2-11N+3}\mathbf{x}_3 - \frac{N}{5N-3}\mathbf{x}_1\right)$	...
$(\mathbf{Q}^t_o)'$	$\frac{1}{2}\lambda_1 W_1 \mathbf{x}_2$	$\frac{1}{2}\lambda_2\left(6 W_3 \mathbf{x}_3 - \frac{N}{5N-3} W_1 \mathbf{x}_1\right)$	...
$\mathbf{Q}''_o$	$\frac{1}{8}\lambda_1\left(6\mathbf{x}_3 - \frac{N}{5N-3}\mathbf{x}_1 - \mathbf{x}_2\right)$	...	...

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{T}_o \mathbf{Q}''_o \mathbf{T}_o$	$\frac{1}{2}N^2\lambda_1\left(\frac{12N^2}{12N^2-11N+3}\mathbf{x}_3 - \frac{N}{5N-3}\mathbf{x}_1 - \frac{N}{3N-1}\mathbf{x}_2\right)$	...	...
$(\mathbf{Q}^t_o)''$	$\frac{1}{8}\lambda_1\left[6\left(\frac{\lambda_2}{\lambda_2-\lambda_3}W_1 - \frac{\lambda_2}{\lambda_1-\lambda_2}W_3 + \frac{\lambda_2^2-\lambda_1\lambda_3}{(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)}W_2\right)\mathbf{x}_3\right.$	...	...
	$\left. - W_1\mathbf{x}_2 - \frac{N}{5N-3}\left(t\lambda_1^{t-1}\frac{(\lambda_1+\lambda_2)}{(\lambda_1-\lambda_2)} - \frac{2\lambda_2}{(\lambda_1-\lambda_2)}W_1\right)\mathbf{x}_1\right]$		

N. B. $\mathbf{T}_o = (\mathbf{I} - \mathbf{Q}_o)^{-1}$

$$W_1 = \left(\frac{\lambda_1^t - \lambda_2^t}{\lambda_1 - \lambda_2}\right) \qquad W_2 = \left(\frac{\lambda_1^t - \lambda_3^t}{\lambda_1 - \lambda_3}\right)$$

$$W_3 = \left(\frac{\lambda_2^t - \lambda_3^t}{\lambda_2 - \lambda_3}\right)$$

TABLE 2

VALUES OF THE α-COEFFICIENTS

	Finite N	$N\to\infty$, $s\to 0$, $Ns=$ Constant
α_{10}	$Ns\left[1-\frac{s}{4}-\left(\frac{N-1}{5N-3}\right)\left(\frac{8N^3-7N+2}{3N-1}\right)\frac{s^2}{8}\right]$	$Ns\,(1-N^2s^2/15)$
α_{20}	$N^2s^2\left[\frac{4N-1}{4(3N-1)}\left(1-\frac{s}{2}\right)\right]$	$\frac{1}{3}N^2s^2$
α_{30}	$N^3s^3\left[\frac{3(4N-1)(N-1)(2N-1)}{2(3N-1)(12N^2-11N+3)}+\frac{3N-1}{12N^2-11N+3}\right]$	$\frac{1}{3}N^3s^3$
α_{11}	$-Ns\left[1-\left(\frac{s}{4}\right)-\frac{(N-1)(2N^2+N-2)+N(2N-1)t}{5N-3}\left(\frac{s^2}{8}\right)\right]$	$-Ns\left[\left(1-\frac{s}{4}-\frac{N^2s^2}{20}\right)e^{-t/2N}-\frac{1}{10}N^2s^2e^{-t/2N}\left(1-e^{-t/2N}\right)\right]$
α_{21}	$-N^2s^2\left[\frac{1}{2}(1-s/2)\right]$	$-\frac{1}{2}N^2s^2$
α_{31}	$-N^3s^3\left[\frac{12N^2-9N+1}{4N(5N-3)}\right]$	$-\frac{3}{5}N^3s^3$
α_{12}	$N^3s^3\left[\frac{(N-1)(2N-1)}{8(5N-3)(3N-1)}\right]$	$\frac{N^3s^3}{60}$
α_{22}	$N^2s^2\left[\frac{(2N-1)}{4(3N-1)}(1-s/2)\right]$	$\frac{1}{6}N^2s^2$
α_{32}	$N^3s^3\left[\frac{(2N-1)}{2(3N-1)}\right]$	$\frac{1}{3}N^3s^3$
α_{33}	$-N^3s^3\left[\frac{48N^5-4N^4-73N^3+74N^2-23N+3}{4N(3N-1)(5N-3)(12N^2-11N+3)}\right]$	$-\frac{1}{15}N^3s^3$

TABLE 3

THE AVERAGE NUMBER OF GENERATIONS $1/N$ UNTIL FIXATION OF A GENE WITH SELECTIVE VALUE s CALCULATED BY THE TRANSITION MATRIX METHOD FOR DIFFERENT POPULATION SIZES (N).

		Initial frequency		
Ns	*N*	0.25	0.50	0.75
16	16	0.46	0.36	0.27
8	16	0.74	0.56	0.40
	8	0.82	0.61	0.43
4	8	1.29	0.94	0.63
	4	1.40	1.02	0.67
2	4	2.03	1.51	0.95
	2	2.07	1.55	1.01

TABLE 4

THE COEFFICIENT OF VARIATION (%) OF THE NUMBER OF GENERATIONS UNTIL FIXATION OF A GENE WITH SELECTIVE VALUE s CALCULATED BY TRANSITION MATRIX METHOD FOR DIFFERENT POPULATION SIZES (N).

		Initial frequency		
Ns	*N*	0.25	0.50	0.75
16	16	28	34	44
8	16	33	39	50
	8	35	42	54
4	8	43	50	64
	4	45	55	69
2	4	53	65	84
	2	55	68	83

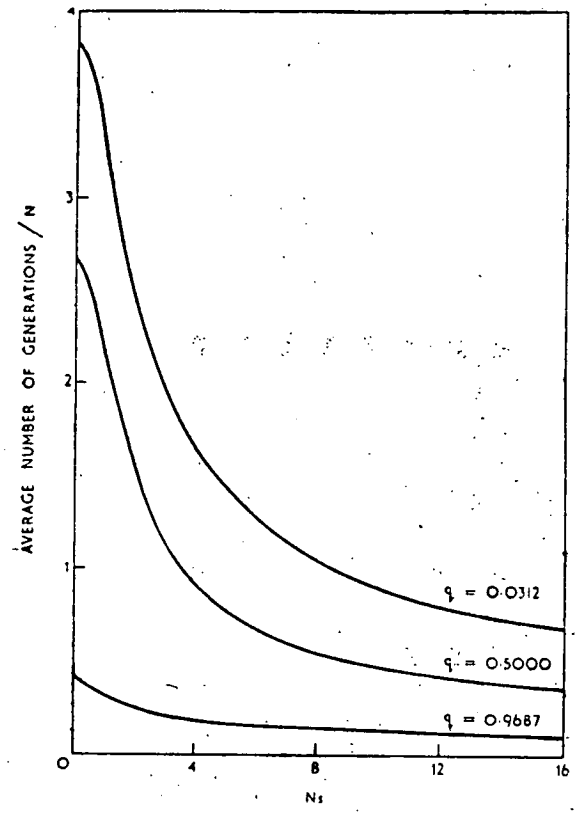

FIG. 1

The average number of generations to fixation divided by the population size for different values of *Ns* and initial gene frequency

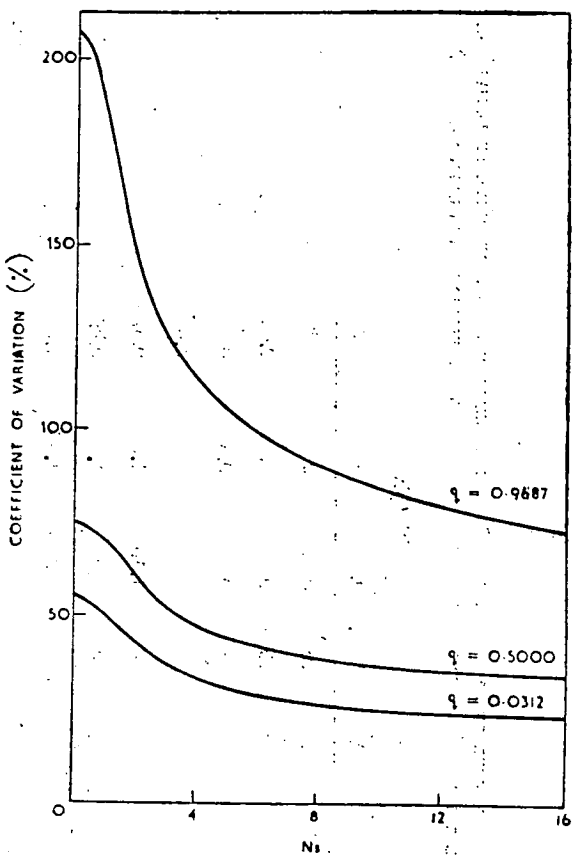

FIG. 2

The coefficient of variation of time to fixation for different values of *Ns* and initial gene frequency

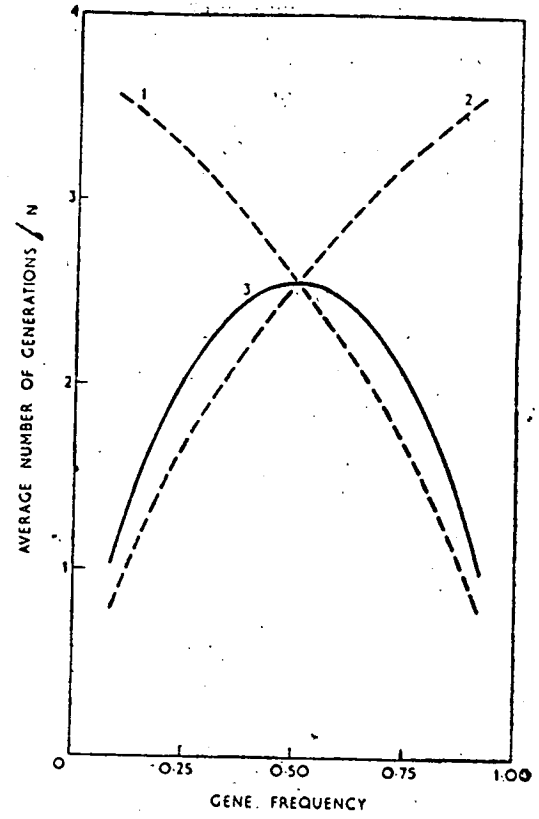

FIG. 3

The mean time to (1) fixation (2) loss and (3) homozygosity respectively, divided by the population size with no selection

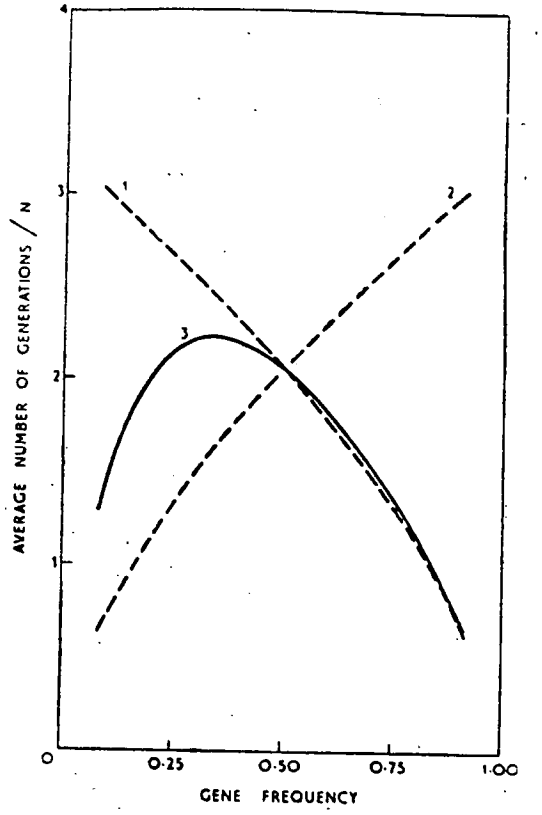

FIG. 4

The mean time to (1) fixation, (2) loss and (3) homozygosity respectively, divided by the population size with $Ns=2$

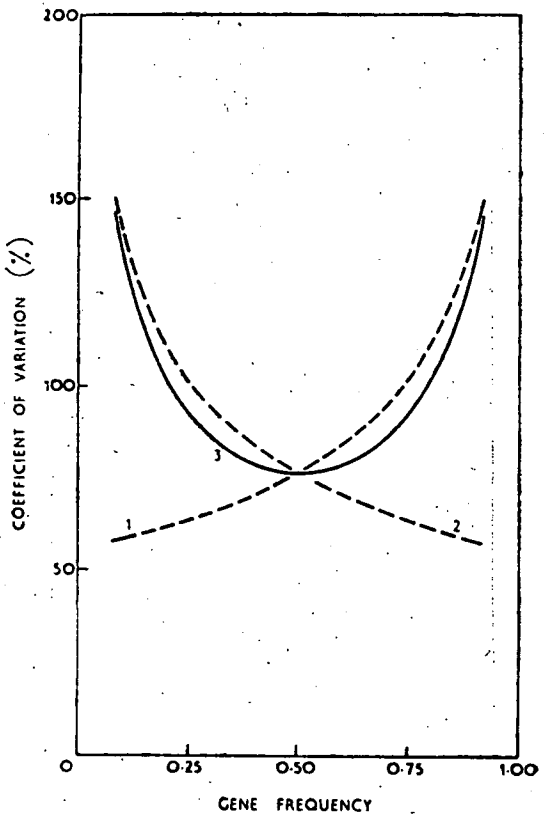

FIG. 5

The coefficient of variation of time until (1) fixation, (2) loss and (3) homozygosity with no selection

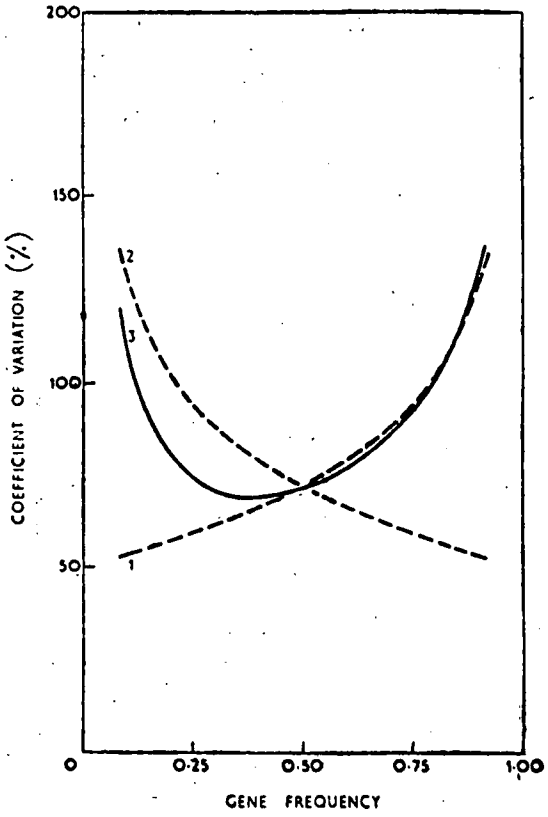

FIG. 6

The coefficient of variation of time until (1) fixation, (2) loss and (3) homozygosity with $Ns=2$

REFERENCES

ALLAN, J. S. AND ROBERTSON, A. 1964. The effect of an initial reverse selection upon total selection response. Genet. Res., Camb. **5**: 68-79.

EWENS, W. J. 1963. Numerical results and diffusion approximations in a genetic process. Biometrika **50**: 241-249.

FADDEEVA, V. N. 1958. *Computational Methods of Linear Algebra.* Dover publications, New York.

FELLER, W. 1951. Diffusion processes in genetics. *Proc. 2nd Berkeley Sym. Mathematical Statistics and Probability.* pp. 227-246. University of California Press.

HILL, W. G. AND ROBERTSON, A. 1966. The effect of linkage on limits to artificial selection. Genet. Res., Camb. **8**: 269-294.

HILL, W. G. AND ROBERTSON, A. 1968. The effects of inbreeding at loci with heterozygote advantage. Genetics **60**: 615-628.

HILL, W. G. 1969. The rate of selection advance for non-additive loci. Genet. Res., Camb. **13**: 165-173.

KEMENY, J. G. AND SNELL, J. L. 1960. *Finite Markov Chains.* Princeton, New York.

KIMURA, M. 1957. Some problems of stochastic processes in genetics. Ann. Math. Statist. **28**: 882-901.

KIMURA, M. 1964. *Diffusion Models in Population Genetics.* Methuen's Review Series in Applied Probability. Vol. 2.

KIMURA, M. AND OHTA, T. 1968. The average number of generations until fixation of a mutant gene in a finite population. Genetics. (In press).

NARAIN, P. 1969. Response to selection in finite populations. Ph.D. thesis, Edinburgh University.

ROBERTSON, A. 1952. The effect of inbreeding on the variation due to recessive genes. Genetics **37**: 199-207.

ROBERTSON, A. 1960. A theory of limits in artificial selection. Proc. Roy. Soc. London B. **153**: 234-249.

WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics **16**: 97-159.

WRIGHT, S. 1945. The differential equation of the distribution of gene frequencies. Proc. Nat. Acad. Sci. **31**: 382-389.

PRINTED IN INDIA AT THE WESLEY PRESS, MYSORE

Genet. Res., Camb. (1970), **15**, *pp.* 251–255
Printed in Great Britain

SHORT PAPERS

A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene

BY P. NARAIN†

Institute of Animal Genetics, Edinburgh, U.K.

(*Received* 1 *May* 1969)

SUMMARY

A general expression is derived for the variance of time to fixation of a neutral gene in a finite population using a diffusion approximation. The results are compared with exact values derived by matrix methods for a population size of 8.

The average number of generations until fixation of a mutant gene in a finite population has recently been given by Kimura & Ohta (1969) with the help of the diffusion model, whereas Narain (1969) has studied the mean and the variance of the number of generations until fixation using transition matrices.

Kimura & Ohta (1969), however, did not derive an expression for the variance of the length of time until fixation, though they did mention that their method could be adapted to obtain the nth moment of the length of time until fixation in terms of the $(n-1)$th moment. The object of this note is, therefore, to derive the variance of the time until fixation using the diffusion approximation.

Let us consider a mutant allele A_2 with frequency p (the normal allele A_1 being at frequency $1-p$) in a diploid population of N individuals with variance effective number N_e, which may differ from N if the mating is not random or if the distribution of the number of offspring does not follow a Poisson distribution. N_e is defined as the size of an idealized population that would have the same variance of change in gene frequency as the population under consideration (Kimura & Crow, 1963). Let $u(p, t)$ be the probability that allele A_2 gets fixed by the tth generation starting with frequency p at $t = 0$. Let

$$T_1(p) = \int_0^\infty t \frac{\partial u(p,t)}{\partial t}\, \mathrm{d}t, \tag{1}$$

$$S_1(p) = \int_0^\infty t^2 \frac{\partial u(p,t)}{\partial t}\, \mathrm{d}t. \tag{2}$$

Then

$$M_1(p) = T_1(p)/u(p), \tag{3}$$

$$V_1(p) = S_1(p)/u(p), \tag{4}$$

represent respectively the average and the second moment about the origin of the length of time until the mutant A_2 becomes fixed in the population, excluding the cases in which it is lost from it. Here $u(p)$ is the probability of ultimate fixation such that

$$u(p) = \lim_{t\to\infty} u(p,t). \tag{5}$$

† On leave from the Institute of Agricultural Research Statistics, (I.C.A.R.), Library Avenue, New Delhi (India).

If $M_{\delta p}$ and $V_{\delta p}$ represent the mean and the variance of the rate of change in the frequency of A_2 per generation, then following Kimura (1962), $u(p,t)$ satisfies the Kolmogorov backward equation

$$\frac{\partial u(p,t)}{\partial t} = \tfrac{1}{2}V_{\delta p}\frac{\partial^2 u(p,t)}{\partial p^2} + M_{\delta p}\frac{\partial u(p,t)}{\partial p}. \tag{6}$$

Following the technique of Kimura & Ohta (1969), the set of differential equations for $T_1(p)$ and $S_1(p)$ are respectively given by

$$\tfrac{1}{2}V_{\delta p}\frac{\mathrm{d}^2T_1(p)}{\mathrm{d}p^2} + M_{\delta p}\frac{\mathrm{d}T_1(p)}{\mathrm{d}p} + u(p) = 0, \tag{7}$$

$$\tfrac{1}{2}V_{\delta p}\frac{\mathrm{d}^2S_1(p)}{\mathrm{d}p^2} + M_{\delta p}\frac{\mathrm{d}S_1(p)}{\mathrm{d}p} + 2T_1(p) = 0. \tag{8}$$

The former differential equation has been derived and solved by Kimura & Ohta (1969) with boundary conditions

$$\lim_{p\to 0} T_1(p) = \text{finite}, \tag{9}$$

$$T_1(1) = 0. \tag{10}$$

The boundary condition (9) means that in a finite population a single mutant gene which appeared in the population reaches fixation within a finite time, whereas (10) is obvious. If we transform (7) into a differential equation for $M_1(p)$ by differentiating

$$T_1(p) = M_1(p).u(p)$$

twice and substituting in (7), we get

$$\tfrac{1}{2}V_{\delta p}\frac{\mathrm{d}^2M_1(p)}{\mathrm{d}p^2} + \left[M_{\delta p} + \frac{V_{\delta p}G(p)}{u(p)}\right]\frac{\mathrm{d}M_1(p)}{\mathrm{d}p} + 1 = 0, \tag{11}$$

$$G(p) = \frac{\mathrm{d}u(p)}{\mathrm{d}p}. \tag{12}$$

Since $\lim_{p\to 0} u(p)$ is finite and $u(1) = 1$, the boundary conditions for the differential equation in $M_1(p)$ are

$$\lim_{p\to\infty} M_1(p) = K_1, \tag{13}$$

where K_1 is a finite quantity $\qquad M_1(1) = 0.$ (14)

In the case of random drift alone, we have

$$M_{\delta_1} = 0, \tag{15}$$

$$V_{\delta p} = p(1-p)/2N_e, \tag{16}$$

$$u(p) = p, \tag{17}$$

$$G(p) = 1. \tag{18}$$

The differential equation (11) then reduces to

$$\frac{\mathrm{d}^2M_1(p)}{\mathrm{d}p^2} + \frac{2}{p}\frac{\mathrm{d}M_1(p)}{\mathrm{d}p} + \frac{4N_e}{p(1-p)} = 0. \tag{19}$$

The solution of this differential equation gives the same result as given by Kimura & Ohta (1969) for a selectively neutral gene. In particular

$$K_1 = 4N_e, \tag{20}$$

$$M_1(p) = -4N_e\left(\frac{1-p}{p}\right)\log_e(1-p). \tag{21}$$

For the second moment about the origin of the length of time until fixation of A_2, we transform (8) into a differential equation for $V_1(p)$ by differentiating $S_1(p) = V_1(p).u(p)$ twice and substituting in (8). This gives

$$\tfrac{1}{2}V_{\delta p}\frac{\mathrm{d}^2V_1(p)}{\mathrm{d}p^2} + \left[M_{\delta p} + \frac{V_{\delta p}G(p)}{u(p)}\right]\frac{\mathrm{d}V_1(p)}{\mathrm{d}p} + 2M_1(p) = 0. \tag{22}$$

It may be noted here that (11) and (22) are similar to the set of differential equations respectively for the mean and the second moment (about the origin) of the length of time until homozygosity (Watterson, 1961, 1962) with the difference that $M_{\delta p}$ has been replaced here by $[M_{\delta p} + (V_{\delta p}G(p))/u(p)]$ as for (7) and (11). The boundary conditions to be imposed are, following the same arguments,

$$\lim_{p\to 0} V_1(p) = K_2, \tag{23}$$

where K_2 is a finite quantity $\qquad V_1(1) = 0. \tag{24}$

In the case of random drift, we apply (15) to (18). The differential equation (22), then, reduces to

$$\frac{\mathrm{d}^2V_1(p)}{\mathrm{d}p^2} + \frac{2}{p}\frac{\mathrm{d}V_1(p)}{\mathrm{d}p} + \frac{8N_eM_1(p)}{p(1-p)} = 0. \tag{25}$$

The solution of (25), after substituting for $M_1(p)$, is given by

$$V_1(p) = B - \frac{A}{p} - 32N_e^2\left[\left(1 - \frac{1}{p} - \log_e p\right)\log_e(1-p) - F(p)\right], \tag{26}$$

where A and B are constants of integration and $F(p)$ is given by

$$F(p) = \int \frac{\log_e p}{1-p}\,\mathrm{d}p. \tag{27}$$

Using the boundary conditions, we get

$$B = A + 32N_e^2F(1), \tag{28}$$

$$K_2 = A - 32N_e^2\left[1 + \int_0^1 \frac{\log_e p}{1-p}\,\mathrm{d}p\right]. \tag{29}$$

Thus $V_1(p)$ is given by

$$V_1(p) = 32N_e^2\left[-\int_p^1 \frac{\log_e p}{1-p}\mathrm{d}p - \left(1 - \frac{1}{p} - \log_e p\right)\log_e(1-p)\right] + \left(1 - \frac{1}{p}\right)\left[K_2 + 32N_e^2\left(\int_0^1 \frac{\log_e p}{1-p}\mathrm{d}p + 1\right)\right]. \tag{30}$$

Since $V_1(p)$ is finite as $p \to 0$, the terms inside the second bracket with the factor $[1-(1/p)]$ in (30) must vanish. This means

$$K_2 = -32N_e^2\left[\int_0^1 \frac{\log_e p}{1-p}\mathrm{d}p + 1\right]. \tag{31}$$

We then get

$$V_1(p) = 32N_e^2\left[-\left(1-\frac{1}{p}\right)\log_e(1-p) + (\log_e p)(\log_e \overline{1-p}) - \int_p^1 \frac{\log_e p}{1-p}\,\mathrm{d}p\right]. \tag{32}$$

Using the results on dilogarithms given in Abramowitz & Stegun (1965), (32) reduces to

$$V_1(p) = 32N_e^2\left[\frac{1-p}{p}\log_e(1-p) + \frac{\pi^2}{6} - \sum_{k=1}^{\infty} p^k \Big/ k^2\right]. \tag{33}$$

The variance is then given by

$$\begin{aligned} V &= V_1(p) - [M_1(p)]^2 \\ &= 32N_e^2\left[\frac{\pi^2}{6} + \left(\frac{1-p}{p}\right)\log_e(1-p)\left\{1 - \frac{1-p}{2p}\log_e(1-p)\right\} - \sum_{k=1}^{\infty} p^k \Big/ k^2\right] \end{aligned} \tag{34}$$

and the coefficient of variation is found to be independent of the effective population size.

Also, in the limit, when $p \to 0$, it follows from Abramowitz & Stegun (1965) that

$$\left.\begin{aligned} \lim_{p\to 0} V_1(p) &= K_2 \\ &= -32N_e^2\left[\int_0^1 \frac{\log_e p}{1-p}\mathrm{d}p + 1\right] \\ &= 32N_e^2[(\pi^2/6) - 1]. \end{aligned}\right\} \tag{35}$$

The variance, in this case, is

$$\left.\begin{aligned} V &= K_2 - K_1^2 \\ &= 16N_e^2[(\pi^2/3) - 3] \\ &= 4{\cdot}64N_e^2, \end{aligned}\right\} \tag{36}$$

giving a coefficient of variation of about 54 %.

DISCUSSION

It is apparent from the preceding derivations that an originally rare neutral mutant gene in a population of effective size N_e takes about $4N_e$ generations on an average, with a standard deviation of about $2N_e$ generations, until it spreads in the whole population. According to Kimura & Ohta (1969) neutral mutation and random drift are of fundamental importance in determining the genetic structure of Mendelian populations. The time that is required for establishing the mutant may average four times the effective population size but may also vary considerably.

Using computer results based on transition matrices it was shown in Narain (1969) that the coefficient of variation of 54 % is obtained when p tends to zero and that it is the minimum possible. It increases as the gene frequency increases. The standard deviation, however, decreases as the gene frequency increases. The following table shows the values of the mean and the standard deviation of time until fixation of A_2 as calculated from (21) and (39) and expressed as multiples of the effective population size for various initial frequencies of A_2.

It is apparent that a larger mean is associated with a larger standard deviation.

The diffusion approximations to the mean and the standard deviations given in Table 1 have been compared with the exact values obtained by the transition matrix approach as developed in Narain (1969). The comparisons are shown in Table 2 for a population of size 8.

Table 1. *Mean and standard deviation of time until fixation of mutant A_2 in terms of the effective population size*

p	$M_1(p)$	$\sqrt{[V(p)]}$
Tending to zero	4·00	2·1536
0·125	3·74	2·1489
0·250	3·45	2·1302
0·375	3·13	2·0947
0·500	2·77	2·0342
0·625	2·35	1·9344
0·750	1·85	1·7678
0·875	1·19	1·4502
1·000	0·00	0·0000

Table 2. *Exact values and diffusion approximation (D.A.) for mean and standard deviation of the number of generations until fixation of A_2 ($N_e = 8$)*

	Mean			Standard deviation		
p	Exact	D.A.	Difference	Exact	D.A.	Difference
0·125	28·30	29·92	1·62	16·58	17·19	0·61
0·250	26·06	27·60	1·54	16·46	17·04	0·58
0·375	23·57	25·04	1·47	16·19	16·76	0·57
0·500	20·76	22·16	1·40	15·73	16·27	0·54
0·625	17·50	18·80	1·30	14·96	15·48	0·52
0·750	13·60	14·80	1·20	13·65	14·14	0·49
0·875	8·55	9·52	0·97	11·14	11·60	0·46

It is quite clear from Table 2 that the diffusion approximations overestimate both the mean and the standard deviation. However, while the overestimation for the mean is, on an average, about one generation, it is only about half-generation for the standard deviation. The former observation is consistent with that obtained by Ewens (1963) in regard to the transition matrix results and diffusion approximation for the mean time until homozygosity. He also observed that the mean error in diffusion approximations is approximately unity.

The author is grateful to Drs M. Kimura and T. Ohta for permission to read the manuscript of their paper before publication and to Professor A. Robertson, and Dr W. G. Hill for valuable discussions.

REFERENCES

Abramowitz, M. & Stegun, I. A. Ed. (1965). Handbook of Mathematical Functions. Dover Publications, New York.

Ewens, W. J. (1963). Numerical results and diffusion approximations in a genetic process. *Biometrika* **50**, 241–249.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

Kimura, M. & Crow, J. F. (1963). The measurement of effective population number. *Evolution* **17**, 279–288.

Kimura, M. & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763–771.

Narain, P. (1969). Response to selection in finite populations. Ph.D. thesis, Edinburgh University.

Watterson, G. (1961). Markov chains with absorbing states (a genetic example). *Ann. Math. Statist.* **32**, 716–729.

Watterson, G. (1962). Some theoretical aspects of diffusion theory in population genetics. *Ann. Math. Statist.* **33**, 939–957.

SABRAO NEWSLETTER
3(2), 135-142, 1971

AVERAGE NUMBER OF GENERATIONS REQUIRED TO ATTAIN LIMITS OF GENETIC IMPROVEMENT

P. NARAIN

(Received April 23, 1971)

The selective capacity of the stock handled by a breeder depends upon the amount of genetic variability present in the stock. The response to selection is, therefore, predicted with the help of an estimate of the genetic variability and selection differential. However, this prediction of response to selection is only for a short term in the sense that it cannot predict the response after several generations of selection. For this purpose it is necessary to know the factors which influence the limit of response to selection and also how many generations of selection are required to achieve the limit. Robertson (1960) developed a theory of limits in artificial selection for predicting the limit of response to selection. Narain and Robertson (1969) developed this theory further so as to predict the average number of generations needed to attain the limit.

Both the limit of selection and the average number of generations required to attain the limits are artifacts of the intensity of selection and the population size. The study of these limits and the time taken to achieve them can also give us some information about the genetic variability in the base population. If the response to selection is mostly due to the fixation of genes present at low frequencies in the initial population, then subjecting the population to a size "bottleneck" of one pair of parents should drastically affect the selection limits and the time taken to attain them. It is, therefore, apparent that the gene frequencies and hence the genetic variability in the base population can be indirectly inferred by measuring the effect of initial restriction of population size on the limits of response to selection and on the time taken to attain these limits.

The theory of limits of genetic improvement and the average number of generations required to attain limits of genetic improvement in artificial selection can therefore be of use to breeders in two ways. Firstly, with the knowledge of the population size and the intensity of selection, the breeder knows in advance the limits of genetic improvement as well as the number of

1. Institute of Agricultural Research Statistics, Library Avenue, New Delhi-12, India

generations required to attain them. Secondly, if the results of introducing an initial restriction of population size on the limits and on the time taken to attain these limits are available, the gene frequencies and hence the genetic variability in the base population can be estimated.

The knowledge of the gene frequencies in the base population can be very helpful to a breeder in sampling, new desirable genes which, by chance, have been lost from the plateaued selection lines. So far as the basic theory of limits useful to the breeder is concerned, Robertson (1960) gives the relevant details. In this article, however, the theory of average number of generations required to attain these limits is discussed briefly with simple applications so as to be of use for breeders. A general and detailed theory in this connection is already reported in Narain (1969) and Narain and Robertson (1969). The estimates of genetic improvement per generation with the help of the limits of selection and the average time required to attain them have also been numerically studied with respect to a single locus with two alleles in this paper. All along, it is assumed that the population under consideration is panmictic so that the theory is relevant to the breeding of animals and outcrossing plants.

Method of Transition Matrix

A general theory of the transition matrix approach to the determination of the chance of fixation of a gene and the average time needed for it in any genetic situation has been developed by Narain and Robertson (1969). Consider a finite population of gametes of constant size 2N in a population of diploid individuals of size N and a single locus with two alleles A and a. The sampling of gametes due to finite size of the population causes a random change, from generation to generation, in the frequency of gene A. The population is subdivided into several lines with different gene frequencies and randomly breeding within lines. These lines can be of three kinds, *i. e.*, lines having gametes A or a only and those in which gametes A and a are segregating. In the third case there are $2N-1$ situations corresponding to the $2N-1$ alternative ways of the mixture of A and a. Lines of the third kind are capable of transition to the same kind or to the first two kinds. When a line attains the state of either of the first two kinds, it is said to be fixed for gene A or a. Theoretically, if P_{ij} represents the conditional probability that there are j A genes out of 2N genes in the line, given that there were i A genes in this line in the previous generation, the $(2N+1)\times(2N+1)$ transition probabilities determine a transition matrix **P** given by

$$\mathbf{P}=\begin{bmatrix}\mathbf{I} & \mathbf{O}\\ \mathbf{R} & \mathbf{Q}\end{bmatrix},$$

where **Q** denotes the transition matrix for transitions between lines of the third kind only, **I** is a 2×2 unit matrix, **O** is a $2\times(2N-1)$ matrix of zeros and **R** is a

$(2N-1)\times 2$ matrix of one-step fixation probabilities of A and a, which can be denoted by $(2N-1)\times 1$ column vectors, $\mathbf{u}(1)$ and $\mathbf{u}(0)$, that is,

$$\mathbf{R}=[\mathbf{u}(0)\ \mathbf{u}(1)]$$

The various elements of a vector correspond to various initial gene frequencies in the base population.

Consider the matrix sum

$$\mathbf{T}=\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\cdots\cdots=(\mathbf{I}-\mathbf{Q})^{-1}.$$

The j^{th} element in the i^{th} row of $\mathbf{T}$ gives the expected total number of generations the population spends in a line with j A genes out of 2N genes on the way to eventual fixation or loss of A, having started from a line with i A genes of 2N genes. The eventual fixation probability of A is then the sum of the expected total number of generations the population spends in the various lines of the third kind in which both A and a gametes are segregating, multiplied by the corresponding probability of fixation in one step. In other words, if $\mathbf{u}(A)$ denotes the vector of the eventual fixation probability of A,

$$\mathbf{u}(A)=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}(1)$$

Further, since the proportion of times a population goes from a particular segregating line into fixation is given by the elements of a particular row of vector $\mathbf{u}(A)$, the vector $\mathbf{m}(A)$ given by

$$\mathbf{m}(A)=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}(A)=(\mathbf{I}-\mathbf{Q})^{-2}\mathbf{u}(1)$$

gives the expected total number of steps in the process which ends with the fixation of A. The mean time until fixation of gene A is therefore given by the ratio of the elements of vectors $\mathbf{m}(A)$ and $\mathbf{u}(A)$. The expected change in the gene frequency of A at the limit, in other words, the selection limit is obtained by subtracting its initial gene frequency from the fixation probability.

Similarly, we have, for the chance of fixation of gene a and the mean time needed for it,

$$\mathbf{u}(a)=(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}(0) \quad \text{and} \quad \mathbf{m}(a)=(\mathbf{I}-\mathbf{Q})^{-2}\mathbf{u}(0).$$

This shows that the computational procedure for the evaluation of the mean time until fixation in a given genetic situation requires the inversion of matrices. These can be easily carried out by a computer. The computation, however, depends on the specification of the P_{ij}s probabilities which depend on the particular genetic situation and the model of sampling. Usually we use the binomial sampling model given by Wright (1931). P_{ij} is given by

$$P_{ij}=\binom{2N}{j}(p_i')^j(1-p_i')^{2N-j}$$

where p_i' is the gene frequency of A after one generation of selection starting with its gene frequency in the previous generation as $p_i=i/(2N)$. The i^{th} element of vectors $\mathbf{u}(1)$ and $\mathbf{u}(0)$ are $(p_i')^{2N}$ and $(1-p_i')^{2N}$, respectively.

In these formulas, p_i' depends on the genetic situation considered, that is, additive, dominance or recessive case. Expressions for p_i' in these cases, with a given selective advantage s of gene A can be found in text-books on population genetics; vide Li (1955) or Falconer (1960).

With the help of the above theory, the limit of selection and the average time needed to attain it were evaluated numerically on a computer for additive, recessive and dominant genes. The estimates of genetic progress per generation were then obtained by taking the ratio of the limit of selection and the average time needed to attain it.

Results

The evaluation of the selection limits and the average time needed for accomplishing them with the help of the matrix formulas by a computer requires that the population size N and the selective advantage of the A gene, s, be specified. But it is known from the investigations carried out by Robertson (1960), and by Hill and Robertson (1966), that the time scale of the selection process is proportional to N, and that if the time is measured by N, the pattern of the selection process is entirely determined by the parameter Ns for a given initial gene frequency. The average time has therefore been expressed by N.

The dependence of selection limits on Ns and initial gene frequency has already been studied by Robertson (1960). The dependence of the mean time on Ns has been shown graphically in Fig. 1 (a, b and c) for additive, recessive and dominant genes. In each case the graphs have been shown for three initial gene frequencies, *viz.*, 0.0312, 0.5000, and 0.9687. It is apparent from these graphs that the mean time decreases as Ns increases in so far as an additive or recessive gene is concerned. But for a dominant gene, the mean

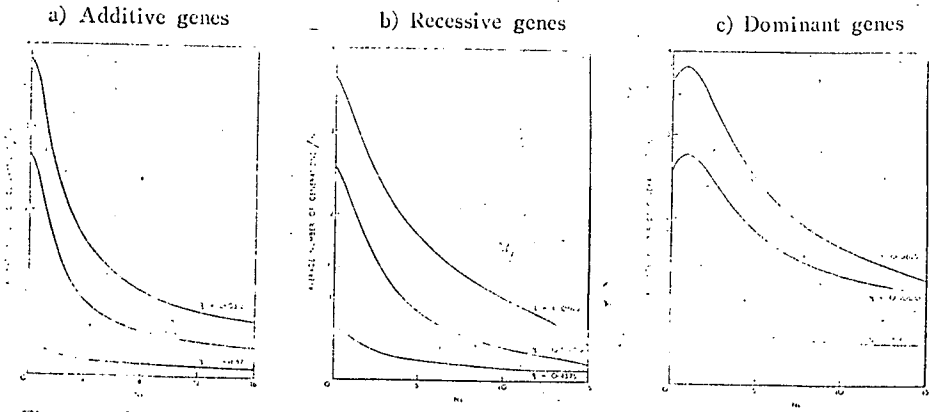

Fig. 1. Average number of generations required for fixation divided by population size, for different values of Ns and initial gene frequency.

time increases initially for small values of Ns and then decreases thereafter. In this case the maximum occurs at $Ns=1$ if the initial gene frequency is 1/2. For a gene with low initial frequency, however, the maximum occurs at a value of Ns less than 1, and for high initial gene frequency it occurs at a value of Ns greater than 1. When a value of Ns is fixed, the mean time is highest at lowest initial frequency, and lowest at highest initial gene frequency. When Ns is 1 and the initial gene frequency is half, the mean time is about 2.25 N, 2.14 N and 2.77 N, respectively for additive, recessive and dominant genes. A gene, therefore, takes less time to reach fixation when it is recessive than when it is dominant.

The estimates of genetic progress per generation in respect of recessive, additive, and dominant genes are shown in Table 1 for $s=0.125$ and $s=0.250$. These estimates have been worked out for each of the three initial gene frequencies 0.0625, 0.5000 and 0.9375 assuming the population size N=8. For each of these cases the genetic change per generation when the population had been infinite has also been included for comparison with N=8. The values for $N=\infty$ can be obtained with the help of the expressions for change in gene frequency after one generation of selection as given by Li (1955) or by Falconer (1960) which depend on s only.

Table 1. Estimates of genetic change per generation

Initial gene frequency	Value of s	Recessive		Additive		Dominant	
		N=8	N=∞	N=8	N=∞	N=8	N=∞
.0625	.125	.00166	.00046	.00259	.00366	.00362	.00686
	.250	.00386	.00092	.00699	.00732	.00935	.01373
.5000	.125	.01298	.01562	.01193	.01562	.01088	.01562
	.250	.02736	.03125	.02328	.03125	.01928	.03125
.9375	.125	.01004	.00686	.00841	.00366	.00593	.00046
	.250	.01866	.01373	.01351	.00732	.00911	.00092

It is apparent from these results that the estimates of genetic change per generation in finite populations are either lower or higher than the genetic change expected in an infinite population depending only on the initial gene frequency and the type of gene action, irrespective of the value of s. For initial gene frequency equal to half, the finite population estimates are lower, but for high initial gene frequencies, they are greater than the corresponding infinite population values regardless of the type of gene action. For rare genes, however, the finite population estimates are greater than the infinite population values only in the case of recessive genes. In the case of a rare recessive gene, the effect of finite population is to increase the estimates of genetic

change per generation over the expected infinite population value by as much as 260% and 320%, respectively when s equals 0.125 and 0.250, respectively. Similarly, a very frequent dominant gene in a finite population gives very high estimates of genetic change per generation compared to infinite population values. For rare additive and dominant genes, the lowering of estimates in finite population is not much, the order of decrease ranging between 5% to 47%. But for very frequent recessive and additive genes, however, the order of increase ranges between 36% to 130%. When a gene is equally frequent, the order of decrease for recessive, additive and dominant genes ranges between 12% to 38%.

Discussion

The behavior of the average time until fixation with variation in Ns is found to be inversely proportional to that of chance of fixation. The chance of fixation of a gene increases but the average time needed for it decreases as Ns increases. This phenomenon has a direct bearing on the estimation of genetic progress per generation obtained by taking the ratio of chance of fixation minus initial gene frequency and the average time. For a gene at low initial frequency, the chance of fixation is smaller but the average time needed is long resulting in very low estimates of genetic progress per generation. On the other hand when the initial gene frequency is high or intermediate, the chance of fixation is substantial and the average time is small which leads to high estimates of genetic change per generation. This has been clearly brought out in Table 1.

It is, however, worth noticing that the estimates of genetic progress follow an increasing trend as we go from dominant to additive and from additive to recessive state provided the initial gene frequencies are higher or intermediate. The trend is, however, reversed if the initial gene frequency is very low. Such a situation also holds true in infinite populations unless the gene frequency is half, when the genetic changes become the same for all the three types of gene action. It is also quite clear from comparisons of the estimates of genetic change per generation with the corresponding infinite population values that the estimate are biased.

The process of random change in gene frequency due to finite size of the population introduces variation in gene frequencies among lines but the average of gene frequencies over lines remains the same as expected in an infinite population. The genetic change per generation in a finite population is therefore expected to be the same as in an infinite population. The bias in the estimation of genetic change per generation is, however, understandable when we

observe that it is the average of the distribution of time to fixation which is in the denominator of the estimates. As shown by Kimura (1970), this distribution is positively skewed. The bias could be due to this.

A highly selected population or a population which has passed through a severe bottleneck in its size would be tolerant to any further size restriction in the sense that the desirable alleles would be hardly lost, because if they are present at all, they would be at a reasonably high frequency. Such populations would, therefore, need a smaller number for maintenance and the limits of genetic improvement would be attained sooner. But in the case of a completely unselected population, the desirable alleles are likely to be at low frequency and, therefore, are easy to be lost by chance unless the population size is kept very large. In such cases longer time would be required for attaining the limit unless the intensity of selection is very high. Unselected populations therefore not only require a larger number for maintenance but also need a substantially higher intensity of selection in order that the limits of genetic improvement are attained sooner.

It has been mentioned in the introduction that the effect of initial restriction of population size on selection limit or the average time needed to attain it are diagonostic of gene frequency in the base population. It is possible, with the help of the graphs presented in this paper, to infer about the frequencies if the average number of generations required to attain the limits, the population size and the intensity of selection are known in a selection programme. With the help of these information, one can determine the relevant points on the graphs and infer what gene frequencies were prevalent in the base population and also what type of gene action, in so far as additiveness or otherwise is concerned, was involved.

Summary

The effects of initial gene frequency, selection intensity and population size on the average number of generations required to attain limits of genetic improvement have been studied by the method of transition matrices. The estimates of genetic progress per generation have been obtained with the help of selection limits and the average time taken to attain them. It is shown that these estimates are considerably biased for rare recessives and dominant genes with high frequency. The effect of restrictions of population size on selection limits and the average time required for their attainment has been discussed in relation to the gene frequency in the base population and the type of gene action involved.

Literature Cited

FALCONER, D. S. (1960) "Introduction to Quantitative Genetics", 365 pp. Oliver and Boyd, Edinburgh.

HILL, W. G. and ROBERTSON, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res., Camb.* 8: 269-294.

KIMURA, M. (1970) The length of the time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet. Res., Camb.* (In press)

LI, C. C. (1955) "Population Genetics", 366 pp. Univ. Chicago Press, Chicago.

NARAIN, P. (1969) Response to selection in finite populations. Ph. D. thesis, Edinburgh University.

NARAIN, P. and ROBERTSON, A. (1969) Limits and duration of response to selection in finite populations: the use of transition probability matrices. *Ind. Jour. Heredity.* 1: 79-97.

ROBERTSON, A. (1960) A theory of limits in artificial selection. *Proc. Roy. Soc. Ser.* B., *Lond.* **153**: 234-249.

WRIGHT, S. (1931) Evolution in Mendelian populations. *Genetics.* **16**: 97-159.

Reprinted from Journal of the Indian Society of Agricultural Statistics, New Delhi Vol. XXIII No. 2 (Silver Jubilee Number) December, 1971

THE USE OF TRANSITION PROBABILITY MATRICES IN STUDIES ON LIMITS OF RESPONSE TO SELECTION

PREM NARAIN
Institute of Agricultural Research Statistic, (I.C.A.R.), New Delhi-12

Introduction

A fundamental problem in population genetics is to describe the changes in the frequency of a gene over time due to systematic forces like selection, mutation and migration. When the size of the given population is very large and the individuals of the population mate at random, the change in the gene frequency is deterministic and can be easily studied by simple algebraic principles. In such cases the population ends up either with the fixation of a favoured gene or with a polymorphic state due to the balancing of various forces. However, when the population is small, consisting of a finite number of individuals, the gene frequency is also subject to fluctuations over time due to random forces created by the random sampling of gametes. The change in the gene frequency over time is then a stochastic process and can only be studied with the help of the mathematical and statistical techniques used for studying a stochastic process. If the rate of change in the gene frequency per generation is very small, the process is approximated by a continuous stochastic process with gene frequency as a random variable lying between 0 and 1. Moreover, in most of the cases encountered, the behaviour of the gene frequency in a generation depends only on its value in the preceeding generation so that the process is Markovian in structure. A Markov process, continuous in gene frequency as well as in time parameter can best be studied with the help of differential equations introduced by Kolmogorov (1931). Making use of these techniques, Wright (1931) and Kimura (1957) respectively gave the concepts of distribution of gene frequency and probability of fixation of a gene. These concepts have proved fundamentals in the theory of population genetics applied to finite populations. In particular, Robertson (1960) made use of these concepts and developed a theory of limits of response to artificial selection with useful applications in animal breeding. When the changes in the gene frequency over time are not small, these cannot be described by a continuous stochastic process. The gene frequency is now a discrete random variable changing by steps between 0 and 1 depending on the population size. The changes in the gene frequency are therefore described exactly by a discrete Markov process with discrete time parameter. Such a process is often referred to as a finite Markov chain and can be studied with the help of transition probability matrices as shown by Narain (1969). Besides translating the concepts of distribution of gene frequency and fixation probability into the transition matrix approach, Narain (1969)

introduced the concept of average time taken for the fixation of a gene. A general theory of limits and duration of response to selection in finite populations was developed subsequently by Narain and Robertson (1969). In this paper an application of the theory to the case of selection at a locus with two alleles, assuming a constant population size, was also studied in detail. Later on, the effect of dominance and recessive nature of the gene on the limits and duration of response was investigated by Narain (1971). In this paper the estimates of genetic change per generation were also obtained with the help of the limits and time taken to achieve them. The transition matrix approach was further used to study the survival of recessive lethals in small populations by Narain (1969) and later by Robertson and Narain (1971). In the later paper, *Monté Carlo* methods were also used to investigate the effect of linkage on the survival of recessive lethals in small populations.

Apart from the above applications of the transition matrix approach to specific genetic problems, a rigorous theory of this approach has not been discussed so far. In this paper, therefore, the basic theory of this approach has been described briefly with special reference to the problem of genetic selection in finite populations. In addition, the theory has been applied to study the effect of linkage on the probability of fixation of a gamete in populations practising self-fertilization.

2. Transition Probability Matrix

Consider a finite population of gametes of size $2N$ and a single locus with two alleles A and a. Such a population can assume $(2N+1)$ states $E_0, E_1, \ldots\ldots E_{2N}$, the i^{th} state E_i respresenting the state of i A genes and $(2N-i)$ a genes. The gene frequency of A, denoted by x_i for the population in the state E_i can then take values, $x_i=i/2N$, $i=0, 1, \ldots\ldots 2N$. The states E_O and E_{2N} represent the states of a and A genes entirely and therefore once the population assumes these states, it gets fixed for either a or A alleles, *i.e.*, the gene frequencies are $x_O=0$ and $x_{2N}=1$ respectively. On the other hand, any state E_i, $i=1, 2, \ldots\ldots(2N-1)$, represents a state segregating for A and a genes and therefore once the population is in such a state, it has a possibility of moving from this state to any other state. Then the gene frequencies of A and a, in this population, are x_i and $(1-x_i)$ respectively. Thus a population of gametes of size $2N$ with two alleles A and a corresponds to a random walk with E_O and E_{2N} as absorbing states and E_i, $i=1, 2, \ldots\ldots(2N-1)$ as transient states.

Let $P_{ij}(t_1, t_2)$ be the conditional probability that the population is in state E_j at time t_1, given that it was in state E_i at time t_2 (less then t_1) *i.e.* it represents the probability of transition from E_i to E_j after a time (t_1-t_2). Mathematically, this means

$$P_{ij}(t_1, t_2)=P\ (x\in E_j \text{ at } t_1/x\in E_i \text{ at } t_2),\ t_1>t_2 \qquad (1)$$

Let the process be homogeneous in time *i.e.* $P_{ij}(t_1, t_2)$ depends only on the difference $(t_1 - t_2)$ and not on t_1 and t_2. We can then denote this probability by $P_{ij}^{(t)}$, representing the probability that the population is in state E_j at time $t+\tau$, greater than t, given that it was in state E_i at time τ for $\tau \geqslant 0$. This is known as t-step transition probability from E_i to E_j so that one-step transition probability can be denoted by P_{ij}. Varying i and j from 0 to $2N$ in steps of 1, we get $(2N+1)\times(2N+1)$ transition probabilities which can conveniently be represented by matrices **P** and **P**(t) for one and t-step transitions respectively. If all the genes are either A or a then for all t, $P^{(t)}_{O,\,O}$ and $P^{(t)}_{2N,\,2N}$ will each be one and $P_O^{(t)}$, $_j$ and $P^{(t)}_{2N,\,j}$ will each be zero for $j=1, 2, \ldots(2N-1)$. Suppose we denote the matrix of transition probabilities associated with transient states by **Q** and **Q**(t) respectively for one and t-step transitions. Further, suppose P_O and $P_O(t)$ represent, respectively for one and t-step transitions, column vectors for transitions from a transient state to E_O. Similarly $P_{2N,}$ and $P_{2N}(t)$ represent the corresponding column vectors for transitions from a transient state to E_{2N}. Then **P** and **P**(t) can be written as

$$\mathbf{P}=\begin{bmatrix} 1 & O' & 0 \\ P_O & \mathbf{Q} & P_{2N} \\ 0 & O' & 1 \end{bmatrix} \tag{2}$$

$$\mathbf{P}(t)=\begin{bmatrix} 1 & O' & 0 \\ P_O(t) & \mathbf{Q}(t) & P_{2N}(t) \\ 0 & O' & 1 \end{bmatrix} \tag{3}$$

The elements of **P**(t) satisfy the condition

$$P_{ij}^{(t)} \geqslant 0 \text{ for all } i, j \tag{4}$$

$$\sum_{j=0}^{2N} P_{ij}^{(t)} = 1 \text{ for all } i \tag{5}$$

Since a transition fram E_i to E_j after t steps means a transition from E_i to E_k in one step and then from E_k to E_j in $(t-1)$ steps, the probabilities of simultaneous realization of these events are

$$P_{ik}\, P_{kj}^{(t-1)} \text{ for } k=0, 1, 2, 3, \ldots\ldots\ldots 2N$$

Hence we have, the Chapman-Kolmogorov equation (Feller 1951)

$$P_{ij}^{(t)} = \sum_{k=0}^{2N} P_{ik}\, P_{kj}^{(t-1)} \tag{6}$$

The corresponding matrix equation is obtained as

$$\begin{aligned}\mathbf{P}(t) &= \mathbf{P}\,\mathbf{P}(t-1)\\ &= \mathbf{P}^2\,\mathbf{P}(t-2)\\ &= \ldots\ldots\ldots\\ &= \mathbf{P}^t\end{aligned} \tag{7}$$

Powering the P-matrix and assuming that inverse of $(\mathbf{I}-\mathbf{Q})$ exists we get the following relations :

$$\mathbf{Q}(t) = \mathbf{Q}^t \tag{8}$$

$$\underline{P_O}(t) = (\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}\,\underline{P_O} \tag{9}$$

$$\underline{P_{2N}}(t) = (\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}\,\underline{P_{2N}} \tag{10}$$

3. Probability of Fixation of a Gene

Let $u_i(t)$ be the probability that at time t, and not sooner, the population with initial gene frequency of A as $i/2N$ becomes fixed for A and let $U_i(t)$ be the probability that it has become fixed for A by t^{th} generation. Then the probabilities of fixation of A with initial gene frequency, $x_i = i/2N$ is given by

$$U_i(t) = \sum_{r=1}^{t} u_i(r) \tag{11}$$

Since fixation at time t in one generation means that transition from initial state to the absorbing state takes place in one step, we have

$$U_i(1) = u_i(1) = P_{i,\,2N} \tag{12}$$

Now fixation at time t can take place in $(2N-1)$ mutually exclusive ways, the k^{th} way being that the initial gene frequency becomes $k/2N$ in the first step and then fixation takes place in $(t-1)$ steps. The probability of simultaneous realization of these two independent events is $P_{ik}\,u_k\,(t-1)$. Hence

$$u_i(t) = \sum_{k=1}^{2N-1} P_{ik}\,u_k\,(t-1) \tag{13}$$

If we denote by $\underline{U}(t)$ and $\underline{u}(t)$ the column vectors of $U_i\,(t)$ and $u_i\,(t)$ respectively for $i = 1, \ldots\ldots\ldots (2N-1)$, we can write these relations as

$$\begin{aligned} \underline{u}(t) &= \mathbf{Q}\underline{u}(t-1) \\ &= \mathbf{Q}^2\underline{u}(t-2) \\ &= \cdots\cdots\cdots \\ &= \mathbf{Q}^{t-1}u(1) \\ &= \mathbf{Q}^{t-1}\underline{P}_{2N} \end{aligned} \tag{14}$$

Then

$$\begin{aligned} \underline{U}(t) &= \sum_{r=1}^{t} \underline{u}(r) \\ &= \sum_{r=1}^{t} \mathbf{Q}^{r-1}\underline{P}_{2N} \\ &= (\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}\underline{P}_{2N} \end{aligned} \tag{15}$$

The expression for $\underline{U}(t)$ is the same as $\underline{P}_{2N}(t)$ given by (10) showing thereby that the fixation probability by t^{th} generation can alternatively be obtained by powering the transition matrix $\mathbf{P}$, t times.

Similarly, if $\underline{I}(t)$ and $\underline{L}(t)$ denote the corresponding vectors for the fixation of gene a, we have

$$\begin{aligned} \underline{L}(t) &= \sum_{r=1}^{t} \underline{1}(r) \\ &= \sum_{r=1}^{t} \mathbf{Q}^{r-1}\underline{P}_O \\ &= (\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}\underline{P}_O \end{aligned} \tag{16}$$

which is the same as $\underline{P}_O(t)$ in view of (9). Also if $\underline{W}(t)$ denotes the vector of probabilities that a population with gene frequency x_i of A is still segregating for it by the t^{th} generation, we have

$$\underline{W}(t) = \mathbf{Q}^t\underline{e} \tag{17}$$

where $\underline{e}$ is a column vector of unities.

As t tends to infinity, we get

$$\underline{U} = \underline{U}(\infty) = (\mathbf{I}-\mathbf{Q})^{-1}\underline{P}_{2N} \tag{18}$$

$$\underline{L} = \underline{L}(\infty) = (\mathbf{I}-\mathbf{Q})^{-1}\underline{P}_O \tag{19}$$

$$\underline{W} = \underline{W}(\infty) = \underline{O} \tag{20}$$

These relations show that ultimately the population is going to be fixed either for A or for a with fixation probabilities given by (18) and (19) respectively.

We can now express (15) and (16) as

$$\underline{U}(t)=(\mathbf{I}-\mathbf{Q}^t)\underline{U} \tag{21}$$

$$\underline{L}(t)=(\mathbf{I}-\mathbf{Q}^t)L \tag{22}$$

4. Expected Change in the Frequency of a Gene

Let the expected frequency of A by the t^{th} generation be denoted by $E[q_i(t)]$ when the initial population had its frequency as $q_i(o)=i/2N$. The expected response in the gene frequency by the t^{th} generation is then

$$E[R_i(t)]=E[q_i(t)]-q_i(o) \tag{23}$$

Let this be represented in vector notation by

$$E[\underline{R}(t)]=E[\underline{q}(t)]-\underline{q}(O) \tag{24}$$

The expected gene frequency by the t^{th} generation, $E[q_i(t)]$ can be obtained by finding the mean of the variate $x_j=j/2\mathrm{N}$ for the distribution given by the i^{th} row of $\mathbf{P}(t)$ i.e.

$$\begin{aligned} E[q_i(t)] &= \sum_{j=0}^{2N} P_{ij}(t)x_j \\ &= \sum_{j=1}^{2N-1} P_{ij}(t)x_j+P_{i,2\mathrm{N}}(t) \end{aligned} \tag{25}$$

In matrix notation, this means

$$\begin{aligned} E[\underline{q}(t)] &= \mathbf{Q}^t\,\underline{q}(O)+\underline{\mathrm{P2N}}(t) \\ &= \mathbf{Q}^t\,\underline{q}(O)+\underline{U}(t) \end{aligned} \tag{26}$$

in view of (10) and (15).

If $\triangle q_i$ denotes the response in the mean gene frequency due to the first generation of selection i.e. the initial response, we have

$$\begin{aligned} q_i(0)+E(\triangle q_i) &= E[q_i(1)] \\ &= \sum_{j=1}^{2N-1} P_{ij}x_j+P_{i,2\mathrm{N}} \end{aligned} \tag{27}$$

In matrix notations, we have

$$\begin{aligned} \underline{q}(0)+E(\underline{\triangle q}) &= E[\underline{q}(1)] \\ &= \mathbf{Q}\underline{q}(0)+\underline{P_{2\mathrm{N}}} \end{aligned} \tag{28}$$

where $E(\underline{\triangle q})$ is the vector of initial expected responses.

This can be manipulated to give

$$(\mathbf{I}-\mathbf{Q})^{-1}\underline{P_{2\mathrm{N}}}-\underline{q}(0)=(\mathbf{I}-\mathbf{Q})^{-1}E(\underline{\triangle q}) \tag{29}$$

Then

$$\begin{aligned} E[\underline{R}(t)] &= \mathbf{Q}^t\underline{q}(0)+\underline{U}(t)-\underline{q}(0) \\ &= -(\mathbf{I}-\mathbf{Q}^t)\underline{q}(0)+(\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}\underline{P_{2N}} \\ &= (\mathbf{I}-\mathbf{Q}^t)[(\mathbf{I}-\mathbf{Q})^{-1}\underline{P_{2N}}-\underline{q}(0)] \\ &= (\mathbf{I}-\mathbf{Q}^t)(\mathbf{I}-\mathbf{Q})^{-1}E(\underline{\triangle q}) \end{aligned} \tag{30}$$

in view of (15) and (29). This shows that the expected response vector by the t^{th} generation is similar to probability fixation vector given by (15) with the difference that P_{2N} in (15) is replaced by $E(\underline{\triangle q})$ in (30). It is also interesting to note that $E(\underline{\triangle q})$ and P_{2N} give respectively the initial expected responses and initial step fixation probabilities.

As t tends to infinity, we get

$$E(\underline{R})=E[\underline{R}(\infty)]=(\mathbf{I}-\mathbf{Q})^{-1}E(\underline{\wedge q}) \tag{31}$$

$$E[\underline{R}(t)] =(\mathbf{I}-\mathbf{Q}^t)E(\underline{R}) \tag{32}$$

which are similar to (18) and (21) respectively. In view of (29) and (18), we also have

$$E(\underline{R})=\underline{U}-\underline{q}(0) \tag{33}$$

This shows that the expected limit of response to selection can, otherwise, be obtained by subtracting the initial gene frequency from the eventual probability of fixation. But it must be noted that the expected response by the t^{th} generation cannot, similarly, be obtained *i. e.* $E[\underline{R}(t)]$ is not equal to $\underline{U}(t)-\underline{q}(0)$ in view of (30). When there is no selection, $E(\underline{\triangle q})=\underline{0}$ and (31) and (33) give the expected result, $\underline{U}=\underline{q}(0)$ i.e. there is no ultimate response, making the fixation probability of a gene equal to its initial gene frequency.

5. Calculation of the probability of fixation and the expected change in the gene frequency

It is apparent from the above derivations that the probability of fixation of a gene and the expected change in the gene frequency by a given number of generations as well as in the limit can be obtained by performing matrix operations on the transition matrix **Q**. However, before these operations, the transition probability P_{ij} is to be specified with the help of the knowledge of the genetic situation involved. Thereafter, matrix functions are to be evaluated either numerically or by analytical methods, for a given population size. For instance, Narain (1969) and Narain and Robertson (1969) used binomial transition probabilities for describing the random sampling of gametes. In some cases they evaluated matrix functions numerically on the computer for a given population size. In some cases, however, the results were also obtained with the help of analytical techniques by using the eigen-roots and eigen-vectors of **Q**. In this section we outline the analytical approach which is applied to a genetic problem studied in the next section.

Let the eigen-roots of $\mathbf{Q}$, a $k \times k$ matrix, obtained by solving the characteristic equation

$$| \mathbf{Q} - \lambda \mathbf{I} | = 0 \tag{34}$$

be given by $\lambda_1, \lambda_2, \ldots\ldots\ldots\ldots\ldots\lambda_k$. Also, let $\underline{x}_i$ and $\underline{y}_i$ be the right and left column vectors respectively corresponding to the root λ_i for $i=1, 2, \ldots\ldots\ldots\ldots k$. The spectral set of matrices for $\mathbf{Q}$ are therefore

$$\mathbf{H}_i = \underline{x}_i \underline{y}_i', \qquad i=1, 2, \ldots\ldots\ldots k. \tag{35}$$

Any function of $\mathbf{Q}$, $f(\mathbf{Q})$, can then be expressed as

$$\begin{aligned} f(\mathbf{Q}) &= \sum_{i=1}^{k} f(\lambda_i)\mathbf{H}_i \\ &= \sum_{i=1}^{k} f(\lambda_i)\underline{x}_i \underline{y}_i' \end{aligned} \tag{36}$$

The formulae for calculating the probability of fixation and the expected change in the gene frequency are then given by

$$\underline{U} = \sum_{i=1}^{k} (1-\lambda)^{-1} \underline{x}_i \underline{y}_i' P_{\underline{2N}} \tag{37}$$

$$\underline{U}(t) = \sum_{i=1}^{k} (1-\lambda_i^t) \underline{x}_i \underline{y}_i' \underline{U} \tag{38}$$

$$E(\underline{R}) = \sum_{i=1}^{k} (1-\lambda_i)^{-1} \underline{x}_i \underline{y}_i' E(\Delta q) \tag{39}$$

$$E[\underline{R}(t)] = \sum_{i=1}^{k} (1-\lambda_i^t) \underline{x}_i \underline{y}_i' E(\underline{R}) \tag{40}$$

6. Effect of linkage on the probability of fixation of a gamete in selfed populations

The case of selfed populations corresponds, in the above discussion, to the situation when $N=1$. The population gets subdivided into lines from each of which two gametes are chosen to form one mature individual only. If we consider a single locus with two alleles A and a, there would be three types of lines. Two of these would have only homozygotes AA and aa respectively and each would occur with frequency 1/4. The third type would have only heterozygotes Aa and would occur with frequency 1/2. It is obvious that the first two types correspond to the absorbing states with frequency of A as 1 and 0 respectively whereas the third type

corresponds to the transient state with frequency of A as 1/2. The matrix **P**, then assumes the form

$$\mathbf{P}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} \tag{41}$$

In this case the matrix **Q** is a scalar quantity equal to 1/2. It is easy to see that (18) and (21) give, for this case, the following results :

$$\text{Probability of fixation of } A \text{ in the limit} = \tfrac{1}{2} \tag{42}$$

$$\text{Probability of fixation of } A \text{ by the } t^{th} \text{ generation} = \tfrac{1}{2}(1-\tfrac{1}{2}^{t}) \tag{43}$$

With two loci each with two alleles $A-a$ and $B-b$ respectively, there would be four gametes AB, Ab, aB, ab. Since two gametes per line are chosen, there would be ten types of lines. Four of these would have only homozygotes respectively as AB/AB, Ab/Ab, aB/aB and ab/ab. These would represent absorbing states. Another four types of lines would contain only single heterozygotes respectively as AB/Ab, AB/aB, Ab/ab, aB/ab. The corresponding states would be transient. The remaining two types of lines would have only double heterozygotes of the types AB/ab (coupling) and Ab/aB (repulsion) respectively and would also correspond to transient states. The linkage between genes at the two loci would affect the contributions of double heterozygotes only. If we denote the probability of crossing-over between two loci by r, with $s=1-r$, the 10 states of the system can be classified according to the distribution amongst the four types of gametes as below :

	AB	Ab	aB	ab
E_1	1	0	0	0
E_2	0	1	0	0
E_3	0	0	1	0
E_4	0	0	0	1
E_5	1/2	1/2	0	0
E_6	1/2	0	1/2	0
E_7	0	1/2	0	1/2
E_8	0	0	1/2	1/2
E_9	s/2	r/2	r/2	s/2
E_{10}	r/2	s/2	s/2	r/2

The P-matrix, then, takes the form :

$$\mathbf{P}=\begin{bmatrix} 1 & 0 & 0 & 0 & \underline{0'} \\ 0 & 1 & 0 & 0 & \underline{0'} \\ 0 & 0 & 1 & 0 & \underline{0'} \\ 0 & 0 & 0 & 1 & \underline{0'} \\ \underline{P_{AB}} & \underline{P_{Ab}} & \underline{P_{aB}} & \underline{P_{ab}} & \mathbf{Q} \end{bmatrix} \tag{44}$$

In this matrix, $\underline{0'}$ is a 1×6 row vector of zeros. $\underline{P_{AB}}$, $\underline{P_{Ab}}$, $\underline{P_{aB}}$, and $\underline{P_{ab}}$, are 6×1 column vectors given by

$$\underline{P'_{AB}}=(1/4,\ 1/4,\ 0,\ 0,\ s^2/4,\ r^2/4) \tag{45}$$

$$\underline{P'_{Ab}}=(1/4,\ 0,\ 1/4,\ 0,\ r^2/4,\ s^2/4) \tag{46}$$

$$\underline{P'_{aB}}=(0,\ 1/4,\ 0,\ 1/4,\ r^2/4,\ s^2/4) \tag{47}$$

$$\underline{P'_{ab}}=(0,\ 0,\ 1/4,\ 1/4,\ s^2/4,\ r^2/4) \tag{48}$$

Further, $\mathbf{Q}$ is a 6×6 matrix given by

$$\mathbf{Q}=\begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 \\ rs/2 & rs/2 & rs/2 & rs/2 & s^2/2 & r^2/2 \\ rs/2 & rs/2 & rs/2 & rs/2 & r^2/2 & s^2/2 \end{bmatrix} \tag{49}$$

The $\mathbf{Q}$-matrix is found to yield the 6 roots λ_1, λ_2, λ_3, λ_4, λ_5 and λ_6 given by

$$\lambda_1=\lambda_2=\lambda_3=\lambda_4=1/2 \tag{50}$$

$$\lambda_5=(1-2rs)/2 \tag{51}$$

$$\lambda_6=(1-2r)/2 \tag{52}$$

The matrix **X** of the right-hand column vectors x_i corresponding to λ_i for $i=1, 2, 3, 4, 5, 6$, is derived from the results given by Puri (1968) and is given by

$$\mathbf{X}=(\underline{x_1}, \underline{x_2}, \underline{x_3}, \underline{x_4}, \underline{x_5}, \underline{x_6})$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \tag{53}$$

Similarly, the matrix **Y** of the left-hand row vectors $\underline{y'_i}$ corresponding to the six roots is given by

$$\mathbf{Y}=(\underline{y_1'}, \underline{y'_2}, \underline{y'_3}, \underline{y'_4}, \underline{y'_5}, \underline{y'_6})'$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -1/\sqrt{2} & -1/\sqrt{2} & -1/\sqrt{2} & -1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \tag{54}$$

With the help of vectors $\underline{x_i}$ and $\underline{y'_i}$, the spectral set of the matrices $\mathbf{H}_i$, $i=1,2,\ldots\ldots 6$ given by (35) are obtained. These matrices, together with the six-roots, provide with the elements of the matrix $(\mathbf{I}-\mathbf{Q})^{-1}$ since

$$(\mathbf{I}-\mathbf{Q})^{-1}=\sum_{i=1}^{6}(1-\lambda_i)^{-1}\mathbf{H}_i$$

$$= \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 1-w & 1-w & 1-w & 1-w & w+v & w-v \\ 1-w & 1-w & 1-w & 1-w & w-v & w+v \end{bmatrix} \tag{55}$$

where $w=1/(1+2rs)$ and $v=1/(1+2r)$.

If the vectors of probabilities of fixation of gametes AB, Ab, aB and ab are denoted respectively by $\underline{U_{AB}}$, $\underline{U_{Ab}}$, $\underline{U_{aB}}$ and $\underline{U_{ab}}$, then, in view of (37) and (45) to (48), we get

$$\underline{U_{AB}} = (1/2 \quad 1/2 \quad 0 \quad 0 \quad v/2 \quad rv)' \tag{56}$$

$$\underline{U_{Ab}} = (1/2 \quad 0 \quad 1/2 \quad 0 \quad rv \quad v/2)' \tag{57}$$

$$\underline{U_{aB}} = (0 \quad 1/2 \quad 0 \quad 1/2 \quad rv \quad v/2)' \tag{58}$$

$$\underline{U_{ab}} = (0 \quad 0 \quad 1/2 \quad 1/2 \quad v/2 \quad rv)' \tag{59}$$

It is confirmed from the above results that the linkage can only have its effect when the population is initially either in state E_9 (coupling phase) or in state E_{10} (repulsion phase). In each of these two situations, the probabilities of fixation of the four types of gametes depend on the recombination fraction r, as shown by the last two elements of each of the vectors given by (56) – (59). As expected, the probability of fixation of gametes AB is the same as that of ab. Similarly, the probability of fixation of Ab is the same as that of aB. The effect of linkage is to increase the probability of fixation of a coupled gamete (AB or ab) if the initial population consists of a double heterozygote in coupling phase. On the other hand with repulsion phase the probability is reduced. For repulsion gametes (Ab or aB), the probability is increased when the initial population consists of a double heterozygote in repulsion phase but decreased with coupling phase of linkage.

7. Summary

A rigorous theory of the transition matrix approach for studying the change in the frequency of a gene in finite populations is developed. The probability of fixation of a gene and the expected change in the gene frequency by a given number of generations as well as in the limit are expressed as functions of the transition probability matrix. The analytical as well as the numerical procedures for the calculation of these quantities are outlined. The theory is applied to study the effect of linkage on the probability of fixation of a gamete in populations practising self-fertilization. It is found that linkage increases or decreases the probability of fixation of a coupled gamete according as the initial population consists of a coupling or repulsion heterozygote respectively.

The Conditioned Diffusion Equation and its Use in Population Genetics

BY

P. NARAIN

Reprinted from

THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY

SERIES B (METHODOLOGICAL)

Volume 36, No. 2, 1974

(pp. 258–266

PRINTED FOR PRIVATE CIRCULATION

1974

The Conditioned Diffusion Equation and its Use in Population Genetics

By P. NARAIN

Institute of Agricultural Research Statistics, New Delhi

[Received October 1972. Revised June 1973]

SUMMARY

The forward and backward conditioned diffusion equations relative to the event of the process attaining absorption in one of the boundaries have been derived from the corresponding Kolmogorov differential equations. The backward conditioned diffusion equation has been used to derive the mean and variance of the length of time until absorption in one of the boundaries. The general results so obtained have been applied to the problem of random drift in population genetics, giving the means and variances of the distributions of time until fixation as well as of time until extinction of a particular gene in a finite population.

Keywords: CONDITIONED DIFFUSION EQUATION; CONDITIONAL DIFFUSION PROCESS IN GENETICS; DIFFUSION APPROACH AND FIXATION OF A GENE; ABSORPTION TIME BY DIFFUSION THEORY

1. INTRODUCTION

WE consider a random variable x at time t whose density function $f(x; t)$ satisfies the forward Kolmogorov (or Fokker–Planck) equation.

$$\frac{\partial f(x;t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{v(x)f(x;\ t)\} - \frac{\partial}{\partial x}\{m(x)f(x;\ t)\}; \tag{1.1}$$

where $m(x)$ and $v(x)$ are the instantaneous drift and diffusion coefficients respectively.

The adjoint of (1.1), known as the backward Kolmogorov equation, is given by

$$-\frac{\partial u(p;\ t)}{\partial t} = \tfrac{1}{2}v(p)\frac{\partial^2}{\partial p^2}u(p;\ t) + m(p)\frac{\partial}{\partial p}u(p;\ t), \tag{1.2}$$

where $u(p; t)$ is the probability that $x = 1$ during a time interval t, given that initially it takes the value p.

In population genetics, we come across situations where we are interested in studying in a population of size N, the behaviour of the frequency of gene A (with a single alternative allele a) over time under the influence of selection, mutation, migration and random sampling of gametes due to finite size of the population. The change in the gene frequency over time is then a stochastic process which can be best studied with the help of the above diffusion equations. The variate x then refers to the gene frequency of A at time t. In such cases, $m(x)$ and $v(x)$ are of the form given by

$$\left.\begin{aligned} m(x) &= x(1-x)\phi(x), \\ v(x) &= x(1-x)/(2N), \end{aligned}\right\} \tag{1.3}$$

where $\phi(x)$ is an arbitrary polynomial, possibly constant or zero. These conditions imply that either A or its alternative allele is fixed in the population in a finite length of time. Mathematically, this means that $x = 0$ and $x = 1$ are exit boundaries such that there is unit probability of one or other boundary being reached in finite time and that once a boundary is reached, the variate remains fixed at that boundary, i.e. the process is said to be absorbed at that boundary. However, in the treatment which follows, we do not impose these conditions until in the final section when we discuss genetical applications.

If the process starts with the value of x as p, the probability $u(p) = \lim_{t\to\infty} u(p; t)$ that the boundary $x = 1$ is reached before the boundary $x = 0$ is a solution of the equation (1.2) with $(\partial u/\partial t) = 0$, i.e. of the equation

$$\tfrac{1}{2}v(p)\frac{d^2u(p)}{dp^2} + m(p)\frac{du(p)}{dp} = 0, \tag{1.4}$$

subject to $u(0) = 0$, $u(1) = 1$. From (1.4), it follows that

$$u(p) = \int_0^p G(x)\,dx \Big/ \int_0^1 G(x)\,dx, \tag{1.5}$$

where

$$\left.\begin{aligned} G(x) &= \exp\left[-2\int\{m(x)/v(x)\}\,dx\right], \\ \frac{du(x)}{dx} &= G(x)\Big/\int_0^1 G(x)\,dx. \end{aligned}\right\} \tag{1.6}$$

The probability that the boundary $x = 0$ is reached before the boundary $x = 1$ is then $1 - u(p)$.

The mean time $M(p)$ until one or other boundary is reached satisfies the differential equation

$$\tfrac{1}{2}v(p)\frac{d^2M(p)}{dp^2} + m(p)\frac{dM(p)}{dp} + 1 = 0, \tag{1.7}$$

subject to $M(0) = M(1) = 0$. Integrating (1.7) and applying the boundary conditions, we obtain

$$M(p) = u(p)\int_p^1 I(x)\{1-u(x)\}\,dx + \{1-u(p)\}\int_0^p I(x)\,u(x)\,dx, \tag{1.8}$$

where

$$I(x) = 2\int_0^1 G(x)\,dx \Big/ \{v(x)\,G(x)\}. \tag{1.9}$$

The second moment (about the origin) of the length of time $V(p)$ until one or other boundary is reached satisfies the differential equation

$$\tfrac{1}{2}v(p)\frac{d^2V(p)}{dp^2} + m(p)\frac{dV(p)}{dp} + 2M(p) = 0, \tag{1.10}$$

subject to $V(0) = V(1) = 0$. Integrating (1.10) and applying the boundary conditions we obtain

$$V(p) = u(p)\int_p^1 J(x)\{1-u(x)\}\,dx + \{1-u(p)\}\int_0^p J(x)\,u(x)\,dx, \tag{1.11}$$

where

$$J(x) = \left\{4\int_0^1 G(x)\,dx\right\}\frac{M(x)}{v(x)\,G(x)}. \tag{1.12}$$

The results mentioned above are discussed in the literature by Kolmogorov (1931), Wright (1931), Feller (1951, 1954), Kimura (1957, 1962, 1964), Watterson (1961, 1962), Ewens (1963, 1969) and Narain (1969, 1970).

2. The Conditioned Diffusion Equations

Let the conditional density function of x at time t, given that its final value at $t = \infty$ will be 1, be denoted by $f_{c1}(x;\, t)$, that is,

$$f_{c1}(x;\, t) = \lim_{t_1\to\infty} f_{c1}(x;\, t, 1;\, t_1), \quad t < t_1. \tag{2.1}$$

Applying Bayes's theorem on conditional probability, we get

$$u(p) f_{c1}(x;\, t) = f(x;\, t)\,u(x). \tag{2.2}$$

Using $u(p)f_{c1}(x;\, t)/u(x)$ for $f(x;\, t)$ in (1.1), we transform the density function $f(x;\, t)$ to $f_{c1}(x;\, t)$ relative to the event of the process attaining absorption at $x = 1$. The transformation from f to f_{c1} gives, on simplification,

$$\frac{\partial}{\partial t}f_{c1} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{v(x)f_{c1}\} - \frac{\partial}{\partial x}\{m(x)f_{c1}\} + v(x)f_{c1}\left\{\frac{d}{dx}\log u(x)\right\}^2 + \left[2m(x)f_{c1} - \frac{\partial}{\partial x}\{v(x)f_{c1}\}\right]\left[\frac{d}{dx}\log u(x)\right]. \tag{2.3}$$

Considering $(\partial/\partial x)\{v(x)f_{c1}(d/dx)\log u(x)\}$, we find

$$-\frac{\partial}{\partial x}\left\{v(x)f_{c1}\frac{d}{dx}\log u(x)\right\} = \left[2m(x)f_{c1} - \frac{\partial}{\partial x}\{v(x)f_{c1}\}\right]\left\{\frac{d}{dx}\log u(x)\right\} + \{v(x)f_{c1}\}\left\{\frac{d}{dx}\log u(x)\right\}^2. \tag{2.4}$$

Substituting (2.4) in (2.3), we get, after simplification,

$$\frac{\partial f_{c1}(x;\, t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{v(x)f_{c1}(x;\, t)\} - \frac{\partial}{\partial x}\left[\left\{m(x) + \frac{v(x)\,G(x)}{u(x)}\right\}f_{c1}(x;\, t)\right]. \tag{2.5}$$

This shows that the density function f_{c1} also satisfies the forward Kolmogorov equation but with the instantaneous drift coefficient $m_1^*(x)$ given by

$$m_1^*(x) = m(x) + \frac{v(x)\,G(x)}{u(x)}. \tag{2.6}$$

The diffusion coefficient $v(x)$, however, does not change. We term (2.5) as the conditioned forward Kolmogorov diffusion equation relative to the event of the process attaining absorption at $x = 1$.

The adjoint of (2.5), which can be termed as the conditioned backward Kolmogorov equation relative to the event of the process attaining absorption at $x = 1$, is then given by (1.2) with $m(p)$ replaced by $m_1^*(p)$ and $u(p; t)$ replaced by $u_{c1}(p; t)$ which is the probability that $x = 1$ during a time interval t, given that initially it takes the value p and relative to the event of the process attaining absorption at $x = 1$. It is obvious then that the probability $u_{c1}(p) = \lim_{t\to\infty} u_{c1}(p; t)$ that the boundary $x = 1$ is reached before the boundary $x = 0$, relative to the event of eventual absorption at $x = 1$, is unity. This can easily be verified by solving the adjoint of (2.5) with $(\partial u_{c1}/\partial t) = 0$, subject to

$$\lim_{p\to 0} u_{c1}(p) = \text{finite} \quad \text{and} \quad u_{c1}(1) = 1.$$

In a similar manner we can derive the conditioned forward and backward diffusion equations relative to the event of the process attaining absorption at $x = 0$.

Let the conditional density function of x at time t, given that its final value at $t = \infty$ will be 0, be denoted by $f_{c0}(x; t)$, that is

$$f_{c0}(x; t) = \lim_{t_1\to\infty} f_{c0}(x; t, 0; t_1), \quad t < t_1. \tag{2.7}$$

Applying Bayes's theorem on conditional probability, we get

$$\{1-u(p)\}f_{c0}(x; t) = f(x; t)\{1-u(x)\}. \tag{2.8}$$

Using (2.8), we transform the density function $f(x; t)$ in (1.1) to $f_{c0}(x; t)$ relative to the event of the process attaining absorption at $x = 0$. This gives

$$\frac{\partial f_{c0}(x; t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{v(x)f_{c0}(x; t)\} - \frac{\partial}{\partial x}\left[\left\{m(x) - \frac{v(x)\,G(x)}{1-u(x)}\right\}f_{c0}(x; t)\right]. \tag{2.9}$$

This shows that f_{c0} also satisfies the forward Kolmogorov equation but with the drift coefficient $m_0^*(x)$ given by

$$m_0^*(x) = m(x) - \frac{v(x)\,G(x)}{1-u(x)}. \tag{2.10}$$

We term (2.10) as the conditioned forward Kolmogorov diffusion equation relative to the event of the process attaining absorption at $x = 0$.

The adjoint of (2.9), termed as the conditioned backward Kolmogorov diffusion equation relative to the event of the process attaining absorption at $x = 0$, is then given by (1.2) with $m(p)$ replaced by $m_0^*(p)$ and $u(p; t)$ replaced by $u_{c0}(p; t)$ which is the probability that $x = 0$ during a time interval t, given that initially it takes the value p and relative to the event of the process attaining absorption at $x = 0$.

The probability $u_{c0}(p) = \lim_{t\to\infty} u_{c0}(p; t)$ that the boundary $x = 0$ is reached before the boundary $x = 1$, relative to the event of eventual absorption at $x = 0$, can be shown to be unity by solving the adjoint of (2.9) with $(\partial u_{c0}/\partial t) = 0$, subject to $u_{c0}(0) = 1$ and $\lim_{p\to 1} u_{c0}(p) = \text{finite}$.

3. MEAN TIME UNTIL ABSORPTION IN ONE OF THE BOUNDARIES

It is apparent that the differential equation for the mean of the distribution of time until absorption in the boundary at $x = 1$, disregarding the cases of absorption at $x = 0$, is given by (1.7) with $m(p)$ replaced by $m_1^*(p)$ and $M(p)$ replaced by $M_{c1}(p)$ where

$$M_{c1}(p) = \int_0^\infty t u_{c1}(p; t)\, dt \tag{3.1}$$

is the mean time until absorption in the boundary at $x = 1$, disregarding the cases of absorption in the boundary at $x = 0$. This was first noticed by Narain (1970) wherein the set of differential equations for the mean and the second moment (about the origin) of the length of time until fixation of a neutral mutant gene, excluding the cases in which it is lost, was obtained from a different approach without invoking a conditioned diffusion equation.

Since $\lim_{p\to 0} U_{c1}(p)$ is finite and $u_{c1}(1) = 1$, the boundary conditions for the differential equation in $M_{c1}(p)$ are

$$\lim_{p\to 0} M_{c1}(p) = K_1, \tag{3.2}$$

where K_1 is a finite quantity and

$$M_{c1}(1) = 0. \tag{3.3}$$

The integration of the differential equation, subject to these conditions, gives

$$K_1 = \int_0^1 I(p)\, u(p)\{1 - u(p)\}\, dp, \tag{3.4}$$

$$M_{c1}(p) = \int_p^1 I(p)\, u(p)\{1 - u(p)\}\, dp + \frac{1 - u(p)}{u(p)} \int_0^p I(p)\{u(p)\}^2\, dp. \tag{3.5}$$

The differential equation for the mean time until absorption in the boundary at $x = 0$ is given by (1.7) with $m(p)$ replaced by $m_0^*(p)$ and $M(p)$ replaced by $M_{c0}(p)$ where

$$M_{c0}(p) = \int_0^\infty t u_{c0}(p; t)\, dt \tag{3.6}$$

is the mean time until absorption in the boundary at $x = 0$, disregarding the cases of absorption in the boundary at $x = 1$.

Since $u_{c0}(0) = 1$ and $\lim_{p\to 1} u_{c0}(p)$ is finite, the boundary conditions for the differential equation in $M_{c0}(p)$ are

$$M_{c0}(0) = 0, \tag{3.7}$$

$$\lim_{p\to 1} M_{c0}(p) = \text{a finite quantity}. \tag{3.8}$$

The integration of the differential equation, subject to these conditions, gives

$$M_{c0}(p) = \int_0^p I(p)\, u(p)\{1 - u(p)\}\, dp + \frac{u(p)}{1 - u(p)} \int_p^1 I(p)\{1 - u(p)\}^2\, dp, \tag{3.9}$$

$$\lim_{p\to 1} M_{c0}(p) = \int_0^1 I(p)\, u(p)\{1 - u(p)\}\, dp. \tag{3.10}$$

The results (3.5) and (3.9) were also obtained by Kimura and Ohta (1969) but from a different approach without involving a conditional diffusion process.

4. Variance of the Length of Time until Absorption in One of the Boundaries

The differential equation for the second moment of the distribution of time until absorption at $x = 1$ is given by (1.10) with $m(p)$ replaced by $m_1^*(p)$, $M(p)$ replaced by $M_{c1}(p)$ and $V(p)$ replaced by $V_{c1}(p)$ where

$$V_{c1}(p) = \int_0^\infty t^2 u_{c1}(p; t)\, dt \tag{4.1}$$

is the second moment (about the origin) of the length of time until absorption in the boundary at $x = 1$.

The boundary conditions for the differential equation in $V_{c1}(p)$ are given by

$$\lim_{p \to 0} V_{c1}(p) = K_2, \tag{4.2}$$

where K_2 is a finite quantity and

$$V_{c1}(1) = 0. \tag{4.3}$$

Integrating the differential equation in $V_{c1}(p)$ twice and applying the above boundary conditions, we obtain

$$V_{c1}(p) = 2\int_p^1 M_{c1}(p)\, I(p)\, u(p)\{1-u(p)\}\, dp + 2\left\{\frac{1-u(p)}{u(p)}\right\}\int_0^p M_{c1}(p)\, I(p)\{u(p)\}^2\, dp, \tag{4.4}$$

$$K_2 = 2\int_0^1 M_{c1}(p)\, I(p)\, u(p)\{1-u(p)\}\, dp. \tag{4.5}$$

Using (4.4) and (3.13), we obtain the variance of the length of time until absorption at $x = 1$ as

$$\text{Variance} = V_{c1}(p) - \{M_{c1}(p)\}^2. \tag{4.6}$$

The differential equation for the second moment of the length of time until absorption at $x = 0$ is given by (1.10) with $m(p)$ replaced by $m_0^*(p)$, $M(p)$ replaced by $M_{c0}(p)$, and $V(p)$ replaced by $V_{c0}(p)$, where

$$V_{c0}(p) = \int_0^\infty t^2 u_{c0}(p; t)\, dt \tag{4.7}$$

is the second moment (about the origin) of the length of time until absorption at $x = 0$ excluding the cases of absorption at $x = 1$. The boundary conditions for the differential equation in $V_{c0}(p)$ are given by

$$V_{c0}(0) = 1, \tag{4.8}$$

$$\lim_{p \to 1} V_{c0}(p) = \text{a finite quantity.} \tag{4.9}$$

Integrating the differential equation in $V_{c0}(p)$ subject to (4.8) and (4.9), we get

$$V_{c0}(p) = 2\int_0^p M_{c0}(p)\, I(p)\, u(p)\{1-u(p)\}\, dp + \left\{\frac{2u(p)}{1-u(p)}\right\}\int_p^1 M_{c0}(p)\, I(p)\{1-u(p)\}^2\, dp. \tag{4.10}$$

Using (4.10) and (3.17), we obtain the variance of the length of time until absorption at $x = 0$.

5. Genetical Applications

It is apparent from the above considerations that the adjoint of the conditioned diffusion equation relative to the event of the process attaining absorption at $x = 1$ (2.5) is particularly useful in population genetics for obtaining the mean and variance of the distribution of time until fixation of gene A, disregarding the cases in which it is lost from the population. Similarly, the adjoint of the conditioned diffusion equation relative to the event of the process attaining absorption at $x = 0$ (2.9) is used for obtaining the mean and variance of the distribution of time until extinction of gene A, disregarding the cases in which it is fixed in the population. Their uses can be illustrated by considering the case of random drift alone.

Let the frequency of gene A be p (with the frequency of its alternative allele a as $1-p$) in a diploid population of N individuals with variance effective number N_e, which may differ from N if the mating is not random or if the distribution of the number of offspring does not follow a Poisson distribution. As defined by Kimura and Crow (1963), N_e is the size of an idealized population that would have the same variance of change in gene frequency as the population under consideration. We assume the absence of directed forces of selection, mutation and migration so that the mean change in the gene frequency over time is zero but the variance of the change in gene frequency follows the binomial law. This means

$$m(p) = 0 \tag{5.1}$$

$$v(p) = p(1-p)/(2N_e). \tag{5.2}$$

This gives, from (1.6), (1.5) and (1.9),

$$G(p) = 1, \tag{5.3}$$

$$u(p) = p, \tag{5.4}$$

$$I(p) = \frac{4N_e}{p(1-p)}. \tag{5.5}$$

Using (2.6),

$$m_1^*(p) = \frac{(1-p)}{2N_e}. \tag{5.6}$$

The differential equation in $M_{c1}(p)$ then reduces to

$$\frac{d^2 M_{c1}(p)}{dp^2} + \frac{2}{p}\frac{dM_{c1}(p)}{dp} + \frac{4N_e}{p(1-p)} = 0, \tag{5.7}$$

as shown in Narain (1970).

With the help of (3.5) and using (5.4) and (5.5) we find the solution of (5.7) as

$$M_{c1}(p) = -4N_e\left(\frac{1-p}{p}\right)\log_e(1-p). \tag{5.8}$$

Similarly,

$$\lim_{p\to 0} M_{c1}(p) = K_1 = \int_0^1 I(p)\,u(p)\{1-u(p)\}\,dp$$
$$= 4N_e. \tag{5.9}$$

Similarly, using (2.10),

$$m_0^*(p) = -p/(2N_e). \tag{5.10}$$

The differential equation in $M_{c0}(p)$ then reduces to

$$\frac{d^2M_{c0}(p)}{dp^2} - \frac{2}{1-p}\frac{dM_{c0}(p)}{dp} + \frac{4N_e}{p(1-p)} = 0, \tag{5.11}$$

For solving this equation, we use (3.9) and obtain,

$$M_{c0}(p) = -4N_e\left(\frac{p}{1-p}\right)\log_e p. \tag{5.12}$$

The results (5.8), (5.9) and (5.12) were obtained earlier by Kimura and Ohta (1969) and Narain (1970) by solving different differential equations without invoking a conditional diffusion process.

In order to obtain the variance of the length of time until absorption in one of the boundaries, we proceed to obtain, using (5.8) and (5.12) and with the help of (4.4) and (4.10),

$$V_{c1}(p) = 32N_e^2\left[(\log_e p)\{\log_e(1-p)\} - \int_p^1 \frac{\log_e p}{1-p}\,dp + \left(\frac{1-p}{p}\right)\log_e(1-p)\right] \tag{5.13}$$

and

$$V_{c0}(p) = 32N_e^2\left\{\left(\frac{p}{1-p}\right)\log_e p - \int_0^p \frac{\log_e p}{p}\,dp\right\}. \tag{5.14}$$

The result (5.13) was earlier obtained by Narain (1970).

With the help of (5.13), (5.14), (5.8) and (5.12), the respective variances can be worked out.

ACKNOWLEDGEMENT

The author is grateful to the referee for helpful comments and for pointing out that the March 1973 issue of *Theoretical Population Biology* contains on pp. 21–30 a paper by W. J. Ewens in which the basic conditional diffusion equation is set out. The priority of this publication is duly acknowledged.

REFERENCES

EWENS, W. J. (1963). Numerical results and diffusion approximation in a genetic process. *Biometrika*, **50**, 241–249.

—— (1969). *Population Genetics*. London: Methuen.

FELLER, W. (1951). Diffusion processes in genetics. *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, pp. 227–246. University of California Press.
—— (1954). Diffusion processes in one dimension. *Trans. Amer. Math. Soc.*, **77**, 1–31.
KIMURA, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Statist.*, **28**, 882–901.
—— (1962). On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–719.
—— (1964). *Diffusion Models in Population Genetics*. London: Methuen. (Review Series in Applied Probability, Vol. 2, pp. 1–56.)
KIMURA, M. and CROW, J. F. (1963). The measurement of effective population number. *Evolution*, **17**, 279–288.
KIMURA, M. and OHTA, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–771.
KOLMOGOROV, A. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, **104**, 415–458.
NARAIN, P. (1969). Response to selection in finite populations. Ph.D. Thesis, Edinburgh University.
—— (1970). A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. *Genet. Res., Camb.*, **15**, 251–255.
WATTERSON, G. (1961). Markov chains with absorbing states; a genetic example. *Ann. Math. Statist.*, **32**, 716–729.
—— (1962). Some theoretical aspects of diffusion theory in population genetics. *Ann. Math. Statist.*, **33**, 939–957.
WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

10

Journal of Genetics Volume 63, Number 2 Dec. 1977

THE CONDITIONAL MARKOV CHAIN IN A GENETIC CONTEXT

PREM NARAIN
Institute of Agricultural Research Statistics, Library Avenue
New Delhi 110 012

Population genetics deals with a study of changes which the gene pool of a Mendelian population may undergo when it is exposed to systematic forces such as selection, mutation and migration. When the population is of limited size, the sample of genes transmitted to the next generation can deviate randomly from the true genetic composition of the parental generation and these random changes can accumulate over several generations. In other words, the change in the gene frequency over time due to systematic as well as random forces is a stochastic process. Usually the behaviour of the gene frequency in a generation depends only on its value in the immediately preceding generation so that the process is Markovian in structure. It can be studied either approximately as a diffusion process in which gene frequency as well as generations are treated as continuous or strictly as a finite Markov chain in which gene frequency is a discrete random variable and generations are discrete. In a series of investigations [Narain (1969), Narain and Robertson (1969), Robertson and Narain (1971), Narain (1971a)], it was shown how the process can be treated as a finite Markov chain and how the use of a transition probability matrix can be helpful in a genetic context. In particular, it was shown how to calculate the probability of fixation of a gene as well as the first two moments of the distribution of time taken for its fixation, disregarding the cases in which it is lost. The calculation of the first two moments of time until fixation of a particular allele was also attempted by the diffusion approach [Narain (1970), Narain (1974)]. The last investigation as well as that of Ewens (1973) have demonstrated that invoking a conditional process facilitates the calculation of the moments of the distribution of time taken for the fixation of a gene. This aspect is intimately connected with the concept of average number of generations required to attain limits of genetic improvement due to artificial selection which was first introduced in Narain (1969) and later elaborated in Narain (1971b). Although the diffusion approach to the conditional process is completely documented in Ewens (1973) and Narain (1974), the transition matrix approach is only briefly indicated in the former reference. The purpose of this paper is therefore to describe the conditional Markov chain and demonstrate its application in a genetic context relating to response to selection in finite populations. In addition, the theory has been applied to study the effect of linkage on the mean and variance of time until fixation of a gamete in populations practising self-fertilization.

CONDITIONAL MARKOV CHAIN

Assuming no mutation, consider a finite population of gametes of size 2N (corresponding to a population of diploid individuals of size N) and a single locus with two alleles A and a. Such a population can assume (2N+1) states $E_0, E_1, \ldots E_{2N-1}, E_{2N}$, the state E_i representing the state of i A genes and (2N-i) a genes. The frequency of A, denoted by x_i for the population in state E_i, can then take values $i/2N$, $i = 0, 1, \ldots (2N-1), 2N$. When $x_i = 0$ or 1 for $i = 0$ and 2N respectively, the population is said to be fixed for A or a respectively. But when $x_i \neq 0$ or 1, the population is said to be segregating for A and a alleles. Such a genetic situation corresponds to a finite absorbing Markov Chain with two absorbing states E_0 and E_{2N} and (2N-1) transient states $E_1, E_2, \ldots E_{2N-1}$. A detailed description of this chain, in such a context, is given in Narain (1971a). If P_{ij} represents one step transition probability for the system to move from E_i to E_j, the transition probability matrix $\underline{P}$ of order $(2N+1) \times (2N+1)$ takes the form

$$\underline{P} = \begin{bmatrix} 1 & \underline{0} & 0 \\ \underline{P}_0 & \underline{Q} & \underline{P}_{2N} \\ 0 & \underline{0} & 1 \end{bmatrix} \tag{1}$$

where $\underline{Q}$ is of order $(2N-1) \times (2N-1)$, giving the one-step transition probabilities amongst the transient states only, $\underline{P}_0$ and $\underline{P}_{2N}$ are column-vectors of order $(2N-1) \times 1$ representing the one-step transition probabilities from a transient state to E_0 and E_{2N} respectively. The vectors of the eventual probabilities of fixation of A and a, denoted by $\underline{U}$ and $\underline{L}$ are respectively given by

$$\underline{U} = (\underline{I} - \underline{Q})^{-1} \underline{P}_{2N} \tag{2}$$

$$\underline{L} = (\underline{I} - \underline{Q})^{-1} \underline{P}_0 \tag{3}$$

Consider now a finite absorbing Markov Chain conditional to the eventual absorption in E_{2N}. We then have only one absorbing state E_{2N} and (2N-1) transient states $E_1, \ldots E_{2N-1}$ from which absorption is only possible in E_{2N}. Let $P_{ij}^{(C1)}$ be the one-step transition probability for the system to move from E_i to E_j relative to the event of ultimate absorption in E_{2N}. Denoting by U_i, the i-th element of vector $\underline{U}$, the eventual probability of fixation of A when initially the population was in state E_i and following Kemeny and Snell (1960), we can define

$$P_{ij}^{(C1)} = P_{ij} U_j / U_i \tag{4}$$

with $U_{2N} = 1$. We then have the conditional one step transition probability matrix $\underline{P}^{(C1)}$, of order 2N x 2N given by

$$\underline{P}^{(C1)} = \begin{bmatrix} \underline{Q}^{(C1)} & \underline{P}_{2N}^{(C1)} \\ \underline{0} & 1 \end{bmatrix} \tag{5}$$

where $\underline{Q}^{(C1)}$ is of order $(2N-1)\times(2N-1)$, giving the one-step transition probabilities amongst the transient states only, conditional to fixation in E_{2N} and $\underline{P}_{2N}^{(C1)}$ is the column vector of order $(2N-1)\times 1$ representing the one-step transition probability from a transient state to E_{2N} relative to the eventual absorption in E_{2N}. The corresponding t-step transition matrix is given by

$$\underline{P}^{(C1)}(t) = \begin{bmatrix} \underline{Q}^{(C1)}(t) & \underline{P}_{2N}^{(C1)}(t) \\ \underline{0}' & 1 \end{bmatrix} \tag{6}$$

and the use of Chapman-Kolmogorov [Feller (1951)] for the Conditional Markov Chain gives

$$\underline{P}^{(C1)}(t) = \left[\underline{P}^{(C1)}\right]^t \tag{7}$$

so that

$$\underline{Q}^{(C1)}(t) = \left[\underline{Q}^{(C1)}\right]^t \tag{8}$$

$$P_{2N}^{(C1)}(t) = \left[I-(\underline{Q}^{(C1)}(t)\right]\left[I-\underline{Q}^{(C1)}\right]^{-1}\underline{P}_{2N}^{(C1)} \tag{9}$$

Following Narain (1971a), the column vector $\underline{U}^{(C1)}(t)$ of the probability of fixation of A by the t-th generation relative to the eventual fixation for A is obtained as

$$\underline{U}^{(C1)}(t) = \underline{P}_{2N}^{(C1)}(t) \tag{10}$$

As t tends to infinity, $(\underline{Q}^{(C1)})^t$ tends to zero so that the vector of the eventual probability of fixation for A, for the conditional process, is given by

$$\underline{U}^{(C1)} = \left[\underline{I}-\underline{Q}^{(C1)}\right]^{-1}\underline{P}_{2N}^{(C1)} \tag{11}$$

Writing $\underline{D}_1 = \mathrm{diag}(U_1, U_2, \ldots, U_{2N-1})$ of order $(2N-1)\times(2N-1)$, we find

$$\underline{Q}^{(C1)} = \underline{D}_1^{-1}\underline{Q}\underline{D}_1 \tag{12}$$

$$\left[\underline{Q}^{(C1)}\right]^t = \underline{D}_1^{-1}\underline{Q}^t\underline{D}_1 \tag{13}$$

$$\left[\underline{I}-\underline{Q}^{(C1)}\right]^{-1} = \underline{D}_1^{-1}(\underline{I}-\underline{Q})^{-1}\underline{D}_1 \tag{14}$$

$$\underline{P}_{2N}^{(C1)} = \underline{D}_1^{-1}\underline{P}_{2N} \tag{15}$$

It, therefore, follows that

$$\underline{U}^{(C1)} = \underline{D}_1^{-1}\cdot\underline{U} = \underline{e} \tag{16}$$

a column vector of unities, as expected due to the conditioning of the process. Further, we get

$$\underline{U}^{(C1)}(t) = \left[\underline{I} - \underline{D}_1^{-1}\, \underline{Q}^t\, \underline{D}_1\right] \underline{e} \qquad (17)$$

for working out the probabilities of fixation of A by the t-th generation.

Similarly, if we consider a finite absorbing Markov Chain conditional to the eventual absorption in E_0, we have again one absorbing state and (2N-1) transient states. Defining the corresponding one-step transition probability $P_{ij}^{(C0)} = P_{ij} L_j / L_i$ with $L_0 = 1$ and proceeding in the same way as above, we get

$$\underline{P}^{(C0)}(t) = \begin{bmatrix} 1 & \underline{0}' \\ \underline{P}^{(C0)}(t) & \underline{Q}^{(C0)}(t) \end{bmatrix} \qquad (19)$$

where, writing $\underline{D}_o = \mathrm{diag}(L_1, L_2, \ldots, L_{2N-1})$,

$$\underline{Q}^{(C0)}(t) = \underline{D}_o^{-1}\, \underline{Q}^t\, \underline{D}_o \qquad (20)$$

and

$$\underline{P}_o^{(C0)}(t) = \left[\underline{I} - \underline{D}_o^{-1}\, \underline{Q}^t\, \underline{D}_o\right] \underline{e}$$

$$= \underline{L}^{(C0)}(t) \,, \qquad (21)$$

giving the probabilities of fixation of a by the t-th generation in the conditional process with $\underline{L}^{(C0)} = \underline{e}$ as usual.

CONDITIONAL EXPECTED RESPONSE DUE TO ARTIFICIAL SELECTION

The random change in gene frequency due to finite population size has important applications in animal breeding as shown by Robertson (1960). The probability of fixation of the desirable allele can be converted into the expected response in the character under selection at the limit by making use of the relation between the selective advantage of a gene with its effect on the metric character under selection given first by Haldane (1931). Under the assumption of independent segregation of several loci affecting the character, the expected response at the limit, expressed in relation to the initial genetic standard deviation, is a function of Nih (the product of population size, intensity of selection and the square-root of heritability) and the initial frequency, p of the desirable allele, assumed equal at all loci. Narain (1971a) showed that this expected limit of response to selection, expressed in terms of the vector of changes in the frequency of desirable allele and denoted by $E(\underline{R})$ is given by

$$E(\underline{R}) = (\underline{I} - \underline{Q})^{-1}\, E(\delta \underline{p}) \qquad (22)$$

where $E(\delta \underline{p})$ is the vector of initial expected responses. Also, the the vector of

the expected response by the t-th generation, $E[\underline{R}(t)]$ was shown to be equal to

$$E[\underline{R}(t)] = (\underline{I} - \underline{Q}^t)\, E(\underline{R}) \tag{23}$$

Invoking a conditional process of selection along the same lines as in the previous section, we get the vector of conditional expected response due to selection by the t-th generation as

$$E\left[\underline{R}^{(C)}(t)\right] = \left[\underline{I} - (\underline{Q}^{(C1)})^t\right] E(\underline{R}^{(C)}) \tag{24}$$

relative to the eventual fixation of A regarded as a desirable allele. The expression for $E(\underline{R}^{(C)})$, however, becomes, as expected

$$\begin{aligned} E(\underline{R}^{(C)}) &= (\underline{I} - \underline{Q}^{(C1)})^{-1} E(\delta \underline{p}) \\ &= (\underline{I} - \underline{Q}^{(C1)})^{-1} \underline{P}_{2N}^{(C1)} - \underline{p}(0) \end{aligned} \tag{25}$$

where $\underline{p}(0)$ is the vector of the frequency of desirable allele in the initial population. We then have,

$$E\left[(\underline{R}^{(C)}(t)\right] = \left[\underline{I} - \underline{R}_1^{-1} \underline{Q}^t \underline{R}_1\right] \left[\underline{e} - \underline{p}(0)\right] \tag{26}$$

PROBABILITY GENERATING FUNCTION OF THE DISTRIBUTION OF TIME UNTIL FIXATION OF A PARTICULAR ALLELE

Let T_i be the time taken to first reach fixation of A, given that the initial population contains i A genes and (2N-i) a genes relative to the hypothesis of eventual absorption in E_{2N}. Let $S_i^{(t)}$ be the probability that $T_i = t$. Then clearly,

$$S_i^{(1)} = P_{i,2N}^{(C1)} \tag{27}$$

The probability generating function $\pi_i^{(C1)}(z)$, in this case, can then be expressed as

$$\begin{aligned} \pi_i^{(C1)}(z) &= z S_i^{(1)} + \sum_{t=2}^{\infty} z^t S_i^{(t)} \\ &= z P_{i,2N}^{(C1)} + z \sum_{t=2}^{\infty} z^{t-1} \sum_{k=1}^{2N-1} P_{ik}^{(C1)} S_k^{(t-1)} \\ &= z P_{i,2N}^{(C1)} + z \sum_{k=1}^{2N-1} P_{ik}^{(C1)} \pi_k^{(C1)}(z) \end{aligned} \tag{28}$$

In matrix notations, we can write it as

$$\underline{\pi}^{(C1)}(z) = z(\underline{I} - z\underline{Q}^{(C1)})^{-1}(\underline{I} - \underline{Q}^{(C1)})\,\underline{e} \tag{29}$$

where z is still a scalar and $\pi^{(C1)}(z)$ is the vector of probability generating functions conditional to fixation of A. Using the relationship between functions of $\underline{Q}^{(C1)}$ and $\underline{Q}$ given in Section 2, we get

$$\underline{\pi}^{(C1)}(z) = z\,\underline{D}_1^{-1}(\underline{I} - z\underline{Q})^{-1}(\underline{I} - \underline{Q})\,\underline{U} \tag{30}$$

The vector of the first moments of the distributions of time until fixation of A is obtained by differentiating $\underline{\pi}^{(C1)}(z)$ once and putting z = 1. This gives

$$\begin{aligned} E(\underline{T}_1) = \left|(d/dz)\,\underline{\pi}^{(C1)}(z)\right|_{z=1} &= \left|(d/dz)\underline{D}_1^{-1}(z^{-1}\underline{I}-\underline{Q})^{-1}(\underline{I}-\underline{Q})\underline{U}\right|_{z=1} \\ &= \left|\underline{D}_1^{-1}z^{-2}(z^{-1}\underline{I}-\underline{Q})^{-2}(\underline{I}-\underline{Q})\underline{U}\right|_{z=1} \\ &= \underline{D}_1^{-1}(\underline{I}-\underline{Q})^{-1}\underline{U} \end{aligned} \tag{31}$$

It is easy to see that this is also equivalent to $(\underline{I}-\underline{Q}^{(C1)})^{-1}\underline{e}$. The vector of the second factorial moment is given by

$$\begin{aligned} E(\underline{T}_1^2) - E(\underline{T}_1) &= \left|(d^2/dz^2)\,\underline{\pi}^{(C1)}(z)\right|_{z=1} \\ &= 2\underline{D}_1^{-1}[(\underline{I}-\underline{Q})^{-2} - (\underline{I}-\underline{Q})^{-1}]\,\underline{U} \end{aligned} \tag{32}$$

Using (31) and (32), the vector of the second moment about origin is obtained as

$$E(\underline{T}_1^2) = \underline{D}_1^{-1}[2(\underline{I}-\underline{Q})^{-2} - (\underline{I}-\underline{Q})^{-1}]\,\underline{U} \tag{33}$$

With the help of the elements of the vectors given by (31) and (33), one can obtain the variances of the time until fixation of A.

In a similar manner, we obtain the vector of the probability generating functions $\underline{\pi}^{(C0)}(z)$, of the distributions of time until fixation of a. This is given by

$$\begin{aligned} \underline{\pi}^{(C0)}(z) &= z(\underline{I} - z\underline{Q}^{(C0)})^{-1}(\underline{I} - \underline{Q}^{(C0)})\,\underline{e} \\ &= z\,\underline{D}_0^{-1}(\underline{I} - z\underline{Q})^{-1}(\underline{I} - \underline{Q})\underline{L} \end{aligned} \tag{34}$$

The vectors of the first and second moments in this case are given by

$$E(\underline{T}_0) = \underline{D}_0^{-1}(\underline{I} - \underline{Q})^{-1}\underline{L} \tag{35}$$

$$E(\underline{T}_0^2) = \underline{D}_0^{-1}[2(\underline{I} - \underline{Q})^{-2} - (\underline{I} - \underline{Q})^{-1}]\,\underline{L} \tag{36}$$

From which one can get the corresponding variances of the time until fixation of a.

EIGEN-ROOTS AND EIGEN-VECTORS OF THE CONDITIONAL MARKOV CHAIN WITH BINOMIAL TRANSITION PROBABILITIES

It is apparent from the above matrix derivations that for applying this theory, the element of $\underline{P}^{(C1)}$, $\underline{P}^{(C0)}$, $\underline{D}_1$ and $\underline{D}_0$ are required to be known. Analytically, this involves working out the eigen-roots and eigen-vectors of $\underline{P}^{(C1)}$ and $\underline{P}^{(C0)}$. Since the conditional transition matrices $\underline{P}^{(C1)}$ and $\underline{P}^{(C0)}$ depend on the conditional transition matrix $\underline{P}$ and since $\underline{D}_1$ and $\underline{D}_0$ are shown to be certain functions of $\underline{P}$ [Narain (1971a)], the problem boils down to a study of $\underline{P}$ or its derivative $\underline{Q}$. For specifying the elements of $\underline{P}$, we consider, as an example, the binomial transition probabilities. This case is commonly known as Wright's model [Wright (1931)]. It assumes absence of selective forces and considers only random drift based on binomial sampling with a constant population size N. The eigen-roots and vectors of $\underline{P}$ in such a case are also known [Feller (1951)]. Extension of such a model so as to involve selection in the context of limits of response to selection has been extensively studied by Narain and Robertson (1969). However, it is still of interest to study the eigen-roots and eigen-vectors of the conditional transition matrices $\underline{P}^{(C1)}$ and $\underline{P}^{(C0)}$. For this purpose we follow the approach given in Feller (1951).

With binomial sampling and no selection, we have

$$P_{ij} = \binom{2N}{j} p_i^j (1-p_i)^{2N-j} \qquad \begin{matrix} i = 0, 1, \ldots 2N \\ j = 0, 1, \ldots 2N \end{matrix} \qquad (37)$$

$$P_{ij}^{(C1)} = \binom{2N}{j} p_i^j (1-p_i)^{2N-j} (p_j/p_i) \quad , \qquad \begin{matrix} i = 1, 2, \ldots, 2N \\ j = 1, 2, \ldots, 2N \end{matrix} \qquad (38)$$

where $p_i = i/2N$.

The eigen-roots of $\underline{P}^{(C1)}$ are obtained by solving the characteristic equation

$$|\underline{P}^{(C1)} - \lambda \underline{I}| = 0 \qquad (39)$$

It is found that the roots are given by

$$\lambda_r^{(C)} = (1 - r/2N) \binom{2N}{r} r! / (2N)^r, \qquad r = 0, 1, \ldots (2N-1) \qquad (40)$$

For $r = 1, 2, \ldots (2N-1)$, the roots are the same as that of $\underline{Q}^{(C1)}$ and similarly as that of $\underline{Q}^{(C0)}$, which in view of (12), are the same as that of $\underline{Q}$ viz.

$$\lambda_r = \binom{2N}{r} r! / (2N)^r \quad , \qquad r = 1, \ldots 2N-1 \quad . \qquad (41)$$

Writing $j_{(v)} = j(j-1)\ldots(j-v+1)$, we get

$$\sum_{j=1}^{2N} P_{ij}^{(C1)} j_{(v)} = [(2N)_{(v+1)} p_i^v + v(2N)_{(v)} p_i^{v-1}] / 2N \qquad (42)$$

This shows that, taking $v = 1$, the expected value of the gene frequency, with pure random drift, will not be simply p_i but instead given by

$$\sum_{j=1}^{2N} p_{ij}^{(C1)} (j/2N) = p_i + (1-p_i)/2N \tag{43}$$

so that

$$E(\delta p_i) = (1-p_i)/2N \tag{44}$$

Similarly, with $v = 2$, we get

$$\sum_{j=1}^{2N} p_{ij}^{(C1)} (j/2N)^2 - (1/2N) \sum_{j=1}^{2N} p_{ij}^{(C1)} (j/2N) = (1-1/2N)(1-2/2N)p_i^2 + (2/2N)(1-1/2N)p_i \tag{45}$$

This shows that, using (43), the variance of the change in the gene frequency, with pure random drift, will not be simply $p_i(1-p_i)/2N$ but, instead, given by

$$V(\delta p_i) = (1/2N)(1-1/2N)p_i(1-p_i) \tag{46}$$

From (44) and (46) it is evident that for a population so large that $(1/2N)^2$ is negligible, the mean and variance of the change in gene frequency due to random drift, in the conditional process, are $(1-p_i)/2N$ and $p_i(1-p_i)/2N$ as against 0 and $p_i(1-p_i)/2N$ respectively in the unconditional case. This is exactly what we get from the diffusion approach for the pure random drift case as shown in Narain (1974). It is however interesting to note that for the exact process, conditioning the process increases the mean but decreases the variance.

Corresponding to each characteristic root $\lambda_r^{(C)}$ given by (40), the system of linear equations

$$\sum_{j=1}^{2N} p_{ij}^{(C1)} x_{jr} = \lambda_r^{(C)} x_{ir}, \qquad i = 1, 2, \ldots, 2N \tag{47}$$

admits a non-trivial solution $\underline{x}_r = (x_{1r} \ldots x_{(2N)r})$ known as the right-hand eigenvector for $\lambda_r^{(C)}$. It is, therefore, always possible to find constants $a_0, a_1, \ldots, a_r$ (not all of them zero) such that

$$x_{jr} = \sum_{v=0}^{r} a_v j^{(v)} \tag{48}$$

is a solution of (47). Substituting (48) in (47) and using (42), we get

$$\sum_{v=0}^{r} a_v[(2N-v)p_i+v]\frac{(2N)_v p_i^{v-1}}{2N} = \lambda_r^{(C)} \sum_{v=0}^{r} a_v i_{(v)} \qquad i = 1, 2, \ldots, 2N \qquad (49)$$

Since coefficients of a_v on the both sides of (49) are polynomials of degree v in i, it is possible to write

$$p_i^v = \sum_{s=0}^{v} C_{s,v}\, i_{(s)} \qquad (50)$$

where $C_{s,v}$'s are independent of i. Substituting (50) in (49) and equating the coefficients of $i_{(t)}$ for $t = 0, 1, \ldots, r$, we get,

$$\lambda_r^{(C)} a_t = a_t (2N)_{(t+1)} C_{tt}/(2N) + \sum_{v=t+1}^{r} a_v[(2N-v)C_{tv}+vC_{t,(v-1)}](2N)_v/(2N), \qquad (t = 0, 1, \ldots, r) \qquad (51)$$

If we take $t = r$ in (51) and use (40) as well as (50), it is found that (51) is satisfied for $v = r$ and arbitrary a_r. We can then put $a_r = 1$ in (51) for $t = (r-1)$ giving a_{r-1}. This procedure allows us to calculate $a_{r-2}, \ldots, a_1$ and a_0 in succession, giving thereby the j-th element of the right-hand eigen-vector corresponding to $\lambda_r^{(C)}$ given by (48).

Using the above procedure of obtaining the eigen-vectors for $r = 1, 2$ and 3, we find that for the three eigen-roots,

$$\begin{aligned} \lambda_1^{(C)} &= (1-1/2N) \\ \lambda_2^{(C)} &= (1-1/2N)(1-2/2N) \\ \lambda_3^{(C)} &= (1-1/2N)(1-2/2N)(1-3/2N) \end{aligned} \qquad (52)$$

the vectors are, respectively, given by

$$\begin{aligned} x_{j1} &= (1-p_j) \\ x_{j2} &= (1-p_j)(1-2p_j) \\ x_{j3} &= (1-p_j)[(2N-1)/(10N-6)-p_j(1-p_j)] \end{aligned} \qquad (53)$$

with $j = 1, 2, \ldots, 2N$. These results can be compared with the corresponding results of the unconditional case detailed in Narain and Robertson (1969). Although the roots are the same, the elements of the vectors are now $(1/p_j)$ of those in the unconditional case. Alternatively, since $\lambda_r^{(C)}$ is an eigenvalue of $\underline{Q}^{(C)}$ with the associated right eigen-vector, $\underline{x}_r$, we have

$$\underline{Q}^{(C1)}\underline{x}_r = \lambda_r^{(C)}\underline{x}_r \ , \qquad r = 1, 2, \ldots, 2N-1$$

In view of (12), we get

$$\underline{Q}\,\underline{D}_1\,\underline{x}_r = \lambda_r^{(C)}\underline{D}_1\,\underline{x}_r \ , \qquad r = 1, 2, \ldots, 2N-1 \ .$$

This shows that $\lambda_r^{(C)}$ is also an eigenvalue of $\underline{Q}$ with the associated right-eigenvector $\underline{z}_r = \underline{D}_1\,\underline{x}_r$, so that by definition of $\underline{D}_1$, the j-th element of $\underline{x}_r$ is 2N/j times as large as the j-th element of $\underline{z}_r$ i.e. those corresponding to the unconditional case.

EFFECT OF LINKAGE ON THE MEAN AND VARIANCE OF TIME UNTIL FIXATION OF A GAMETE IN SELFED POPULATIONS

The effect of linkage on the probability of fixation of a gamete in populations practising self-fertilization was studied in Narain (1971a) which can be consulted for details. The case of self-fertilization corresponds to the situation when N=1. The population is sub-divided into lines from each of which two gametes are chosen to form one mature individual only. With two linked loci each with two alleles A-a and B-b respectively with recombination probability r, with s = 1-r, and assuming no mutation, there are 10 states of the system corresponding to 10 types of lines out of which four homozygous ones represent absorbing states and the remaining six are trasient states. Amongst the transient states, the two corresponding to two double heterozygotes, AB/ab (coupling) and Ab/aB (repulsion) are important from the point of view of linkage, the remaining four involving single heterozygotes only. Taking the P-matrix of the process and the $\underline{U}_{AB}$ vector of the probabilities of fixation of gamete AB from Narain (1971a), the $\underline{P}^{(C11)}$ matrix for the process, conditional to absorption in AB/AB, has the form

$$\underline{P}^{(C11)} = \begin{bmatrix} \underline{Q}^{(C11)} & \underline{P}_{AB}^{(C)} \\ \underline{0}' & 1 \end{bmatrix} \tag{54}$$

where

$$\underline{P}_{AB}^{(C)'} = (1/2, 1/2, 0, 0, s^2(1+2r)/2, r(1+2r)/4) \tag{55}$$

$$\underline{Q}^{(C11)} = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ rs(1+2r)/2 & rs(1+2r)/2 & 0 & 0 & s^2/2 & r^3 \\ s(1+2r)/4 & s(1+2r)/4 & 0 & 0 & r/4 & s^2/2 \end{bmatrix} \tag{56}$$

The ordering of the states being AB/Ab, AB/aB, Ab/ab, aB/ab, AB/ab and Ab/aB. With the help of the results given in Narain (1971a) and using (31) & (33), the vectors of mean as well as second moment about origin of time until fixation of AB

are respectively given by

$$E(\underline{T}'_{AB}) = [2, 2, 0, 0, \alpha^{(c)}_{AB}, \alpha^{(r)}_{AB}] \tag{57}$$

$$E(\underline{T}^{2}_{AB}) = [6, 6, 0, 0, \beta^{(c)}_{AB}, \beta^{(r)}_{AB}] \tag{58}$$

where $\alpha^{(c)}_{AB}$, $\beta^{(c)}_{AB}$ corresponding to the situation where the population is initially in the coupling phase, are given by

$$\alpha^{(c)}_{AB} = (1+2r)(1+4rs)/(1+2rs) + (1-2r)/(1+2r) \tag{59}$$

$$\beta^{(c)}_{AB} = (1+2r)(3+26rs+24r^2s^2)/(1+2rs)^2 + (1-2r)(3-2r)/(1+2r)^2 \tag{60}$$

and $\alpha^{(r)}_{AB}$, $\beta^{(r)}_{AB}$ corresponding to the situation when the population is initially in the repulsion phase, are given by

$$\alpha^{(r)}_{AB} = (1+2r)(1+4rs)/2r(1+2rs) - (1-2r)/2r(1+2r) \tag{61}$$

$$\beta^{(r)}_{AB} = (1+2r)(3+26rs+24r^2s^2)/2r(1+2rs)^2 - (1-2r)(3-2r)/2r(1+2r)^2 \tag{62}$$

In a similar manner, we get the corresponding vectors for moments of time until fixation of Ab, aB and ab. It is found that when the population is initially in the coupling phase, the means and second moments about origin of time until fixation of Ab as well as aB are the same as that given by (61) and (62) respectively whereas when the initial population is in repulsion phase, these are correspondingly given by (59) and (60). These results for the time until fixation of ab are exactly the same as that until fixation of AB given by (59) to (62). In each case, the variance of time until fixation is calculated by subtracting the square of α from β.

It is interesting to note that mean and variance of time until homozygosity can further be obtained by multiplying the mean and variance of time until fixation of a gamete by the corresponding probability of fixation and adding over the four possible cases. For the situation when the initial population is in the coupling phase, these are given by

$$E(T) = U^{(c)}_{AB}\,\alpha^{(c)}_{AB} + U^{(c)}_{Ab}\,\alpha^{(c)}_{Ab} + U^{(c)}_{aB}\,\alpha^{(c)}_{aB} + U^{(c)}_{ab}\,\alpha^{(c)}_{ab} = 2(1+4rs)/(1+2rs) \tag{63}$$

$$\begin{aligned} Var(T) &= U^{(c)}_{AB}\,\beta^{(c)}_{AB} + U^{(c)}_{Ab}\,\beta^{(c)}_{Ab} + U^{(c)}_{aB}\,\beta^{(c)}_{aB} + U^{(c)}_{ab}\,\beta^{(c)}_{ab} - [E(T)]^2 \\ &= 2(1+10rs-8r^2s^2)/(1+2rs)^2 \end{aligned} \tag{64}$$

where $U^{(c)}_{AB} = U^{(c)}_{ab} = 1/2(1+2r)$ and $U^{(c)}_{Ab} = U^{(c)}_{aB} = r/(1+2r)$ [Narain (1971a)]. Because of symmetry, (63) and (64) hold for the repulsion phase also. The values of E(T) and Var(T) obtained here are exactly the same as those obtained by Puri (1968) who obtained them directly without working out the time until fixation of a particular gamete.

The effect of the recombination fraction r on the mean and the standard deviation of time until fixation was numerically studied with the help of expressions (59) to (62). The results for the case when the initial population is in coupling phase are presented in Table 1. For the case when the initial population is in repulsion phase, the results are obtainable from the Table by interchanging either A and a or B and b.

Table 1: Mean and standard deviation of time (number of generations) until fixation of a gamete for the initial population with heterozygotes in coupling phase.

r	AB or ab		Ab or aB	
	Mean	s.d.	Mean	s.d.
0.0000	2.0000	1.4142	4.0000	2.0000
0.0625	2.0208	1.4337	3.7244	1.8622
0.1250	2.0743	1.4753	3.4973	1.7682
0.1875	2.1506	1.5213	3.3103	1.7078
0.2500	2.2424	1.5635	3.1516	1.6709
0.3125	2.3441	1.5959	3.0120	1.6486
0.3750	2.4513	1.6187	2.8873	1.6382
0.4375	2.5602	1.6306	2.7734	1.6354
0.5000	2.6666	1.6329	2.6666	1.6329

It is found that when the initial population is in coupling phase, the effect of linkage is to decrease the average and standard deviation of the number of generations until fixation for a coupled gamete (AB or ab) but to increase the same for a repulsed gamete (Ab or aB). It may be noted that when we consider independently segregating loci ($r = 0.50$) and fixation of coupled gametes with initial population in coupling phase (or of repulsed gametes with initial population in repulsion phase), the average time to fixation is about 1.33 times that for completely linked loci whereas the chance of fixation is half of its value for the completely linked case. As expected, with independent segregation, the average time to fixation of corresponding repulsed (or coupled) gametes, the chance of fixation being the same viz. 0.25 in all the four cases. But for very tight linkage (r approaching zero) average time to fixation of a repulsed gamete with initial population in coupling phase (or of a coupled gamete with initial population in repulsion phase) tends to a limiting value of 4 with chance of its fixation becoming very very small. As regards variability in time to fixation, a characteristic feature, true for all situations, is that a larger mean is accompanied by a larger standard deviation.

SUMMARY

A theory of the stochastic change in the frequency of a gene in finite populations conditional to its eventual fixation has been developed employing a Condi-

ditional Markov Chain. The probability generating function of the distribution of time until fixation of a particular allele as well as the eigen-roots and eigen-vectors of the conditional process with binomial transition probabilities have been studied. The theory has been applied to investigate the effect of linkage on the mean and standard deviation of time until fixation of a gamete in populations practising self-fertilization. It has been found that linkage decreases or increases the average and standard deviation of time to fixation of a coupled gamete according as the initial population consists of a coupling or repulsion heterozygote respectively.

ACKNOWLEDGEMENTS

The author is grateful to Professor E. Pollak, Iowa State University, U.S.A. as well as the referee for helpful suggestions which have considerably improved the presentation.

REFERENCES

EWENS, W.G., 1973. Conditional diffusion processes in population genetics. Theor. Pop. Bio. 4, 21-30.

FELLER, W., 1951. Diffusion processes in genetics. Proc. 2nd Berkeley Symp. Math. Statist. and Prob., pp. 227-246, Univ. of California Press, Berkeley, C.A.

HALDANE, J.B.S., 1931. A mathematical theory of natural and artificial selection. VII. Selection intensity as a function of mortality rate. Proc. Camb. Phil. Soc. 27, 131-136.

KEMENY, J.G. and SNELL, J.L., 1960. Finite Markov Chains. Princeton, New York.

NARAIN, P., 1969. Response to selection in finite populations. Ph.D. thesis, Edinburgh University.

NARAIN, P., 1970. A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. Genet. Res., Camb. 15, 251-255.

NARAIN, P., 1971a. The use of transition probability matrices in studies on limits of response to selection. J. Indian Soc. Agri. Statist. 23, 48-60.

NARAIN, P., 1971b. Average number of generations required to attain limits of genetic improvement. SABRAO Newsletter. 3, 135-142.

NARAIN, P., 1974. The conditioned diffusion equation and its uses in population genetics. J. Royal Statist. Soc. London B 36, 258-266.

NARAIN, P. and ROBERTSON, A., 1969. Limits and duration of response to selection in finite populations: the use of transition probability matrices. Indian J. Heredity, 1, 79-97.

PURI, P.S., 1968. A study through a Markov Chain of a population undergoing certain mating systems in the presence of linkage. Contributions in Statist. and Agri. Sciences. Vol. presented to Dr. V.G. Panse at his 62nd birthday: 127-141.

ROBERTSON, A., 1960. A theory of limits in artificial selection. Proc. Roy. Soc. London B 153, 234-249.

ROBERTSON, A. and NARAIN, P., 1971. Survival of recessive lethals in finite populations. Theor. Pop. Bio. 2, 24-50.

WRIGHT, S., 1931. Evolution in Mendelian populations. Genetics, 16, 97-159.

Genet. Res., Camb. (1977), 29, pp. 113–121
Printed in Great Britain

11

On the fixation probability of a gene under random fluctuations in selection intensities in small populations*

BY PREM NARAIN† AND EDWARD POLLAK
Statistical Laboratory, Iowa State University, Ames, Iowa 50011, U.S.A.

DEDICATED TO THE MEMORY OF THEODOSIUS DOBZHANSKY

(*Received* 22 *June* 1976)

SUMMARY

A population with N monoecious individuals, and having two alleles, is considered. The problem of calculating the fixation probability of a particular allele under random fluctuation of selection intensities is re-examined, employing finite Markov chain methods. An approximate but general expression for this probability is obtained and the results obtained by previous workers are shown to be special cases of this result.

1. INTRODUCTION

A problem in population genetics, which has received considerable attention in recent times, is that of the computation of the fixation probability of a gene and, along with it, the average time until its fixation. This is usually dealt with in two different ways. On the one hand, there is the diffusion approximation method, wherein the gene frequency is treated as a continuous random variable, between 0 and 1, and the time parameter of the underlying Markovian process is also taken as continuously varying. This method involves the use of the diffusion equations of Kolmogorov (1931), a mode of attack initiated for genetical problems by Fisher (1922), and the conditioned diffusion equations dealt with by Ewens (1973) and Narain (1974). On the other hand, a more exact treatment is that by a finite Markov chain involving the use of transition matrices. This method is discussed extensively in Kemeny & Snell (1960). Whichever of the two approaches is adopted, there is a basic question of whether the selection coefficients attached to the genes are constant over time or are fluctuating randomly. In regard to the latter, some results have recently been published by Jensen & Pollak (1969), Ohta (1972), Gillespie (1973) and Jensen (1973) as well as by Karlin & Levikson (1974). By and large, these results are based on the method of diffusion approximations and iteration procedures on the computer using transition matrices. According to Ohta (1972), if the ratio of the mean and variance of the selection coefficient is small, a mutant, even if

* Journal Paper No. J-8525 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project 1669.

† On leave from the Institute of Agricultural Research Statistics, Library Avenue, New Delhi, India.

selected against, becomes fixed in the population like a selectively neutral mutant. Contrary to this result, Jensen (1973) showed that the variability in the selection coefficient increases the chance of fixation of a rare gene. The problem was, however, attacked in a much more comprehensive manner by Karlin & Levikson (1974). In particular, they formulated a haploid model allowing for variability in the selection coefficients of both the types as well as for correlation between the two and showed that the variance in selection expression reduces and mitigates the mean effects of selection differentials, so that the fixation probability of the abundant allele is diminished. It seems that these different results on the fixation probability are probably due to the difference in the forms of the mean as well as variance functions for the change in gene frequency used in the diffusion approximation approach. However, if one does exact computations on the finite Markov chain, it seems that the choice of the appropriate mean and variance functions could be easily, as well as accurately, resolved and that an algebraic expression for the fixation probability could be obtained. According to Karlin & Levikson (1974), the work of Ohta (1972) suffers from an incorrect mean function. In this paper, it is shown that not only the mean but the variance function also needs correction, particularly for extremely small populations. An approximate but general expression for the fixation probability of a gene, in the haploid case, allowing for the variability in the selection coefficients of both the types as well as for the correlation between the two, is also derived, and the results obtained by the previous workers are shown to be special cases of this result.

2. THE MODEL

Consider a haploid population of $2N$ genes, corresponding to a monoecious population of individuals, of constant size N, reproducing in discrete generations. Let there be two alleles A and a with fitness coefficients in generation n as follows:

$$\begin{array}{cc} A & a \\ 1+s_1 & 1+s_2. \end{array} \tag{2.1}$$

The selection intensities, s_1 and s_2, are assumed to fluctuate over time in a random manner with identical distribution functions in all generations and independence between generations. The means, variances and covariances of these variables are

$$E(s_1) = \bar{s}_1, \quad E(s_2) = \bar{s}_2, \tag{2.2}$$

$$\text{var}\,(s_1) = v_1, \quad \text{var}\,(s_2) = v_2, \tag{2.3}$$

$$\text{cov}\,(s_1, s_2) = r, \quad \text{where} \quad |r| \leqslant \sqrt{(v_1 v_2)}. \tag{2.4}$$

We are virtually considering the haploid model of Karlin & Levikson (1974) but with the difference that means of the selection effects and their variances and covariances are not taken to be of the order of magnitude of $1/2N$.

The frequency of the A-gene in generation n, given its frequency before selection was $p_i = i/2N$ and given the selection parameters (s_1, s_2) is

$$p_i^{(n)} = p_i + p_i(1-p_i)(s_1 - s_2)/[1 + s_2 + (s_1 - s_2)p_i]. \tag{2.5}$$

According to the standard Wright–Fisher Markov chain process, the distribution of the proportion of A-genes in generation n follows the binomial distribution with parameters $(2N, p_i^{(n)})$. In other words, the transition probability P_{ij} representing the conditional probability that there are j A-genes out of $2N$ genes, given that there were i A-genes in the population in the previous generation, is given by

$$P_{ij} = \binom{2N}{j} [p_i^{(n)}]^j [1-p_i^{(n)}]^{2N-j}, \tag{2.6}$$

where $i, j = 0, 1, 2, \ldots, 2N$.

The finite Markov chain, thus generated, could be studied by using results of Kemeny & Snell (1960). However, Narain & Robertson (1969) gave an analytical treatment of such processes in a genetic context and we follow the same procedure here.

3. EXPANSION OF THE TRANSITION PROBABILITY

Treating P_{ij} as a function of (s_1, s_2), we expand it, by Taylor's expansion, as a series in two variables. Neglecting terms involving powers and products of s_1 and s_2 greater than 2, we get

$$P_{ij} = P_{ij}(0)\,[1 + s_1 a_1 + s_2 a_2 + s_1^2 b_{11} + s_1 s_2 b_{12} + s_2^2 b_{22}], \tag{3.1}$$

where

$$P_{ij}(0) = \binom{2N}{j} p_i^j (1-p_i)^{2N-j}, \tag{3.2}$$

$$a_1 = 2N(p_j - p_i), \tag{3.3}$$

$$a_2 = -2N(p_j - p_i), \tag{3.4}$$

$$b_{11} = N[(2N-1)(p_j-p_i)^2 - p_j(1-p_j) - 2p_i(p_j-p_i)], \tag{3.5}$$

$$b_{12} = 2N[-(2N-1)(p_j-p_i)^2 + p_j(1-p_j) - (1-2p_i)(p_j-p_i)], \tag{3.6}$$

$$b_{22} = N[(2N-1)(p_j-p_i)^2 - p_j(1-p_j) + 2(1-p_i)(p_j-p_i)]. \tag{3.7}$$

If we now compute the expectation of (3.1) with respect to the distribution of s_1 and s_2, we have, in view of (2.2), (2.3) and (2.4),

$$\begin{aligned} E(P_{ij}) = P_{ij}(0)\,[1 &+ 2N\{(\bar{s}_1 - \bar{s}_2)(1-\bar{s}_2) + v_2 - r\}(p_j - p_i) \\ &+ N\{(\bar{s}_1 - \bar{s}_2)^2 + v_1 + v_2 - 2r\} \\ &\times \{(2N-1)(p_j-p_i)^2 - p_j(1-p_j) - 2p_i(p_j-p_i)\}], \end{aligned} \tag{3.8}$$

where the expectation is taken over the distribution of s_1 and s_2. Because the selection intensities are independent between generations and have the same distribution in each generation, the expression (3.8) holds as a one-step transition probability for any n. In other words, a Markov chain characterized by (3.8) is homogeneous in time.

4. MEAN AND VARIANCE FUNCTION FOR THE FREQUENCY OF A

In order to obtain the various moments of the frequency of the A-gene in the next generation, we proceed to obtain expressions for the expected values of expressions $j(j-1)\dots(j-k+1)$, given $p_i = i/2N$. Such expectations will be denoted by $E_i(j(j-1)\dots(j-k+1))$.

$$E_i[j(j-1)\dots(j-k+1)]$$
$$= \binom{2N}{j} k!\, p_i^k[1+k(1-p_i)\{1-\tfrac{1}{2}(\bar{s}_1+\bar{s}_2)+\tfrac{1}{2}(k(1-p_i)-p_i)(\bar{s}_1-\bar{s}_2)\}(\bar{s}_1-\bar{s}_2)$$
$$+\tfrac{1}{2}k(1-p_i)(k(1-p_i)-p_i)(v_1+v_2-2r)-\tfrac{1}{2}k(1-p_i)(v_1-v_2)]. \quad (4.1)$$

The gene frequency expected in the next generation on the basis of binomial sampling is therefore obtained by putting $k = 1$ in (4.2). Denoting it by p_i^*, we get

$$p_i^* = p_i + E_i(\Delta p) = p_i + p_i(1-p_i)$$
$$\times\left[(\bar{s}_1-\bar{s}_2)[1-\{\bar{s}_2+(\bar{s}_1-\bar{s}_2)p_i\}]+\frac{v_1+v_2-2r}{2}(1-2p_i)-\frac{v_1-v_2}{2}\right], \quad (4.2)$$

where $E_i(\Delta p)$ is the expected mean change in the gene frequency in one generation, given that the frequency is $p_i = i/2N$.

Putting $k = 2$ in (4.1) and using (4.2), we get the variance of the change in gene frequency per generation, given that $p_i = i/(2N)$, $V_i(\Delta p_i)$, to the same degree of approximation, as

$$V_i[\Delta p] = \frac{p_i(1-p_i)}{2N}[1+(1-\bar{s}_2)(\bar{s}_1-\bar{s}_2)(1-2p_i)-(\bar{s}_1-\bar{s}_2)^2 p_i(2-3p_i)$$
$$-(v_1+v_2-2r)p_i(2-3p_i)+(v_2-r)(1-2p_i)]+(v_1+v_2-2r)p_i^2(1-p_i)^2. \quad (4.3)$$

Expressions (4.2) and (4.3) can also be derived directly by manipulating expectations. An alternative expression for $V_i[\Delta p]$, in terms of p_i given by (4.2), can be written as

$$V_i[\Delta p] = \frac{p_i^*(1-p_i^*)}{2N}+\left(1-\frac{1}{2N}\right)(v_1+v_2-2r)p_i^2(1-p_i)^2. \quad (4.4)$$

It is evident from the expressions for $E_i(\Delta p)$ and $V_i[\Delta p]$ derived above that random fluctuations in selection intensities affects both of them. These expressions can be compared with (3.7) and (3.8) on page 392 in the paper by Karlin & Levikson (1974) which may be written in the notation of this paper as

$$E_i(\Delta p) = p_i(1-p_i)[\bar{s}_1-\bar{s}_2-\tfrac{1}{2}(v_1-v_2)+\tfrac{1}{2}(v_1+v_2-2r)(1-2p_i)]$$

and

$$V_i(\Delta p) = \frac{p_i(1-p_i)}{2N}+(v_1+v_2-2r)p_i^2(1-p_i)^2$$

respectively. It is clear from (4.2) that our expression for $E_i(\Delta p)$ has an extra term: $-(\bar{s}_1-\bar{s}_2)(\bar{s}_2+(\bar{s}_1-\bar{s}_2)p_i)\,p_i(1-p_i)$. In expression (4.4) there are two types of terms associated with the non-additivity of the parts of $V_i(\Delta p)$ that arise if we consider

separately random changes due to (i) random sampling of gametes and (ii) random fluctuations in selection. In all the studies made so far, i.e. by Kimura (1962), Ohta (1972), Jensen (1973), as well as Karlin & Levikson (1974), this non-additivity is ignored.

Apart from this, the expression for $V_i(\Delta p)$ needs to be reconsidered, even when there is only random sampling of gametes and no random fluctuations in selection. With non-random selection ($v_1 = v_2 = r = 0$) the expressions for $E_i(\Delta p)$ and $V_i(\Delta p)$ become

$$E_i[\Delta p] = (\bar{s}_1 - \bar{s}_2)(1 - \bar{s}_2 - (\bar{s}_1 - \bar{s}_2)p_i)p_i(1-p_i), \tag{4.5}$$

$$V_i[\Delta p] = \frac{p_i(1-p_i)}{2N} + (1 - 2p_i - E_i[\Delta p])\frac{E_i[\Delta p]}{2N}. \tag{4.6}$$

It may be observed that the usual variance due to binomial sampling is strictly true only when $E_i[\Delta p] = 0$, i.e. the selectively neutral case. Even when the changes in gene frequency per generation due to selection are very small so that squares of $E_i[\Delta p]$ can be neglected, the binomial sampling variance holds only when $p_i = \frac{1}{2}$.

5. FIXATION PROBABILITY OF THE *A*-GENE

Let $u(p_i)$ be the fixation probability of the A-gene, given that initially at $t = 0$, it had frequency p_i, so that the total expected change in the frequency at the limit is

$$L_i = u(p_i) - p_i. \tag{5.1}$$

Narain & Robertson (1969) showed that the vector $\mathbf{L} = (L_1, L_2, \ldots, L_{2N-1})'$ is obtained by operating the matrix $\mathbf{T} = (\mathbf{I} - \mathbf{Q})^{-1}$ on to the vector

$$\mathbf{E}(\Delta p) = [E_1(\Delta p), E_2(\Delta p), \ldots, E_{2N-1}(\Delta p)]',$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{Q}$ is the matrix of transition probabilities when we consider transitions between the transient states only; i.e. $i, j = 1, 2, \ldots, 2N-1$. Expressed in powers and products of s_1 and s_2 up to terms involving $\{s_1\}^2$, $\{s_2\}^2$ and $\{s_1 s_2\}$, $\mathbf{T}$ is given by

$$\begin{aligned}\mathbf{T} = \mathbf{T}_0 &+ s_1 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 + s_2 \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0 + s_1^2[\mathbf{T}_0 \mathbf{B}_{11} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0] \\ &+ s_1 s_2[\mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 + \mathbf{T}_0 \mathbf{B}_{12} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0] \\ &+ s_2^2[\mathbf{T}_0 \mathbf{B}_{22} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0],\end{aligned}$$

where $\mathbf{T}_0 = (\mathbf{I} - \mathbf{Q})^{-1}$ and $\mathbf{Q}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_{11}$, $\mathbf{B}_{12}$ and $\mathbf{B}_{22}$ are $(2N-1) \times (2N-1)$ matrices with $i-j$th elements $P_{ij}(0)$, $a_1 P_{ij}(0)$, $a_2 P_{ij}(0)$, $b_{11} P_{ij}(0)$, $b_{12} P_{ij}(0)$ and $b_{22} P_{ij}(0)$ respectively. Allowing for random variations in s_1 and s_2, as before, we get

$$\begin{aligned}E_{s_1, s_2}(\mathbf{T}) = \mathbf{T}_0 &+ (\bar{s}_1 - \bar{s}_2)(\mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0) + (v_1 + \bar{s}_1^2)[\mathbf{T}_0 \mathbf{B}_{11} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0] \\ &+ (r + \bar{s}_1 \bar{s}_2)[\mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 + \mathbf{T}_0 \mathbf{B}_{12} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0] \\ &+ (v_2 + \bar{s}_2^2)[\mathbf{T}_0 \mathbf{B}_{22} \mathbf{T}_0 + \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0 \mathbf{A}_2 \mathbf{T}_0].\end{aligned} \tag{5.2}$$

The expression $E(\Delta p)$ in powers and products of s_1 and s_2 up to terms involving

$\{s_1\}^2$, $\{s_2\}^2$ and $\{s_1 s_2\}$ and allowing for random variations in s_1 and s_2, in the same manner as before, is obtained by converting (4.2) into vector notation as

$$E[\Delta p] = \left(1 - \frac{\bar{s}_1 + \bar{s}_2}{2}\right)(\bar{s}_1 - \bar{s}_2)\,\mathbf{x}_1 + \tfrac{1}{2}(\bar{s}_1 - \bar{s}_2)^2\,\mathbf{x}_2 + \tfrac{1}{2}(v_1 + v_2 - 2r)\,\mathbf{x}_2 - \tfrac{1}{2}(v_1 - v_2)\,\mathbf{x}_1, \tag{5.3}$$

where vectors

$$\mathbf{x}_1 = [p_1(1-p_1), p_2(1-p_2), \ldots, p_{2N-1}(1-p_{2N-1})]', \tag{5.4}$$

$$\mathbf{x}_2 = [p_1(1-p_1)(1-2p_1), p_2(1-p_2)(1-2p_2), \ldots, p_{2N-1}(1-p_{2N-1})(1-2p_{2N-1})]', \tag{5.5}$$

correspond to the two eigenvalues of $\mathbf{Q}_0$ given by

$$\lambda_1 = (1 - 1/2N), \tag{5.6}$$

$$\lambda_2 = (1 - 1/2N)(1 - 2/2N). \tag{5.7}$$

Operating (5.2) on to (5.3) then gives, to the same degree of approximation,

$$\mathbf{L} = (\bar{s}_1 - \bar{s}_2)(\mathbf{T}_0 \mathbf{x}_1) + (\bar{s}_1 - \bar{s}_2)^2 \mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{x}_1) - \tfrac{1}{2}(v_1 + \bar{s}_1^2)\{\mathbf{T}_0(\mathbf{x}_1 - \mathbf{x}_2)\} - (r + \bar{s}_1 \bar{s}_2)(\mathbf{T}_0 \mathbf{x}_2) + \tfrac{1}{2}(v_2 + \bar{s}_2^2)\{\mathbf{T}_0(\mathbf{x}_1 + \mathbf{x}_2)\}. \tag{5.8}$$

Using Table 1 of Narain & Robertson (1969), while noting that $\mathbf{A}_1$ is twice as large as $2\mathbf{Q}_0'$, we obtain

$$\mathbf{T}_0 \mathbf{x}_1 = 2N\mathbf{x}_1, \tag{5.9}$$

$$\mathbf{T}_0 \mathbf{A}_1 \mathbf{T}_0 \mathbf{x}_1 = \left(\frac{4N^3}{3N-1}\right)\lambda_1 \mathbf{x}_2, \tag{5.10}$$

$$\mathbf{T}_0 \mathbf{x}_2 = \left(\frac{2N^2}{3N-1}\right)\mathbf{x}_2. \tag{5.11}$$

With use of (5.9) to (5.11), the expression (5.8) becomes

$$\mathbf{L} = 2N(\bar{s}_1 - \bar{s}_2)\left(1 - \frac{\bar{s}_1 + \bar{s}_2}{2}\right)\mathbf{x}_1 + \frac{N^2(4N-1)}{3N-1}(\bar{s}_1 - \bar{s}_2)^2 \mathbf{x}_2 + \frac{N^2}{3N-1}(v_1 + v_2 - 2r)\,\mathbf{x}_2 - N(v_1 - v_2)\,\mathbf{x}_1. \tag{5.12}$$

With the non-random selection model, $v_1 = v_2 = r = 0$, so that if $\mathbf{L}_0$ denotes the corresponding vector of the expected change in the frequency of the A-gene at the limit, we have

$$\mathbf{L}_0 = 2N(\bar{s}_1 - \bar{s}_2)\left(1 - \frac{\bar{s}_1 + \bar{s}_2}{2}\right)\mathbf{x}_1 + \frac{N^2(4N-1)}{3N-1}(\bar{s}_1 - \bar{s}_2)^2 \mathbf{x}_2. \tag{5.13}$$

We can then express (5.12) as

$$\mathbf{L} = \mathbf{L}_0 - N(v_1 - v_2)\,\mathbf{x}_1 + \frac{N^2}{3N-1}(v_1 + v_2 - 2r)\,\mathbf{x}_2. \tag{5.14}$$

We can now examine $\mathbf{L}$ for three cases considered by previous workers on the problem of random variations in selection intensities.

(a) *Asymmetric case*

$\bar{s}_2 = v_2 = r = 0$, $v_1 > 0$. The changing environment allows variability only in the selection coefficient attached to the A-gene while its counterpart a-gene maintains the same constant selective value independent of the environmental background. We have

$$\mathbf{L}_0 = 2N\bar{s}_1(1-\bar{s}_1/2)\,\mathbf{x}_1 + \frac{N^2(4N-1)}{3N-1}\,\bar{s}_1^2\,\mathbf{x}_2, \tag{5.15}$$

$$\mathbf{L} = \mathbf{L}_0 - Nv_1\left[\mathbf{x}_1 - \left(\frac{N}{3N-1}\right)\mathbf{x}_2\right]. \tag{5.16}$$

If the initial frequency of **A** is $p_i = i/2N$, these expressions reduce to

$$u_0(p_i) = p_i + 2N\bar{s}_1(1-\bar{s}_1/2)\,p_i(1-p_i) + N^2\bar{s}_1^2\left(\frac{4N-1}{3N-1}\right)p_i(1-p_i)(1-2p_i), \tag{5.17}$$

$$u(p_i) = u_0(p_i) - Nv_1\left[\frac{2N(1+p_i)-1}{3N-1}\right]p_i(1-p_i). \tag{5.18}$$

This shows that the fixation probability $u(p_i)$ under random fluctuations in s_1 is always smaller than its value $u_0(p_i)$ under non-random selection. This is in conformity with the result of Karlin & Levikson (1974) obtained from the diffusion approach. However, they do not give any explicit formula for this case. We can, however, give such a formula by assuming N to be very large while $N\bar{s}_1$ and Nv_1 remain constant. The expressions (5.17) and (5.18) then give

$$u(p_i) = p_i + 2(N\bar{s}_1)\,p_i(1-p_i) + \tfrac{4}{3}(N\bar{s}_1)^2\,p_i(1-p_i)(1-2p_i) - \tfrac{2}{3}(Nv_1)\,p_i(1-p_i^2), \tag{5.19}$$

showing that the fixation probability is now a function of p_i, $N\bar{s}_1$ and Nv_1. For extremely small populations we have to use (5.18) in conjunction with (5.17).

(b) *Symmetric case in the sense of Karlin & Levikson (1974)*

In this case $\bar{s}_1 = \bar{s}_2$, $v_1 = v_2 = v > 0$, $r = 0$. Thus (5.13) and (5.14) respectively reduce to

$$\mathbf{L}_0 = \mathbf{0}, \tag{5.20}$$

$$\mathbf{L} = \frac{2N^2}{3N-1}\,v\mathbf{x}_2. \tag{5.21}$$

For the initial frequency p_i, we get

$$u_0(p_i) = p_i, \tag{5.22}$$

$$u(p_i) = p_i + \frac{2N^2v}{3N-1}\,p_i(1-p_i)(1-2p_i). \tag{5.23}$$

This shows that the fixation probability is smaller than p_i for $p_i > \frac{1}{2}$ and greater than p_i if $p_i < \frac{1}{2}$, which is consistent with what was found by Karlin & Levikson (1974). For large N and small v such that Nv remains constant, we get the approximation

$$u(p_i) = p_i + \tfrac{2}{3}(Nv)\,p_i(1-p_i)(1-2p_i) \tag{5.24}$$

giving that the fixation probability is a function of p_i and Nv.

(c) *Symmetric case in the sense of Jensen & Pollak (1969)*

We now have $\bar{s}_1 = \bar{s}_2 = \bar{s}$, $v_1 = v_2 = v > 0$, $r = -\bar{s}_1\bar{s}_2 = -\bar{s}^2 \geqslant -v$, which would result if either $s_1 > 0$, $s_2 = 0$ or $s_1 = 0$, $s_2 > 0$ in every generation. Expressions (5.13) and (5.14) now take the form

$$\mathbf{L}_0 = \mathbf{0}, \tag{5.25}$$

$$\mathbf{L} = \left(\frac{2N^2}{3N-1}\right)(v+\bar{s}_2)\,\mathbf{x}_2. \tag{5.26}$$

For initial frequency p_i, we get $u_0(p_i) = p_i$ as in Case (*b*) but $u(p_i)$ becomes

$$u(p_i) = p_i + \frac{2N^2(v+\bar{s}^2)}{3N-1}\,p_i(1-p_i)(1-2p_i), \tag{5.27}$$

which is the same as (5.23) except that v is replaced by $(v+\bar{s}^2)$. Thus we have essentially the same result as in case (*b*). Once again $u(p_i) > p_i$ if $p_i < \frac{1}{2}$ and $u(p_i) < p_i$ for $p_i > \frac{1}{2}$. This agrees with what was found by Jensen (1973). For large N and small v as well as $\bar{s}$ such that Nv and $N\bar{s}$ remain constant, the fixation probability is now a function of p_i, $N\bar{s}^2$ and Nv.

Going back to the general expressions (5.13) and (5.14) we can derive results which cover all the three cases mentioned above.

If the initial frequency of the A-gene is $p_i = i/2N$, (5.13) and (5.14) reduce to

$$u_0(p_i) = p_i + 2N(\bar{s}_1-\bar{s}_2)\left(1-\frac{\bar{s}_1+\bar{s}_2}{2}\right)p_i(1-p_i) + N^2(\bar{s}_1-\bar{s}_2)^2\left(\frac{4N-1}{3N-1}\right)p_i(1-p_i)(1-2p_i), \tag{5.28}$$

$$u(p_i) = u_0(p_i) - \frac{N}{3N-1}\,[(3N-1)(v_1-v_2) - N(v_1+v_2-2r)(1-2p_i)]\,p_i(1-p_i). \tag{5.29}$$

For large values of N and small values of $\bar{s}_1$, $\bar{s}_2$, v_1, v_2 and r such that $N(\bar{s}_1-\bar{s}_2)$, Nv_1, Nv_2 and Nr remain constant, we get the approximations:

$$u_0(p_i) = p_i + 2N(\bar{s}_1-\bar{s}_2)\,p_i(1-p_i) + \tfrac{4}{3}\{N(\bar{s}_1-\bar{s}_2)\}^2\,p_i(1-p_i)(1-2p), \tag{5.30}$$

$$u(p_i) = u_0(p_i) + \tfrac{2}{3}[(Nv_1+Nv_2-2Nr)\,p_i - (2Nv_2-Nv_1-Nr)]\,p_i(1-p_i). \tag{5.31}$$

We thus find that $u(p_i)$ is smaller or larger than $u_0(p_i)$ depending upon whether p_i is less than or greater than $(2v_2-v_1-r)/(v_1+v_2-2r)$ because $(v_1+v_2-2r) \geqslant 0$ in view of (2.3) and (2.4). If we define

$$\alpha = (2v_2-v_1-r)/(v_1+v_2-2r) \tag{5.32}$$

we find that, in general, the fixation probability of the A-gene, under random fluctuation in selection intensities, is determined solely by p_i, $N(\bar{s}_1-\bar{s}_2)$,

$N(v_1+v_2-2r)$ and α. It is thus given by

$$u(p_i) = p_i + 2\{N(\bar{s}_1-\bar{s}_2)\}p_i(1-p_i) + \tfrac{4}{3}\{N(\bar{s}_1-\bar{s}_2)\}^2 p_i(1-p_i)(1-2p_i) - \tfrac{2}{3}\{N(v_1+v_2-2r)\}p_i(1-p_i)(\alpha-p_i), \quad (5.33)$$

where α is given by (5.32).

While expression (5.33) is only approximate, it is a generally applicable explicit expression that holds regardless of the nature of the joint distribution of s_1 and s_2 in a generation. Previous authors have only been able to obtain an explicit solution in the symmetric cases. Moreover, expression (5.29) is valid for any population size, however small, with the restriction that we neglect moments of the third and higher orders in the selection intensities. This appears to be a new result in the literature.

The authors are grateful to Professor Oscar Kempthorne for helpful advice and valuable discussions. They are also grateful to Professor Alan Robertson for editorial suggestions which resulted in a considerable improvement in presentation.

This work received partial support from the National Institute of Health, Grant GM 13827.

REFERENCES

Ewens, W. J. (1973). Conditional diffusion processes in population genetics. *Theoretical Population Biology* **4**, 21–30.

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**, 321–341.

Gillespie, J. H. (1973). Natural selection with varying selection coefficients: a haploid model. *Genetical Research* **21**, 115–120.

Jensen, L. (1973). Random selective advantages of genes and their probabilities of fixation. *Genetical Research* **21**, 215–219.

Jensen, L. & Pollak, E. (1969). Random selective advantages of a gene in a finite population. *Journal of Applied Probability* **6**, 19–37.

Karlin, S. & Levikson, B. (1974). Temporal fluctuations in selection intensities: Case of small population size. *Theoretical Population Biology* **6**, 383–412.

Kemeny, J. G. & Snell, J. L. (1960). *Finite Markov Chains*. New York: D. van Nostrand.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719.

Kolmogorov, A. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen* **104**, 415–458.

Narain, P. & Robertson, A. (1969). Limits and duration of response to selection in finite populations: The use of transition probability matrices. *Indian Journal of Heredity* **1**, 1–19.

Narain, P. (1974). The conditional diffusion equation and its use in population genetics. *Journal of the Royal Statistical Society* B **36**, 258–266.

Ohta, T. (1972). Fixation probability of a mutant influenced by random fluctuation of selection intensity. *Genetical Research* **19**, 33–38.

Bull. Int. Stat. Inst., 1977, XLVII (2) 165–170.

12

ON THE AVERAGE AGE OF A MUTANT IN FINITE POPULATIONS

PREM NARAIN

Institute of Agricultural Research Statistics
New Delhi-110012, India

1. Introduction

Investigations made by Kimura and Ohta (1973) as well as Maruyama (1974) showed that the average age of mutants at particular frequency segregating in a finite population with effective size N_e can be quite old. The theory presented therein consider a situation in which a mutant allele with initial frequency p subsequently reaches a higher frequency x before it is either lost from the population or fixed in it. However, if $4N_ev$ is less than 1 (v being the mutation rate per locus per generation) and the number of possible allelic states per locus is so large, that, whenever, mutation occurs, it leads to a new (not a pre-existing) allele, there would be a dynamic equilibrium between the occurrence of new alleles by mutation and the loss of existing ones by random drift. In such a case, it is more appropriate to consider only those sample paths of the process which lead to the loss of the mutant allele by random drift. In a study on the survival of recessive lethals in finite populations by Robertson and Narain (1971), such a requirement was automatically met, since there was only one absorbing state *viz.* the status of the population with no heterozygotes. But in general there could be several absorbing states and it becomes necessary to condition the process relative to the event of the process attaining absorption at one of the boundaries. In this paper, therefore, a theory is presented which could give the average age of a mutant at a particular frequency x conditional to its eventual loss from the finite population.

2. Theory

We follow a method similar to that developed in Kimura and Ohta (1973) but with the difference that instead of forward Kolmogorov equation, a conditioned forward diffusion equation relative to the event of the process attaining absorption at $x = 0$ (introduced in Narain, 1974), is used. Let $f_{co}\ (p, x; t)$ be the probability density that the frequency of the mutant becomes x at time t, given that it was p at the start ($t = 0$) and that it would be zero

at the end ($t = \infty$). This density function satisfies the partial differential equation

$$\frac{\partial f_{co}(p, x; t)}{\partial t} = (1/2)\frac{\partial^2}{\partial x^2}[v(x)f_{co}(p, x; t)]$$

$$-\frac{\partial}{\partial x}\left[\left\{m(x)-\frac{v(x)G(x)}{1-u(x)}\right\}f_{co}(p, x; t)\right] \quad \ldots \quad (1)$$

where $m(x)$ is the instantaneous drift coefficient, $v(x)$ is diffusion coefficient and equals $x(1-x)/2N_e$, $u(x)$ is the probability of fixation of an allele with initial frequency x and $G(x) = \frac{du(x)}{dx}$. The initial and final conditions for the above equations are $f_{co}(p, x; 0) = \delta(x-p)$, $f_{co}(p, x; \infty) = \delta(x)$ where $\delta(.)$ stands for Dirac's delta function. Let

$$T_{co}^{(i)}(p, x) = \int_0^\infty t^i f_{co}(p, x; t)\, dt \quad \ldots \quad (2)$$

be the i-th moment ($i = 0, 1, 2, \ldots$) of t, the time interval in gen erationsfor an allele to have frequency x starting with initial frequency p and relative to the event that the allele eventually disappears from the population. It is found that $T_{co}^{(i)}(p, x)$ satisfies the differential equation

$$\left(\frac{1}{4N_e}\right)\frac{d^2}{dx^2}[x(1-x)T_{co}^{(i)}(p, x)]$$

$$-\frac{d}{dx}\left[\left\{m(x)-\frac{x(1-x)G(x)}{2N_e(1-u(x))}\right\}T_{co}^{(i)}(p, x)\right]+i\,T_{co}^{(i-1)}(p, x) = 0 \quad \ldots \quad (3)$$

with the boundary condition that as x approaches zero, $T_{co}^{(i)}(p, x)$ approaches a finite quantity.

With $i = 0$ we get $T_{co}^{(0)}(p, x)$, the sojourn time conditional to the loss of the allele or the expected number of visits to a particular frequency x in respect of those sample paths which lead to the loss of the allele from the population. For the neutral case, $m(x) = 0$ and $u(x) = x$ and the solution of the resulting differential equation turns out to be $4N_e$, being independent of p and x. But with $4Nv$ less than 1, $m(x) = -vx$ and $u(x) = 1-(1-x)^F$ where $F = 1-4N_ev$ and we get the solution

$$T_{co}^{(0)}(p, x, v) = \frac{4N_e}{F}\left[\frac{1-(1-x)^F}{x}\right] \quad \ldots \quad (4)$$

Since $2Nv$ mutants occur in a population of size N, on an average, in each generation, $2Nv\ T_{\infty}^{(0)}(p, x, v)$ gives the density of the gene frequency distribution, at equilibrium, among those alleles which are ultimately extinct from the population.

For obtaining the average age of an allele having frequency x at present given that it started with an initial frequency p and that it would ultimately disappear from the population, we take $i = 1$ in (3) and solve the differential equation for $T_{\infty}^{(1)}(p, x)$.

The average age of the allele at frequency x given that it disappears ultimately from the population is then given by

$$E[t_{\infty}(p, x)] = T_{\infty}^{(1)}(p, x)/T_{\infty}^{(0)}(p, x) \qquad \ldots (5)$$

It may be noted that the boundary conditions imposed for solving the differential equation are now such that as x approaches zero, the average age approaches the mean time until loss of the allele as given in Narain (1974). For neutral alleles, we obtain finally

$$E[t_{\infty}(p, x)] = -4N_e \left[\left(\frac{p}{1-p} \right) \log_e p + \left(\frac{1-x}{x} \right) \log_e (1-x) + 1 \right] \qquad \ldots (6)$$

When a neutral mutant is represented only once at the moment of its appearance in the population, we may put $p = 1/2N$ in (6) and obtain

$$E\left[t_{\infty} \left(\frac{1}{2N}, x \right) \right] = 2 \left(\frac{N_e}{N} \right) \log_e (2N) - 4N_e \left[\left(\frac{1-x}{x} \right) \log_e (1-x) + 1 \right] \qquad \ldots (7)$$

For the model of infinite isoalleles with $4N_v v$ less than unity and initial frequency as p, we get, approximately

$$E[t_{\infty}\ (p, x, v)]$$

$$= \frac{4N_e}{F} \left[\int_x^1 \frac{1-(1-x)^F}{x}\, dx + \frac{(1-x)^F}{1-(1-x)^F} \int_0^x \frac{\{1-(1-x)^F\}^2}{x(1-x)^F}\, dx \right.$$

$$\left. -(1-p)^{-F} \int_p^1 \frac{(1-p)^F - (1-x)^F}{x}\, dx \right] \qquad \ldots (8)$$

As v approaches zero, (8) reduces to (6).

For obtaining the variance of the age of an allele having frequency x at present, given that it started with frequency p and that it would eventually disappear from the population, we take $i = 2$ in (3) and solve the differential

equation for $T_{co}^{(2)}(p, x)$. The boundary conditions imposed are such that as x approaches zero,

$$E[t_{co}^2(p, x)] = T_{co}^{(2)}(p, x)/T_{co}(p, x) \quad \ldots \quad (9)$$

approaches the mean square time until loss of the allele as given in Narain (1974). The variance of the age is then obtained from

$$V[t_{co}(p, x)] = E[t_{co}^2(p, x)] - \{E[t_{co}(p, x)]\}^2 \quad \ldots \quad (10)$$

For neutral alleles, we obtain finally

$$E[t_{co}^2(p, x)] = 32N_e^2 \Big[\Big\{ 2 + \Big(\frac{1-x}{x} \Big) \log_e(1-x) \Big\} \Big\{ 1 + \Big(\frac{p}{1-p} \Big) \log_e p \Big\}$$

$$- \int_x^p \frac{\log x}{1-x} dx + \Big(\frac{1-x}{x} \Big) \log_e (1-x) + (\log_e x)(\log_e (1-x)) \Big] \quad \ldots \quad (11)$$

3. Discussion

The sojourn time of a neutral mutant at frequency x conditional to its eventual loss under the assumption of no further mutation is shown to be independent of the initial frequency as well as the present frequency. The quantity $T_{co}^{(0)}(p, x)/2N$ then becomes $(2N_e/N)$ and therefore gives the average number of generations spent at gene frequency $i/2N$, $i = 1, 2, \ldots (2N-1)$ by the mutant allele, provided that it starts with any intermediate frequency and is eventually lost from the population. This result is in contrast with the result given in Pollak and Arnold (1975) when eventual fixation is considered. When we consider the model of infinite isoalleles with mutation rate v, the sojourn time conditional to the loss, as given by relation (4), is found to be different from relation (8) of Maruyama (1974). The average age of a neutral allele under the assumption of no further mutation ($v = 0$) as given by relation (6) is similar to relation (11) of Kimura and Ohta (1973) if we inter-change p with $(1-p)$ and x with $(1-x)$. It appears therefore that the average age discussed by them is conditional to eventual fixation of the gene. However, they do not clarify this point in their discussion. If we allow p to approach unity, the relation (6) gives $-4N_e \left(\frac{1-x}{x} \right) \log_e(1-x)$ which happens to be the average time until fixation of an allele with initial frequency x. This result is quite expected since the ultimate value of x is zero. For neutral mutants occuring initially once only, the average age until extinction is given by relation (7) which approach, as expected, to average time until loss of the gene

if we let x approach zero. In a similar way we find that for the model of infinite isoalleles, if we let p approach unity in (8), the average age is exactly the same as relation (13b) of Kimura and Ohta (1973). But if we take $p = (1/2N)$ and let x approach zero, we find that the average age approaches $2(N_e/N)(1-2N_ev/N)\log_e(2N)$. This corresponds to the average time until loss of the gene and could be smaller than the average extinction time for neutral mutants with $v = 0$. However, if one is interested in working out the average age irrespective of the fact whether the allele is destined to be fixed or lost from the population, the mathematical treatment given in Kimura and Ohta (1973) would still hold but the condition of the average age approaching average time until fixation would have to be replaced by the average age approaching average time until homozygosity as x approaches unity. The mean square age, for neutral alleles, as given by (11), approaches the mean square time until fixation of the allele with initial frequency x as given in Narain (1974) if we let p approach unity. On the other hand, if we let x approach zero in (11), we get the mean square time until loss of the allele with initial frequency p as given in Narain (1974).

Although the approach presented in this paper relates to conditioning the underlying process relative to the event of eventual loss of the allele from the population, a similar method could be followed for the case of determining the average age of a mutant at a given frequency conditional to its eventual fixation in the finite population. Instead of the partial differential equation given in (1), we would have to take the conditioned forward diffusion equation relative to the event of the process attaining absorption at $x = 1$ as introduced in Narain (1974).

References

KIMURA, M. and OHTA, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics*. **75**, 199-212.

MARUYAMA, T. (1972). The average number and the variance of generations at particular gene frequency in the course of fixation of a mutant gene in a finite population. *Genet. Res., Camb.*, **19**, 109-113.

MARUYAMA, T.(1974). The age of an allele in a finite population. *Genet. Res., Camb.*, **23**, 137-143.

NARAIN, P. (1974). The conditional diffusion equation and its use in population genetics. *J.R.S.S.*, **B**. **36**, 258-266.

POLLAK, E. and ARNOLD, B. C. (1975). On sojourn times at particular gene frequencies. *Genet. Res., Camb.*, **25**, 89-94.

ROBERTSON, A. and NARAIN, P. (1971). The survival of recessive lethals in finite populations. *Theor. Pop. Biol.*, **2**, 24-50.

Key words

Age of a mutant, conditioned sojourn time, expectation of age conditional to extinction.

Abstract

The expectation of the age of a mutant at particular frequency in finite populations is discussed. With dynamic equilibrium in e model of infinite isoalleles but with low mutation rate to allow temporary fixation or loss, age of a mutant conditional to its loss is considered to be more appropriat than its age irrespective of fixation or loss.

Résumé

On a étudié la probabilité de l' âge d' un mutant ā une fréquence particulière dans des populations limitées. Dans des conditions d'équilibre dynamique dams un modèle ā isolallèles en nombre infini mais avec un taux de mutation peu élevé afin de permettre la fixation ou la perte provisoire, on considère que l' âge d' un mutant dépendant de sa perte est plus approprié que son âge ne tenant pas compte de la fixation ou de la perte.

[REPRINTED FROM THE *Indian Journal of Genetics and Plant Breeding*, Vol. 38, No. 1, March 1978, pp. 57-62]

13

AVERAGE TIME UNTIL FIXATION OF A MUTANT AT A TRI-ALLELIC LOCUS IN A FINITE POPULATION

PREM NARAIN

Institute of Agricultural Research Statistics, Library Avenue, New Delhi-110 012 (*India*)

FOR a selectively neutral di-allelic locus and initial gene frequency tending to zero, Kimura and Ohta (1969) as well as Narain (1970) found that it takes, on an average, $4N_e$ generations to reach fixation in a population of effective size N_e. These studies utilised diffusion approximations. On the other hand, Narain and Robertson (1969) developed a more exact method of transition matrix using finite Markov Chain to obtain similar results. The advantage of this method is that one can extend it to more complex genetic situations and use a powerful computer to obtain numerical results. In this paper we employ this method for working out the average number of generations required to fix a mutant at a tri-allelic locus.

METHOD OF TRANSITION MATRIX FOR A TRI-ALLELIC LOCUS

Consider a finite population of gametes of constant size 2N in a population of diploid individuals of size N and a single locus with three alleles A_1, A_2 and A_3. The sampling of gametes due to finite size of the population causes a random change, from generation to generation, in the frequency of a given allele. The population gets sub-divided into several lines with different gene frequencies but with random breeding within the lines. These lines can be of seven kinds. The first three kinds of lines have alleles A_1 or A_2 or A_3 only representing respectively the fixation of the corresponding allele. The next three kinds of lines are those in which either alleles A_1 and A_2 or alleles A_1 and A_3 or alleles A_2 and A_3 are segregating. The seventh type is such that all the three alleles are segregating. In this case there are (N-1) (2N-1) possibilities corresponding to (N-1) (2N-1) alternative ways of the mixture of A_1, A_2 and A_3. Lines of this kind are capable of transition to the same kind or to any of the other six kinds. On the other hand, lines of the 4th, 5th and 6th kind are capable of transition to the same kind or to any two of the first three kinds of lines depending upon the types of segregating alleles. Theoretically, if $P_{(i_1 i_2)(j_1 j_2)}$ represents the conditional probability that there are j_1, A_1 genes, j_2, A_2 genes out of 2N genes in the line, given that there were i_1, A_1 genes and i_2, A_2 genes in this line in the previous generation, the $(N+1)(2N+1) \times (N+1)(2N+1)$ transition probabilities determine a transition matrix **P** given by

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}$$

where $\mathbf{I}$ is a 3×3 unit matrix, $\mathbf{O}$ is $3 \times (N+2)(2N-1)$ matrix of zeros, $\mathbf{R}$ is a $(N+2)(2N-1) \times 3$ matrix of one step fixation probabilities of A_1, A_2 and A_3 and $\mathbf{Q}$ is a $(N+2)(2N-1) \times (N+2)(2N-1)$ matrix given by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{12} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{Q}_{13} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{Q}_{23} & \mathbf{O} \\ \mathbf{D}_{12} & \mathbf{D}_{13} & \mathbf{D}_{23} & \mathbf{Q}^* \end{bmatrix}$$

where $\mathbf{Q}_{ij}$ gives the transition probabilities between states of 4th or 5th or 6th kinds of lines, $\mathbf{D}_{ij}$ gives the transitions from the seventh type to any one of 4th or 5th or 6th types and $\mathbf{Q}^*$ represents the $(N-1)(2N-1) \times (N-1)(2N-1)$ matrix of transitions between lines of the seventh kind only.

If we define the matrix sum $\mathbf{T}=\mathbf{I}+\mathbf{Q}+\mathbf{Q}^2+\ldots=(\mathbf{I}-\mathbf{Q})^{-1}$, the nature of the partitioned matrices given above shows that the structure of $\mathbf{T}$ would be

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{12} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{T}_{13} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{T}_{23} & \mathbf{O} \\ \mathbf{T}^*\mathbf{D}_{12}\mathbf{T}_{12} & \mathbf{T}^*\mathbf{D}_{13}\mathbf{T}_{13} & \mathbf{T}^*\mathbf{D}_{23}\mathbf{T}_{23} & \mathbf{T}^* \end{bmatrix}$$

where $\mathbf{T}_{12}=(\mathbf{I}-\mathbf{Q}_{12})^{-1}$; $\mathbf{T}_{13}=(\mathbf{I}-\mathbf{Q}_{13})^{-1}$; $\mathbf{T}_{23}=(\mathbf{I}-\mathbf{Q}_{23})^{-1}$; $\mathbf{T}^*=(\mathbf{I}-\mathbf{Q}^*)^{-1}$.

The fixation probability vector U_1 corresponding to A_1 allele consists of four component vectors U_{12}, U_{13}, U_{23} and U^* and are obtained by operating $\mathbf{T}$ on the corresponding component of $\mathbf{R}$. The mean time until fixation of the allele A_1, disregarding the cases in which it is lost, is then given by the ratio of elements of vector M_1 and U_1 where M_1 consists of four component vectors M_{12}, M_{13}, M_{23} and M^* which are given by $M_{12}=\mathbf{T}_{12}\,U_{12}$; $M_{13}=\mathbf{T}_{13}\,U_{13}$; $M_{23}=O$; $M^*=\mathbf{T}^*(\mathbf{D}_{12}M_{12}+\mathbf{D}_{13}M_{13}+U^*)$.

The above method shows that the computational procedure for the evaluation of the mean time until fixation in a given genetic situation requires the inversion and various operations on matrices of orders depending upon the population size. As such, computations can easily be carried out on a powerful computer. However, this computation depends on the specification of the transition probabilities which ultimately depend on the particular genetic situation and the model of sampling of gametes. Here we use trinomial sampling model where $P_{(i_1\,i_2)\,(j_1\,j_2)}$ is given by

$$P_{(i_1\,i_2)\,(j_1\,j_2)} = \frac{(2N)!}{j_1!j_2!(2N-j_1-j_2)!}\,(q'_{1_i})^{j_1}\,(q'_{2_i})^{j_2}\,(1-q'_{1_i}-q'_{2_i})^{(2N-j_1-j_2)}$$

where q'_{1i} and q'_{2i} are the respective frequencies of A_1 and A_2 after one generation of selection, starting with their frequencies in the previous generation as $q_1=i_1/2N$ and $q=i/2N$. In what follows we assume that the relative selective advantages of A_1, A and A_3 alleles are respectively $(1+s_1/2)$, $(1+s/2)$ and 1. In such a genetic situation, the changes in the gene frequencies due to selection are functions of q_1, q, the average superiority $\bar{s}$ and the variance in superiority V_m of A_1 over A_2 and A_3 taken together as a group. These are given by $\bar{s}=(s_1-rs)/2$; $V_m=r(1-r)s^2_2/4$; $r=q_2/(1-q_1)$.

The variance in selective advantage of the three alleles can be partitioned as

$$V_s = q_1(1-q_1)s^{-2} + (1-q_1)V_m$$
$$=\sigma_a^2/2 + \sigma^2_m/2$$

where σ_a^2 is the contribution of the locus to the additive genetic variance when the three system A_1–A_2–A_3 is collapsed into two allele system A_1–$\bar{A}_2$ with $\bar{A}_2$ referring to the group of A_2 and A_3 and σ^2_m is the component of total genetic variance due to the distinction made between A_2 and A_3 on the basis of their different selective advantages. This idea of distinguishing multiple alleles on the basis of their quantitative effects was given by Narain (1965).

With finite population, the diffusion approach shows that the selection process is entirely governed by Ns, and Ns_2 on a time scale proportional to 1/N starting from the given initial configuration (q_1, q_2). It then follows that the probability of fixation of A_1 and hence the average time until its fixation would depend on $N\bar{s}$, N^2V_m as well as the initial gene frequencies q_1 and q_2.

Results and Discussion

For a neutral tri-allelic locus, when $s_1=s_2=0$, the mean time until fixation of A_1 is found to be independent of the frequency of alleles A_2 or A_3 and is the same as that for a neutral di-allelic locus. Thus we have from the diffusion approach, the mean time $m(q_1)$ as

$$-4N_e\left(\frac{1-q_1}{q_1}\right)\log_e(1-q_1).$$

When selective forces are operating the mean time was calculated for $N=6$ and all possible initial gene frequencies. As noted earlier, it is affected by $N\bar{s}$, N^2V_m and the initial frequencies q_1 and q_2 so that one could apply the results to other population sizes. Since the mutants produced artificially would be at low frequencies initially, we present the results only for the following three cases: (a) $q_1=0\cdot0833$, $q_2=0\cdot0833$; (b) $q_1=0\cdot0833$, $q_2=0\cdot4167$; (c) $q_1=0\cdot3333$, $q_2=0\cdot3333$.

The results obtained are given in Tables 1, 2 and 3.

(a) *Average number of generations/N until fixation of* A_1 *with* $q_1=q_2=0\cdot0833$.

From Table 1, it is found that the mean time decreases with increase in $N\bar{s}$ at all values of N^2V_m. For values of $N\bar{s}$ from 0 to 4, the effect of multiple

allelic variance is to decrease the mean timc. But when $\bar{Ns}$ lies between 5 to 8, the mean time decreases first with increase in N^2V_m but increases thereafter. When $\bar{Ns}=5$, the minimum lies at $N^2V_m=4$ whereas for $\bar{Ns}$ lying between 6 and 8 it lies at $N^2V_m=1$.

TABLE 1

Average number of generations/N until fixation of A_1 *with* $q_1=0\cdot0833$

$\bar{Ns}$	N^2V_m 0	1	4	9	16
0	3·57	3·39	3·01	2·64	2·33
1	3·40	3·08	2·72	2·40	2·14
2	3·03	2·76	2·52	2·28	2·07
3	2·66	2·46	2·35	2·22	2·06
4	2·36	2·21	2·17	2·15	2·08
5	2·11	2·01	2·00	2·04	2·06
6	1·92	1·84	1·85	1·91	1·99
7	1·77	1·70	1·71	1·78	1·89
8	1·64	1·59	1·60	1·67	1·77

(b) *Average number of generations/N until fixation of* A_1 *with* $q_1=0\cdot0833$, $q_2=0\cdot4167$

From Table 2, it is apparent that for N^2V_m equal to 0 and 1, the mean time decreases with increase in $\bar{Ns}$. However, for higher values of N^2V_m, it increases first and decreases thereafter with increase in $\bar{Ns}$. The maximum occurs at $\bar{Ns}=2$ with $N^2V_m=4$, at $\bar{Ns}=3$ with $N^2V_m=9$ and at $\bar{Ns}=4$ with $N^2V_m=16$. For $\bar{Ns}=0$ and 1, the multiple allelic effects decrease the mean time but at higher values of $\bar{Ns}$, the mean increases first and decreases thereafter. At $\bar{Ns}=2$, the maximum occurs when N^2V_m is equal to 1 and 4 whereas at $\bar{Ns}=3$, it occurs when $N^2V_m=9$. At higher values of $\bar{Ns}$, the maximum occurs presumably when N^2V_m is greater than 16.

TABLE 2

Average number of generations/N until fixation of A_1 *with* $q_1=0\cdot0833$ *and* $q_2=0\cdot4167$

$\bar{Ns}$	N^2V_m 0	1	4	9	16
0	3·57	3·37	2·99	2·62	2·33
1	3·40	3·29	3·04	2·75	2·50
2	3·03	3·09	3·09	2·96	2·76
3	2·66	2·81	3·01	3·07	3·00
4	2·36	2·53	2·81	3·04	3·11
5	2·11	2·28	2·57	2·88	3·08
6	1·92	2·07	2·34	2·66	2·95
7	1·77	1·90	2·14	2·45	2·76
8	1·64	1·76	1·97	2·25	2·55

(c) *Average number of generations/N until fixation of* A_1 *with* $q_1=q_2=0\cdot3333$.

It is apparent from Table 3 that when $N^2V_m=0$ or 1, the mean time decreases with increase in $\bar{Ns}$ but at higher values of N^2V_m, it increases first and decreases thereafter The maximum occurs at $Ns=2$ when $N^2V_m=4$, at $\bar{Ns}=3$ when $N^2V_m=9$ and at $\bar{Ns}=4$ when $N^2V_m=16$. The effect of multiple allelic variance is to decrease the mean time when $\bar{Ns}=0$ or 1, but at higher values of Ns, the mean time increases first with increase in N^2V_m and decreases thereafter. The maximum occurs at $N^2V_m=1$ when $\bar{Ns}=2$, at $N^2V_m=9$ when $\bar{Ns}=3$ and at N^2V_m greater than 16 for higher values of $\bar{Ns}$.

TABLE 3

Average number of generations until fixation of A_1 *with* $q_1=q_2=0\cdot3333$

$\bar{Ns}$	N^2V_m 0	1	4	9	16
0	3·00	2·81	2·45	2·11	1·84
1	2·83	2·75	2·51	2·23	1·99
2	2·48	2·55	2·54	2·40	2·20
3	2·12	2·28	2·44	2·48	2·39
4	1·84	2·00	2·24	2·42	2·47
5	1·61	1·76	2·01	2·26	2·43
6	1·44	1·57	1·79	2·06	2·29
7	1·31	1·42	1·61	1·86	2·11
8	1·20	1·30	1·46	1·68	1·92

From the above results, it is apparent that the behaviour of the mean time over variations in $\bar{Ns}$ for a fixed N^2V_m is almost the same as that of the mean time over variations in N^2V_m for a fixed $\bar{Ns}$ in the (b) and (c) cases. In the case at (a) where A_1 occurs once in a population having one representative of A_2 and the rest A_3, the behaviour of the mean time over variations in N^2V_m for a fixed but high value of $\bar{Ns}$ is reversed. It is also of interest to compare the results given in Tables 1 and 2. In both the cases, A_1 allele occurs initially once in the population but the frequency of A_2 allele differs. When N^2V_m is zero, the mean time, in both the cases, are exactly the same. This is according to expectation since with no multiple allelic effects, the system reduces to a 2-allele situation and the mean time is dependent on q_1 and $\bar{Ns}$ only. However, when N^2V_m is greater than zero and $\bar{Ns}$ is not zero, a greater mean time for fixation of A_1 is required when the frequency of A_2 is more. But when $\bar{Ns}=0$, slightly more time is needed for fixation of A_1 when frequency of A_2 is less. It is thus seen that production of a mutant artificially at a locus segregating already for two alleles will decrease the fixation time, on an average, provided there is variability in the superiority of the new mutant over its two pre-existing alleles. This finding may find important applications in mutation breeding in the sense that useful and viable mutants forming a multiple allelic series may take lesser number of generations to fix than those involving a di-allelic locus.

Summary

The average number of generations involved in fixing an artificially produced mutant at a tri-allelic locus has been studied mathematically. The results have been found to be affected by a combination of parameters describing population size and multiple allelic effect.

REFERENCES

Kimura, M. and T. Ohta (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61:** 763–771.

Narain, P. (1965). The description of gene action and interaction with multiple alleles in continuous variation. *Genetics*, **52:** 43–53.

Narain, P. (1970). A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. *Genetical Res. Camb.*, **15:** 251–255.

Narain, P. and A. Robertson. (1969). Limits and duration of response to selection in finite populations: the use of transition probability matrices. *Indian J. Hered.*, **1:** 79–97.

14

Proc. Indian natn. Sci. Acad. B 45 No. 3 pp 239-246 (1979)

On the Statistical Properties of the Conditional Equilibrium Distribution under Steady Flux of Mutations*

PREM NARAIN**, FNA
Statistical Laboratory, Iowa State University
Ames, Iowa 50011, USA

(Received 26 September 1978; after revision 27 March 1979)

The statistical properties of a conditional equilibrium distribution of mutant frequency resulting from the balance between the continued production of new mutants over many generations and their loss from the population because of random drift are discussed. The revised estimates of the average number of heterozygous sites in mammals are found to be lower than those given earlier in which the underlying stochastic process is not conditioned.

Introduction

Evidence from such diverse organisms as man, mouse, fruit fly and horseshoe crab has accumulated to show that there exists very high variability at the molecular level (Selander et al. 1970). Various mathematical models have been developed to give a framework in which molecular polymorphisms could be discussed. One such model is "the model of infinite sites" (Kimura 1969). In this model, it is assumed that the total number of nucleotide sites available for mutation is so large and that the mutation rate per site is so low that whenever a mutant appears, it represents a mutation at a new site. Using this model, Kimura (1969) obtained a formula for $H(p)$, the expected total number of heterozygous sites per individual maintained in a finite population because of steady flux of mutations with frequency of the mutant at the moment of its occurrence at each site as p. The method of obtaining this formula is based on considering all the sample paths of the underlying stochastic process resulting in either loss of the mutant from the population or fixation in it within a finite length of time. Although we do not know whether the mutant, at its initial occurrence, with frequency p, is going to be eventually lost or fixed, we do know that the probabilities of these two eventualities are $1-u(p)$ and $u(p)$ respectively, where $u(p)$ stands for the probability of fixation and equals p for neutral genes. Using these probabilities, one can invoke a

*Journal Paper No. J-8637 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa Project 1669. Partial support by National Institute of Health, Grant GM 13827.
**On leave from the Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India.

conditioned process in which the loss of the allele is made certain if there is no production of new mutants. The recurrence of new mutants is then balanced by their loss only and not by both fixation as well as loss. Such a situation would occur if the population size is very large and the selection forces are weak. $u(p)$ would then be small. However, if the event of fixation does occur, the conditional expectation would be much larger than if the gene is lost. In such a situation, unconditional expectations of Kimura (1969) would be misleading because they may give too heavy a weight to sample paths that rarely occur. It is, therefore, more appropriate to consider only such sample paths that lead to the loss of the mutant form from the population disregarding those in which they are fixed. The purpose of this paper is therefore to study the statistical properties of the conditional equilibrium distribution under steady flux of mutation. In particular, a formula for the average number of heterozygous nucleotide sites per individual maintained in a finite population because of steady flux of mutations conditional to their random extinction will be derived, and revised estimates of the average number of heterozygous sites in mammals will be presented. The theory developed is also applicable to a cistron consisting of at least several hundred nucleotide sites.

The Theory

We assume that, on the average, in each generation mutant forms appear in a population of constant size N in ν_m nucleotide sites so that mutation rate per gamete is $\nu = \nu_m/2N$. We assume also "the model of infinite sites", viz., the total number of sites per individual is so large and mutation rate per site is so small that whenever a mutant appears, it represents a mutation at a previously homoallelic site. Let $f_{co}(p, x; t)$ be the conditional probability density that the frequency of the mutant in the population is p at the start ($t = 0$), given that it would be x at time t as well as that it would be zero at the end ($t = \infty$). This means that the process is viewed retrospectively, in the reserve time sequence so that x is regarded as fixed and p is taken as a random variable varying between 0 and 1. Then $f_{co}(p, x; t)$ satisfies the following conditional backward diffusion equation introduced by Narain (1974):

$$-\partial f_{co}(p, x; t)/\partial t = \overset{*}{L_0} f_{co}(p, x; t), \qquad \ldots(1)$$

where $\overset{*}{L_0}$ is an operator given by

$$\overset{*}{L_0} \equiv (1/2)\, V_{\delta p}\, (\partial^2/\partial p^2) + {}_0\overset{*}{M}_{\delta p}\, (\partial/\partial p), \qquad \ldots(2)$$

and

$${}_0\overset{*}{M}_{\delta p} = M_{\delta p} - V_{\delta p}\, G(p)/[1-u(p)], \qquad \ldots(3)$$

$$G(p) = \exp\left[-2\int_0^p (M_{\delta y}/V_{\delta y})\, dy\right]. \qquad \ldots(4)$$

where $M_{\delta p}$ and $V_{\delta p}$ denote the mean and variance of the change in the mutant frequency p per generation assumed same for all the sites. In other words, the mean and variance of the amount of change in mutant frequency p during a short interval from t to $t + \delta t$ are $M_{\delta p}\, \delta t$ and $V_{\delta p}\, \delta t$ respectively, both being independent of time parameter t, so that the conditioned process under study is time-homogenous. The boundary conditions for this process are

$$f_{co}(p, x; 0) = \delta(x - p), \qquad \ldots(5)$$

where $\delta(\,.\,)$ is Dirac delta function,

$$f_{co}(p, x; \infty) = 0, \ (0 < x < 1). \qquad \ldots(6)$$

Further, $u(p)$ is the eventual probability of fixation of the mutant given by

$$u(p) = \int_0^p G(x)\, dx \Big/ \int_0^1 G(x)\, dx. \qquad \ldots(7)$$

We consider only those sample paths of the mutant appearing in the finite population in which it is lost from the population within a finite time. A balance between the continued production of new mutants over

many generations and their loss from the population will then be established. We can therefore envisage a conditional stable distribution of the mutant frequencies in different sites, considering only those sites in which mutants are not lost. Since ν_m is the number of sites in which new mutations appear in the population in each generation, $\nu_m f_{co}(p, x; t)dx$ represents the contribution made by mutants which appeared t generations earlier with initial frequency p to the present frequency class in which the mutant frequencies are in the range from x to $x + dx$. Thus, considering all the contributions made by mutations in the past, the expected number of sites in which the mutants are in frequency range x to $x + dx$ in the present generation conditional to their loss from the population, is $\phi_{co}(p, x)\,dx$ where

$$\phi_{co}(p, x) = \nu_m \int_0^\infty f_{co}(p, x; t)\,dt, \quad (0<x<1) \qquad ...(8)$$

is the conditional stable distribution under steady flux of mutations. The statistical properties of distribution can be studied by deriving the expectation of an arbitrary function $g(x)$, differentiable up to the second order at p, with respect to this distribution. We denote such an expectation (functional) by $\overset{g}{I}_{co}(p)$, that is,

$$\overset{g}{I}_{co}(p) \equiv \int_0^1 g(x)\,\phi_{co}(p, x)\,dx$$

$$= \nu_m \int_0^\infty \left[\int_0^1 g(x) f_{co}(p, x; t)\,dx\right] dt \qquad ...(9)$$

Regarding the process in the change of gene frequency as a collection of sample paths $\{w\}$ and denoting by $x(w, t)$ the position of a particular path w at time t, the above expectation can also be expressed alternatively, in accordance with the theory developed in Maruyama and Kimura (1971, 1975), as

$$\overset{g}{I}_{co}(p) = E\left[\int_0^{\tau(w)} g(x(w, t))\,dt / x(w, \tau(w)) = 0\right] \qquad ...(10)$$

where $\tau(w)$ is the time when path w exits from the interval (0, 1) with $x(w, \tau(w))=0$ and $E[\ \ldots\ ]$ stands for the expectation with respect to the sample paths that start from p at time 0, i.e. $x(w, 0)=p$ and lead to eventual extinction of the allele.

Multiplying each term of (1) by $\nu_m g(x)$ and then integrating each of the resulting term first with respect to x in the open interval (0, 1) and then with respect to t over $(0, \infty)$ gives

$$\int_0^\infty (\partial/\partial t)\left[\nu_m \int_0^1 g(x) f_{co}(p, x; t)\,dx\right] dt$$

$$= \overset{*}{L}_0 \overset{g}{I}_{co}(p) \qquad ...(11)$$

The L.H.S. of this equation gives

$$\left|\nu_m \int_0^1 g(x) f_{co}(p, x; t)\,dx\right|_0^\infty$$

$$= -\nu_m \int_0^1 g(x)\,\delta(x-p)\,dx \qquad ...(12)$$

in view of (5) and (6). It reduces to $-\nu_m g(p)$ because of

$$\int_0^1 g(x)\,\delta(x-y)\,dx = g(y). \qquad ...(13)$$

We thus see that $\overset{g}{I}_{co}(p)$ satisfies the ordinary differential equation

$$(1/2) V_{sp} (d^2 \overset{g}{I}_{co}(p)/dp^2) + {}_0\overset{*}{M}_{sp} (d \overset{g}{I}_{co}(p)/dp) + \nu_m g(p) = 0 \qquad ...(14)$$

Now mutations at $p=0$ do not contribute to segregating sites so that $f_{co}(0, x; t)=0$ for $0<x<1$, giving one of the boundary conditions as

$$\overset{g}{I}_{co}(0) = 0 \qquad ...(15)$$

However, because a mutant appearing in the population is destined to be lost in the conditional process under study, mutants at p tending to 1 will contribute to the segregating sites so that $\lim_{p\to 1} f_{co}(p, x; t)$ for $0<x<1$ will tend to be finite. This would give the other boundary condition as

$$\lim_{p\to 1} {}^{g}I_{c0}(p) = K, \qquad \ldots(16)$$

a finite quantity whose value can only be obtained by taking the limit in the final expression for ${}^{g}I_{c0}(p)$. The solution of eq. (14) which satisfies the boundary conditions (15) and (16) is given by

$${}^{g}I_{c0}(p) = \nu_m \int_0^p g(y)\, I(y)\, u(y)\,[1-u(y)]\, dy$$

$$+\nu_m \left[u(p)/(1-u(p))\right] \int_p^1 g(y)\, I(y) \left[1-u(y)\right]^2 dy \qquad \ldots(17)$$

$$\lim_{p\to 1} {}^{g}I_{c0}(p) = K = \nu_m \int_0^1 g(y)\, I(y)\, u(y)\, [1-u(y)]\, dy \qquad \ldots(18)$$

where

$$I(y) = \left[2 \int_0^1 G(x)\, dx\right] \Big/ (V_{sy})\, G(y)$$

$$= 2/V_{sy}\, u'(y) \qquad \ldots(19)$$

in which $u'(y) = du(y)/dy$.

Although of no immediate interest, we can similarly deal with the situation in which only those sample paths of the process which lead to the fixation of the mutants are considered. This gives a stable distribution $\phi_{c1}(p, x)$ as

$$\phi_{c1}(p, x) = \nu_m \int_0^\infty f_{c1}(p, x; t)\, dt, \quad (0<x<1) \qquad \ldots(20)$$

where $f_{c1}(p, x; t)$ satisfies the following backward diffusion equation conditional to fixation given by Narain (1974) :

$$-\partial f_{c1}(p, x; t)/\partial t = L_1^* f_{c1}(p, x; t) \qquad \ldots(21)$$

where the operator L_1^* is given by

$$L_1^* \equiv [(1/2) V_{sp} (\partial^2/\partial p^2) + {}_1M_{sp}^* (\partial/\partial p)] \qquad \ldots(22)$$

and

$${}_1M_{sp}^* = M_{sp} + V_{sp}\, G(p)/u(p) \qquad \ldots(23)$$

The expectation of an arbitrary function $g(x)$, differentiable up to the second order at p, with respect to $\phi_{c1}(p, x)$ and denoted by ${}^{g}I_{c1}(p)$ then satisfies the ordinary differential equation

$$(1/2)\, V_{sp}\, (d^2\, {}^{g}I_{c1}(p)/dp^2) + {}_1M_{sp}^* \, (d\, {}^{g}I_{c1}(p)/dp) + \nu_m\, g(p) = 0 \qquad \ldots(24)$$

subject to

$$\lim_{p\to 0} {}^{g}I_{c1}(p) = K, \text{ a finite quantity} \qquad \ldots(25)$$

$${}^{g}I_{c1}(1) = 0 \qquad \ldots(26)$$

The appropriate solution is found to be

$${}^{g}I_{c1}(p) = \nu_m \left[(1-u(p))/u(p)\right] \int_0^p g(y)\, I(y)\, [u(y)]^2\, dy + \nu_m \int_p^1 g(y)\, I(y)\, u(y)[1-u(y)]\, dy \qquad \ldots(27)$$

$$\lim_{p\to 0} {}^{g}I_{c1}(p) = K = \nu_m \int_0^1 g(y)\, I(y)\, u(y)\, [1-u(y)]\, dy \qquad \ldots(28)$$

Statistical Properties of the Conditional Stable Distribution

To study the statistical properties of the distribution, we have to specify the forms of the functions of M_{sx} and V_{sx} which depend on the genetic situation. We consider here the case of no dominance and assume that random fluctuation in mutant frequency is due to random sampling of gametes. Then

$$M_{s}T = \frac{s}{2}\, x(1-x) \qquad \ldots(29)$$

$$V_{sx} = x(1-x)/2N_e \qquad \ldots(30)$$

where N_e is the variance effective population size which may differ from actual size N if the distribution of the number of offspring does not follow Poisson distribution, and $(1+s)$, $(1+\frac{1}{2}s)$ and 1 are the respective fitness of the three genotypes AA, Aa and aa. With these forms of M_{sx} and V_{sx}, we have

$$G(x) = \exp(-2Sx) \quad \ldots(31)$$

$$u(x) = (1 - \exp(-2Sx))/(1 - \exp(-2S)) \quad \ldots(32)$$

$$I(x) = 2N_e(1 - \exp(-2S))(\exp(2Sx))/Sx(1 - x) \quad \ldots(33)$$

$$S = N_e S \quad \ldots(34)$$

The specification of the form of $g(p)$ depends on the statistical property of the distribution in which we are interested. For instance, if we put $g(x) = 2x$ in (9), we get $M_0(p)$ the mean of the number of mutants per individual but if we put $g(x) = 2x(1-x)$, we get $H_0(p)$, the mean of the number of heterozygous sites per individual. The statistical properties are functions of the initial frequency p. Here we consider only three statistical properties viz. when $g(x) = 1$, $g(x) = 2x(1-x)$ and $g(x) = s(1-x)$ in relation (9). These give respectively the total number of segregating sites in the population, the mean number of heterozygous sites per individual and the substitutional load in a finite population. These properties can however be obtained directly by putting $g(p)=1$, $g(p)=2p(1-p)$ and $g(p)=s(1-p)$ in (17).

(i) *Total number of segregating sites in the population*

Taking $g(p) = 1$ in (17) and using (29) to (34), we get

$$I_{c0}(p) = [2N_e v_m / S(1-\exp-(2S))][1-\exp(-2Sp))\exp(2Sp)/(1-\exp(-2S(1-p)))\int_p^1 [\{1-\exp(-2S(1-y))\}^2 .\exp(-2Sy)/y(1-y)]\,dy + \int_0^p [1-\exp(-2Sy)).(1-\exp(-2S(1-y)))/y(1-y)]\,dy] \quad \ldots(35)$$

If the mutant is represented only once at the moment of its occurrence, $p = 1/(2N)$, and we have, approximately,

$$I_{c0}^{1}(1/2N) = [2 v_m (N_e/N)/(1-\exp(-2S)(1-(S/N)-\exp(-2S))][(1-\exp(-2S)) - (1-(S/N)-\exp(-2S)) - 2\exp(-2S)\log_e(2N) + \int_{S/N}^{2S}(\exp(-y)/y)\,dy + \exp(-4S)\int_{S/N}^{2S}/\exp(y)/y)\,dy + \exp(-2S)\int_0^{(2S-S/N)}((\exp(y)-1)/y)\,dy - \exp(-2S)\int_0^{(2S-S/N)}((1-\exp(-y))/y)\,dy] \quad \ldots(36)$$

The integrals on the right-hand side of (36) can be evaluated by using

$$\int_{S/N}^{2S}(\exp(-y)/y)\,dy = E_i(S/N) - E_i(2S) \quad \ldots(37)$$

$$\int_{S/N}^{2S}(\exp(y)/y)\,dy = E_i(2S) - E_i(S/N) \quad \ldots(38)$$

$$\int_0^{(2S-S/N)}((\exp(y)-1)/y)\,dy = E_i(2S-S/N) - \log_e(2S-S/N) - \gamma \quad \ldots(39)$$

$$\int_0^{(2S-S/N)}(1-\exp(-y))/y)\,dy = E_i(2S-S/N) + \log_e(2S-S/N) + \gamma \quad \ldots(40)$$

In these relations, γ is Euler's constant and equals 0.57721..., $E_1(.)$ and $E_i(.)$ are exponential integrals defined by

$$E_1(x) = \int_x^\infty (\exp(-y)/y)\,dy = -E_i(-x),\ x>0 \quad \ldots(41)$$

for which fairly extensive tabulations are available in Abramowitz and Stegun (1964). Thus, if the mutant is advantageous, such that $2S = N_e S \gg 1$ but $S/N = (N_e/N)s \ll 1$, we get approximately

$$\overset{1}{I}_{co}(1/2N) \approx 2\nu_m (N_e/N)\,[1-\log_e (S/N)-\gamma] \quad \ldots(42)$$

However, if both $2S$ and (S/N) are much smaller than unity, we get, approximately,

$$\overset{1}{I}_{co}(1/2N) \approx (\nu_m/s)\,[2(1-S)/N - (\gamma+\log_e 2S)/NS] \quad \ldots(43)$$

When the mutant is neutral, $s=0$ and (35) reduces, in the limit, to

$$\overset{1}{I}_{co}(p) = -4N_e \nu_m\,(p/(1-p))\log_e p \quad \ldots(44)$$

For $p=1/2N$, this becomes, approximately for large N,

$$\overset{1}{I}_{co}(1/2N) \approx 2\nu_m (N_e/N)\log_e (2N) \quad \ldots(45)$$

(ii) *Expected number of heterozygous sites per individual*

We now take $g(p)=2p(1-p)$ in (17) for obtaining the mean number of heterozygous nucleotide sites per individual conditional to loss of mutants. Denoting it by $H_0(p)$ and using (29) to (34), we get

$$H_0(p) = (4N_e\nu_m/S)[\{(1+\exp(-2S))/(1-\exp(-2S)\} - (1-p)(\exp(-2Sp) + \exp(-2S)/(\exp(-2Sp) - \exp(-2S))] \quad \ldots(46)$$

The limiting value of $H_0(p)$ when p tends to 1, is found to be

$$\lim_{p\to 1} H_0(p) = 4N_e\nu_m\,[S(1+\exp(-2S)) - (1-\exp(-2S))]/S^2(1-\exp(-2S)) \quad \ldots(47)$$

For neutral mutants ($s=0$), we get the corresponding results as

$$H_0(p) = (4/3)\,N_e\nu_m\,p(2-p) \quad \ldots(48)$$

$$\lim_{p\to 1} H_0(p) = (4/3)\,N_e\nu_m. \quad \ldots(49)$$

In a population consisting of N individuals, if the mutant form in each site is represented only once at the moment of its occurrence, $p=1/(2N)$ and the mean number of heterozygous sites per individual, conditional to loss, becomes

$$H_0(1/2N) \approx (4/3)\,\nu_m(N_e/N) \quad \ldots(50)$$

if the mutant is neutral. However, if the mutant is advantageous, such that $2S \gg 1$ but $(S/N) \ll 1$, we have

$$H_0(1/2N) \approx 2\nu_m\,(N_e/NS) \quad \ldots(51)$$

(iii) *Substitutional load*

For this property, we take $g(p)=s(1-p)$ in (17). Denoting it by $L_0(p)$, we get

$$L_0(p) = [2\nu_m/(1-\exp(-2S))][\{(1-\exp(-2Sp))\exp(2Sp)/(1-\exp(-2S(1-p)))\} \int_p^1 [(1-\exp(-2S(1-y)))^2 \exp(-2Sy)/y]\,dy + \int_0^p \{(1-\exp(-2Sy)).(1-\exp(-2S(1-y)))/y\}\,dy]. \quad \ldots(52)$$

If the mutant form appears once in the population at the time of its occurrence, $p=1/2N$, and $L_0(1/2N)$ becomes

$$L_0(1/2N) = [2\nu_m(S/N)/(1-\exp(-2S))(1-\exp(-2S)-S/N)]\,[(1-\exp(-2S))(1-\exp(-2S)-S/N)-2\exp(-2S)\log_e(2N) + \int_{S/N}^{2S} (\exp(-y)/y)\,dy + \exp(-4S) \int_{S/N}^{2S} (\exp(y)/y)\,dy] \quad \ldots(53)$$

When the mutant is advantageous so that $2S$ is much greater than unity but (S/N) is much smaller than unity, we get

$$L_0(1/2N) \approx 2\nu_m\,(S/N)\,[1-\gamma-\log_e(S/N)] \quad \ldots(54)$$

On the other hand, if both $2S$ and (S/N) are much smaller than unity, we get

$$L_0(1/2N) \approx \nu_m\,[2(1-S)/N-(\gamma+\log_e 2S)/NS] \quad \ldots(55)$$

Discussion

The behaviour of the genetic composition of Mendelian populations over time is determined by the principles of stochastic process.

The mathematical theory of population genetics treats such processes as Markov processes with gene frequency as a random variable subject to the influence of mutation, migration, selection and random sampling of gametes in reproduction. In the context of understanding the mechanics of evolution, this theory could not be very helpful because of the difficulty in relating the gene frequency with the phenotypic level on which the evolutionary data were collected. Fortunately, the recent study of molecular evolution has opened a field in which this theory could be introduced with advantage (Kimura 1971). Most of the studies on mathematical theory of population genetics, in the context of evolution, deal with diffusion models in which gene frequency is treated as a continuous random variable, with time also as continuous. This stands to reason in evolutionary studies because of the 'slow change', of the order of about 0.1 Darwin units (a Darwin unit amounts to a change of $e=2.17$ per million years) and because of the population size, though finite, being considerably large. Diffusion models lean heavily on the forward and backward diffusion equations introduced by Kolmogorov (1931) and used very widely in physics. In this paper, it has been shown how conditioning a diffusion model and making use of conditioned diffusion equations introduced by Narain (1974) could affect the results, particularly the properties of the equilibrium distribution under steady flux of mutations.

In mammals, the total number of nucleotide sites for the haploid chromosome set (T) is estimated to be about 4×10^9. These are sufficient to code for 2×10^6 polypeptides, each consisting of 500 amino acids. If the number of sites for cistron (C) is taken to be about 1,000, the total number of cistrons would be as large as 2×10^6. Let us assume that, in each generation, one advantageous mutant gene appears within the population ($\nu_m=1$) consisting of $N=2\times10^5$ individuals and having effective population size $N_e=10^5$ half as large so that $(N_e/N)=0.5$. This means the mutation rate per gamete $\nu=\nu_m/2N=0.25\times10^{-5}$, whereas the mutation rate per site, denoted by $\mu=\nu/T$, is as small as 0.0625×10^{-14}. The mutation rate per cistron, $U=C\mu$, is then 0.0625×10^{-11}. For $s=0.01$, we get from (42), $I'_{co}(1/2N)\approx5.72$. This estimate is about one-fifth of 28.95, the value we obtain by using the approach of Kimura (1969) and therefore even much smaller than 2×10^6, the total number of cistrons. This justifies the assumption that the total number of sites available for mutation is very much larger than the number of temporarily segregating sites. For neutral mutations, however, we have to take a considerably higher rate of about 2 per gamete per generation. This means $\mu=0.5\times10^{-9}$ and $U=0.5\times10^{-6}$. Now, $\nu_m=2N\nu=8\times10^5$ and, from (45), we get $I'_{co}(1/2N)\approx8N_e\log_e 2N=1.4\times10^7$. This would be a very negligible fraction (0.003) of the total number of segregating sites, and the model could be appropriate if the individual nucleotide site is taken as the unit of mutation.

In regard to the second property (viz, average number of heterozygous nucleotide sites per individual), we get from (50), $H_0(1/2N)\approx5.3\ N_e$ if we assume that molecular mutations are neutral and occur at the rate of 2 per gamete per generation so that $\nu_m=2N\nu=4N$. This means, in a population of effective size as 10^5, the average number of heterozygous nucleotide sites per individual conditional to the ultimate loss of the mutants from the population is about 5.3×10^5. This estimate would be about two-thirds of that obtained by the approach of Kimura (1969). The proportion of heterozygous sites can be obtained by dividing the average number of heterozygous sites by the total number of sites; i.e., $H_0(1/2N)/T=(4/3)(N_e/N)(\nu_m/T)=(8/3)\ N_e\mu$. The probability for a particular site being heterozygous for a selectively neutral mutant, given that the

mutant is destined to be lost, is (8/3) $N_e\mu = 1.33 \times 10^{-4}$ instead of $4N_e\mu = 2 \times 10^{-4}$ on the basis of the unconditional approach of Kimura (1969). The probability that a cistron with 10^3 sites would be heterozygous at one or more sites would then be $1-[1-(8/3)\,N_e\mu]^C \approx 1-\exp[-(8/3)\,N_e\,C\mu] = 1-\exp(-0.133) = 0.1245$ as against the value of 0.1813, which we would get if we follow Kimura (1969)in which sample paths leading to fixation and loss are both taken into account. In either case therefore the conditional approach leads to estimates that are lower than those obtained by the unconditional approach.

For substitutional load, Evens (1972) discussed the conditional case when the favoured allele is eventually fixed. It was pointed out that load obtained by the conditional argument is smaller than that obtained by the unconditional approach used by Kimura and Maruyama (1969). However, as s increases considerably, the two loads become closer. In this paper also, we find the conditional approach giving values smaller than those given by Kimura (1969). However, the conditional load considered relates to the case when the substitutional process is such that the favoured mutant is eventually lost from the population. If we take $p = 1/2N$, $\nu_{in} = 1$ in a population of effective size 10^5 with $(N_e/N) = 0.5$ and $s = 0.01$, we get from (54) $L_0(1/2N) \approx 0.0572$. However, if we take $s = 10^{-6}$ with $N_e = 10^5$ so that $S = N_e s = 0.1$ and $S/N = 0.5 \times 10^{-6}$, we get from (55), $L_0(1/2N) \approx 15 \times 10^{-6}$. Such considerably smaller substitutional loads indicate that there may not be any limit to the rate of gene substitution. It might therefore be desirable to consider conditional substitutional load in determining whether load limits the rate of selectively controlled gene substitutions.

Acknowledgement

The author is grateful to Professors O. Kempthorne and E. Pollak of Iowa State University, Ames and Professor W. J. Ewens of University of Pennsylvania, Philadelphia, USA for valuable discussions.

References

Abramowitz M and Stegun I A (Ed.) 1964 *Handbook of Mathematical Functions with Formulas. Graphs and Mathematical Tables*. Washington U.S. Department of Commerce

Ewens W J 1972 Concepts of substitutional load in finite populations; *Theor. Popu. Biol.* **3** 153

Kimura M 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations; *Genetics* **61** 893

Kimura M 1971 Theoretical foundations of population genetics at the molecular level; *Theor. Popu. Biol.* **2** 174

Kimura M and Maruyama T 1969 The substitutional load in a finite population; *Heredity* **24** 101

Kolmogorov A 1931 Uber die analytischen Methoden in der Wahrscheinlichkeitsr echnung; *Math. Ann.* **104** 415

Maruyama T and Kimura M 1971 Some methods for treating continuous stochastic processes in population genetics; *Jap. J. Genet.* **46** 407

Maruyama T and Kimura M 1975 Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change; *Proc. natn. Acad. Sci.* U.S.A. **72** 1602

Narain P 1974 The conditioned diffusion equation and its use in population genetics; *J.R. Stat. Soc. Ser.* **B36** 258

Selander R K, Yang S Y, Lewontin R C and Johnson W E 1970 Genetic variation in the horseshoe crab (*Limulus polyphomus*) a phylogenetic "relic"; *Evolution* **24** 402

15

THE DESCRIPTION OF GENE ACTION AND INTERACTION WITH MULTIPLE ALLELES IN CONTINUOUS VARIATION

P. NARAIN

Institute of Agricultural Research Statistics, New Delhi, India

Received January 8, 1965

THE description of gene control in characters showing continuous variation was first considered by FISHER (1918). He showed how to represent and estimate the average main and dominance effects of a large number of genes controlling the character under consideration, even when the genes are unequal in effect and show incomplete dominance. His theory was, later on, elaborated in the form of a "Biometrical" method by MATHER (1949) who showed how main and dominance variation can be estimated from various genetic experiments. HAYMAN (1955), however, gave an illuminating device of representing the gene action and interaction by making use of the mathematical representation of laws of genetics. His method gave a general formula for the variance of the family of a cross which leads to the formulae given by MATHER (1949) in particular cases. In all these studies, however, only two alleles at each locus were assumed. If one is dealing with a F_2 derived from two homozygous strains this assumption is quite justified, as only two alleles are then possible at each locus. But when several F_2 populations are considered or when one is dealing with a randomly breeding population, it is necessary to consider more than two alleles at each locus.

The main difficulty in considering more than two alleles at a locus lies in the fact that while with two alleles A_1, A_2 we can represent the three genotypes A_1A_1, A_1A_2 and A_2A_2 by a stochastic variable θ taking values -1, 0 and 1 respectively so that the heterozygote is located at a point intermediate between the two homozygotes, we cannot do so with three alleles or more at a locus. STANTON (1960) suggested that with multiple alleles the homozygotes may be situated at the vertices of a regular simplex whereas the heterozygotes are at the mid-point of the edges. In particular, with three alleles A_1, A_2, A_3 involved at a locus there are six genotypes A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3, A_1A_3 which can be arranged symmetrically by placing the homozygotes at the vertices of an equilateral triangle and the heterozygotes at the midpoint of the edges in a two-dimensional plane. Making use of this fact and representing the six genotypes by a two-dimensional stochastic vector $\theta = (x,y)$, NARAIN (1963) arrived at a mathematical formulation of a law of segregation analogous to that given by HAYMAN (1955) with two alleles at a locus. The genotypic value or the metric of an individual was then expressed as a polynomial function of the variables x and y. In this paper, this scheme is described and the effect of considering multiple alleles in continuous variation is discussed. While the results hold good

for an arbitrary number of alleles at a locus, only three alleles at a locus are assumed to illustrate the underlying principle. Linkage and nonallelic interactions are, however, not taken into account in this discussion.

Representation of genotype: Consider a panmictic population containing three alleles at a given locus. The six genotypes A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3, and A_1A_3 can be represented by the vector variable (X_1, X_2, X_3) where X_i is 0, 1, or 2, and denotes the number of A_i genes in the genotype. All these vectors lie on the plane $X_1 + X_2 + X_3 = 2$. Taking the origin at (1, 1, 0) and making all the vectors lie on the plane $X_3 = 0$, gives two-dimensional vector variables (1, 0), (0,0), (−1, 0), (−1, 1), (−1, 2) and (0, 1) which represent respectively the six genotypes A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3 and A_1A_3. An individual can then be represented by $\theta = (x, y)$ which takes anyone of the six pairs of values depending upon its genotype. It is obvious that x takes the value −1 when A_1 gene is absent, 0 when it is present in single dose and +1 when it is present in double dose. $\frac{1}{2}(1+x)$ will therefore take the values 0, ½ and 1 respectively in the three cases, and these happen to be the corresponding probabilities for the production of an A_1 gamete by the individual. For if the A_1 gene is absent, the genotype will not produce a gamete carrying an A_1 gene, or it can be said to produce gamete A_1 with zero probability. If, however, the A_1 gene is present in single dose, the genotype (being a diploid) will produce a gamete carrying an A_1 gene with probability ½, whereas when the gene is present in double dose, the genotype is certain to produce an A_1 gamete, or in the other words, the probability of its production is unity. The probability that an individual $\theta = (x, y)$ produces a gamete carrying the A_1 gene can, therefore, be represented by $\frac{1}{2}(1+x)$. Similar considerations show that the probabilities for the production of gametes carrying A_2 and A_3 genes can be represented respectively by $\frac{1}{2}(1-z)$ and $\frac{1}{2}y$ where $z = x+y$. When two individuals with genotypes $\theta_1 = (x_1, y_1)$ and $\theta_2 = (x_2, y_2)$ are crossed, the distribution of θ in the offspring can then be obtained as given in Table 1, where $z_1 = x_1+y_1$, $z_2 = x_2+y_2$

In particular, if A_1A_2 and A_2A_3 are crossed, $x_1 = y_1 = 0$, $x_2 = -1$ and $y_2 = 1$ and we obtain from the Table 1 that the resulting generation will have offsprings with genotypes A_1A_2, A_2A_2, A_2A_3 and A_1A_3 each with probability ¼. This is what is expected otherwise from simple genetic principles. The result given in Table 1 can thus be regarded as a mathematical representation of the Mendel's law of

TABLE 1

Distribution of θ in the offspring of the cross $\theta_1 \times \theta_2$

Genotype	Value of θ	Probability
A_1A_1	(1,0)	$\frac{1}{4}[(1+x_1)(1+x_2)]$
A_1A_2	(0,0)	$\frac{1}{4}[(1+x_1)(1-z_2) + (1+x_2)(1-z_1)]$
A_2A_2	(−1,0)	$\frac{1}{4}[(1-z_1)(1-z_2)]$
A_2A_3	(−1,1)	$\frac{1}{4}[(1-z_1)y_2 + (1-z_2)y_1]$
A_3A_3	(−1,2)	$\frac{1}{4}[y_1 y_2]$
A_1A_3	(0,1)	$\frac{1}{4}[(1+x_1)y_2 + (1+x_2)y_1]$

segregation with three alleles. It can easily be generalised to any number of alleles at a locus. For, in a panmictic population containing n alleles $A_1, A_2, \ldots A_n$ at a given locus, there will be $[n(n+1)]/2$ genotypes of form A_iA_j which can be represented by the vector variable $(X_1, X_2, \ldots X_n)$ where X_i is 0, 1 or 2 and denotes the number of A_i genes in the genotype. All these vectors lie on the flat $X_1+X_2+\ldots+X_n=2$. Taking the origin at (1, 1, 0, . . . 0) and making all the vectors lie on the flat $X_2=0$ gives the $(n-1)$-dimensional vector variables of the type $(u_1, u_2, \ldots, u_{n-1})$. Here u_1 takes the values -1, 0 or 1 according as the gene A_1 is absent, present in single dose, or present in double dose whereas (u_1+u_2) takes the values 1, 0 and -1 according as gene A_2 is absent, present in single dose or present in double dose. Also (u_2-u_3), (u_3-u_4), . . . , $(u_{n-2}-u_{n-1})$ take the values 0, 1, or 2 in the three respective phases for the genes $A_3, A_4, \ldots A_{n-1}$ respectively, whereas u_{n-1} takes the values 0, 1, or 2 in the three phases for the gene A_n. By the considerations given above, it can be seen that the individual $\theta=(u_1, u_2, \ldots u_{n-1})$ produces gametes $A_1, A_2, A_3, \ldots A_{n-1}$ and A_n with probabilities $\frac{1}{2}(1+u_1)$, $\frac{1}{2}(1-u_1-u_2)$, $\frac{1}{2}(u_2-u_3)$, . . . $\frac{1}{2}(u_{n-2}-u_{n-1})$ and $\frac{1}{2}u_{n-1}$ respectively. These probabilities give the distribution of θ in the offspring of the cross $\theta_1 \times \theta_2$ which can be regarded as a mathematical representation of the law of segregation with an arbitrary number of alleles at a given locus.

When the two individuals with genotypes θ_1 and θ_2 happen to be members of a random mating population with gene frequencies p, q and r for the three alleles A_1, A_2 and A_3 respectively, it can be seen that the $E(\theta_1)=E(\theta_2)$ i.e. $E(x_1)=E(x_2)=p-q-r$ and $E(y_1)=E(y_2)=2r$ so that in the next generation the six genotypes given in the Table 1 occur with frequencies respectively as

$$\begin{aligned}
\tfrac{1}{4}E[(1+x_1)(1+x_2)] &= p^2 \\
\tfrac{1}{4}E[(1+x_1)(1-z_2)+(1+x_2)(1-z_1)] &= 2pq \\
\tfrac{1}{4}E[(1-z_1)(1-z_2)] &= q^2 \\
\tfrac{1}{4}E[(1-z_1)y_2+(1-z_2)y_1] &= 2qr \\
\tfrac{1}{4}E[y_1y_2] &= r^2 \\
\text{and } \tfrac{1}{4}E[(1+x_1)y_2+(1+x_2)y_1] &= 2pr.
\end{aligned}$$

The gene frequencies in this generation thus turns out to be the same as in the previous generation as is otherwise expected in view of the HARDY-WEINBERG law of equilibrium in random mating populations. This further corroborates the soundness of the proposed representation.

From the distribution given above it is possible to work out the expectations of the stochastic variables x and y and their various powers, as well as of various products between them. These expectations can be used to obtain the variances of x, y, x^2, y^2, xy, and covariances between x and y, x and x^2, x and y^2, x and xy, y and x^2, y and y^2, y and xy, x^2 and y^2, x^2 and xy and y^2 and xy. The results obtained for various expectations, variances and covariances are given by

Expectations

$$\begin{aligned}
E(x) = E(x^3) &= \tfrac{1}{2}(x_1+x_2) \\
E(x^2) = E(x^4) &= \tfrac{1}{2}(1+x_1x_2)
\end{aligned}$$

$$
\begin{aligned}
E(y) &= \tfrac{1}{2}(y_1+y_2) \\
E(y^2) &= \tfrac{1}{2}(y_1+y_2+y_1y_2) \\
E(y^3) &= \tfrac{1}{2}(y_1+y_2+3y_1y_2) \\
E(y^4) &= \tfrac{1}{2}(y_1+y_2+7y_1y_2) \\
-E(xy) = -E(x^3y) &= E(x^2y) = \tfrac{1}{4}\{y_1(1-x_2) + y_2(1-x_1)\} \\
-E(xy^2) = E(x^2y^2) &= \tfrac{1}{4}\{y_1(1-x_2) + y_2(1-x_1) + 2y_1y_2\}
\end{aligned} \tag{1}
$$

Variances

$$
\begin{aligned}
V(x) &= \tfrac{1}{4}(2-x_1^2-x_2^2) \\
V(y) &= \tfrac{1}{4}\{2(y_1+y_2) - y_1^2-y_2^2\} \\
V(x^2) &= \tfrac{1}{4}(1-x_1^2x_2^2) \\
V(y^2) &= \tfrac{1}{4}\{1 - (y_1+y_2+y_1y_2-1)^2 + 12y_1y_2\} \\
V(xy) &= 1/16\{y_1(1-x_2)(4-y_1+y_1x_2) + y_2(1-x_1)(4-y_2+y_2x_1) \\
&\quad + 2y_1y_2(3-x_1x_2+x_1+x_2)\}
\end{aligned} \tag{2}
$$

Covariances

$$
\begin{aligned}
Cov(x, y) &= -\tfrac{1}{4}\{y_1(1+x_1) + y_2(1+x_2)\} \\
Cov(x, x^2) &= \tfrac{1}{4}(x_1+x_2)(1-x_1x_2) \\
Cov(x, y^2) &= -\tfrac{1}{4}\{y_1(1+x_1) + y_2(1+x_2) + y_1y_2(x_1+x_2+2)\} \\
Cov(x, xy) &= \tfrac{1}{8}\{y_1(1-x_2) + y_2(1-x_1)\}(2+x_1+x_2) \\
Cov(x^2, y) &= -\tfrac{1}{4}\{x_1(y_2-x_2) + x_2(y_1-x_1) + x_1x_2(y_1+y_2+2)\} \\
Cov(y, y^2) &= \tfrac{1}{4}\{1 - (y_1+y_2+y_1y_2-1)(y_1+y_2-1) + 5y_1y_2\} \\
Cov(y, xy) &= \tfrac{1}{8}\{(y_1-y_2)^2 - 2(y_1+y_2) - (y_1+y_2-2)(x_2y_1+x_1y_2)\} \\
Cov(x^2, y^2) &= \tfrac{1}{4}\{y_1y_2(1-x_1x_2) - y_1x_2(1+x_1) - y_2x_1(1+x_2)\} \\
Cov(x^2, xy) &= \tfrac{1}{8}\{y_1(1-x_2) + y_2(1-x_1)\}(x_1x_2-1) \\
Cov(y^2, xy) &= \tfrac{1}{8}\{(y_1+y_2+y_1y_2-2)(y_1+y_2-y_1x_2-y_2x_1) - 12y_1y_2\}
\end{aligned} \tag{3}
$$

Now consider k loci, $A, B, C, \ldots K$ each with three alleles. The whole genotype can be represented by a set of k two-dimensional stochastic variables $G = (\theta_a, \theta_b, \ldots \theta_k)$ where $\theta_a = (x_a, y_a)$, $\theta_b = (x_b, y_b)$, $\ldots$, $\theta_k = (x_k, y_k)$. Since we are not taking into account linkage between loci, x_a, y_a can be assumed to be uncorrelated with x_b, y_b etc. so that $Cov(x_a, x_b) = Cov(x_a, y_b) = Cov(x_b, y_a) = Cov(x_b, y_b) = 0$ etc.

Relationship between the metric and the genotype: Consider first a single locus involving three alleles A_1, A_2, A_3 determining the genotype $\theta_a = (x_a, y_a)$. Since the genotype is two-dimensional, the metric $M(\theta_a)$, in accordance with HAYMAN's notations is described as a suitable polynomial function in x_a and y_a given by

$$
M(x_a, y_a) = \lambda_{1a}x_a + \lambda_{2a}y_a + \lambda_{3a}x_a^2 + \lambda_{4a}y_a^2 + \lambda_{5a}x_ay_a \tag{4}
$$

In this case five parameters are needed to describe the metric completely. There are two main effects which may be represented by d_{1a} and d_{2a}. While d_{1a} measures the effect of A_1 relative to A_2, d_{2a} measures that of A_2 relative to A_3. The main effect of A_3 relative to A_1, would obviously be $(d_{1a}-d_{2a})$. As regards dominance deviations, we may represent them by h_{1a}, h_{2a} and h_{3a} respectively measuring the

dominance deviations for the pairs A_1-A_2, A_2-A_3 and A_3-A_1. The relationship between the metrices of the various genotypes and the parameters should, therefore, satisfy the following relations.

$$\begin{aligned}
&\tfrac{1}{2}\{M(1,0) - M(-1,0)\} = d_{1a}\\
&\tfrac{1}{2}\{M(-1,2) - M(-1,0)\} = d_{2a}\\
&\tfrac{1}{2}\{M(1;0) - M(-1,2\} = (d_{1a} - d_{2a})\\
&M(0,0) - \tfrac{1}{2}\{M(1,0) + M(-1,0)\} = h_{1a}\\
&M(-1,1) - \tfrac{1}{2}\{M(-1,0) + M(-1,2)\} = h_{2a}\\
&M(0,1) - \tfrac{1}{2}\{M(-1,2) + M(1,0)\} = h_{3a}
\end{aligned} \tag{5}$$

Consistent with (5) the metric (4) can be shown to take the six values corresponding to the six genotypes as given in Table 2.

The coefficients in the polynomial (4) when expressed in terms of the five parameters are given by

$$\begin{aligned}
\lambda_{1a} &= d_{1a}\\
\lambda_{2a} &= d_{2a} - h_{1a} + h_{2a} + h_{3a}\\
\lambda_{3a} &= -h_{1a}\\
\lambda_{4a} &= -h_{2a}\\
\lambda_{5a} &= -h_{1a} - h_{2a} + h_{3a}
\end{aligned} \tag{6}$$

When k loci are considered simultaneously the metric is given by

$$\sum_a [\lambda_{1a}x_a + \lambda_{2a}\gamma_a + \lambda_{3a}x_a{}^2 + \lambda_{4a}\gamma_a{}^2 + \lambda_{5a}x_a\gamma_a] \tag{7}$$

where nonallelic interactions are assumed to be absent.

The relationship between the metric and the genotype given above for the case of three alleles at each locus can be generalised to any number of alleles at each of the loci. Consider first a single locus involving n alleles $A_1, A_2 \ldots A_n$ determining the genotype $\theta_a = (u_{1a}, u_{2a}, \ldots u_{(n-1),a})$. The metric $M(\theta_a)$ is now described as a suitable polynominal function in $(n-1)$ variables $u_{1a}, u_{2a}, \ldots u_{(n-1),a}$ given by

$$M(u_{1a}, u_{2a}, \ldots u_{(n-1),a}) = \sum_{j=1}^{(n-1)} \lambda_{ja}u_{ja} + \sum_{j=1}^{(n-1)} \lambda_{2ja}u^2{}_{ja} + \sum_{i=1}^{(n-2)} \sum_{j>i}^{(n-1)} \lambda_{(i+2),(j-i),a}u_{ia}u_{ja} \tag{4a}$$

In this case $(n-1)(n+2)/2$ parameters are needed to describe the metric com-

TABLE 2

Value of the metric corresponding to different genotypes

Genotype	(x_a, γ_a)	Metric $M(x_a, \gamma_a)$ or the genotypic value
A_1A_1	(1,0)	$d_{1a} - h_{1a}$
A_1A_2	(0,0)	0
A_2A_2	(−1,0)	$-d_{1a} - h_{1a}$
A_2A_3	(−1,1)	$-d_{1a} - h_{1a} + d_{2a} + h_{2a}$
A_3A_3	(−1,2)	$-d_{1a} - h_{1a} + 2d_{2a}$
A_1A_3	(0,1)	$-h_{1a} + d_{2a} + h_{3a}$

pletely. There are now $(n-1)$ main effects and $n(n-1)/2$ dominance deviations which make up the total of $(n-1)(n+2)/2$. The coefficients λ's in (4a) can be expressed as linear functions of these main effects and dominance deviations. When k loci are considered simultaneously the metric is given by

$$\sum_a \sum_j \lambda_{1ja} u_{ja} + \sum_a \sum_j \lambda_{2ja} u^2_{ja} + \sum_a \sum_i \sum_{j>i} \lambda_{(i+2),(j-i),a} u_{ja} u_{ia} \qquad (7a)$$

where nonallelic interactions are assumed to be absent.

Mean and variance of the offspring of the cross $\theta_1 \times \theta_2$: The mean of the offspring generation resulting from a cross $\theta_1 \times \theta_2$ can be obtained by finding the expectation of the metric $M(x, y)$ with the help of (1). Thus

$$\begin{aligned} E\{M(x,y)\} = {} & \tfrac{1}{2}\lambda_1(x_1+x_2) + \tfrac{1}{2}\lambda_3(1+x_1x_2) \\ & + \tfrac{1}{4}\{2(\lambda_2+\lambda_4) - \lambda_5\}(y_1+y_2) \\ & + \tfrac{1}{2}\lambda_4 y_1y_2 + \tfrac{1}{4}\lambda_5(y_1x_2+y_2x_1) \end{aligned} \qquad (8)$$

The suffix "a" is dropped for simplicity, it being understood that we are considering only a single locus. It is interesting to note the difference between the mean given by (8) and the parental mean given by.

$$\begin{aligned} \tfrac{1}{2}\{M(x_1,y_1) + M(x_2,y_2)\} = {} & \tfrac{1}{2}\lambda_1(x_1+x_2) + \tfrac{1}{2}\lambda_2(y_1+y_2) \\ & + \tfrac{1}{2}\lambda_3(x^2_1+x^2_2) + \tfrac{1}{2}\lambda_4(y^2_1+y^2_2) \\ & + \tfrac{1}{2}\lambda_5(x_1y_1+x_2y_2) \end{aligned} \qquad (9)$$

This difference is given by

$$\begin{aligned} & + \tfrac{1}{4}h_1[2\{(1+x_1x_2)-x^2_1(1-y_1)-x^2_2(1-y_2)\}-\{y_1(1-x_2)+y_2(1-x_1)\}] \\ & + \tfrac{1}{4}h_2[2y_1y_2+y_1(1+2x^2_1)+y_2(1+2x^2_2)+y_1(x_2-2y_1)+y_2(x_1-2y_2)] \\ & + \tfrac{1}{4}h_3[y_1(1-2x^2_1)+y_2(1-2x^2_2)-(y_1x_2+y_2x_1)] \end{aligned} \qquad (10)$$

It can be seen that the difference given by (10) clearly depends on the three dominance deviations and the component variables of the genotype of the parents. That this expression would reduce to the ones expected in particular cases is demonstrated in Table 3.

The variance of the offspring generation from the cross $\theta_1 \times \theta_2$ is obtained by finding the variance of the metric $M(x,y)$ with the help of (2), (3) and (6). The resulting expression is given by

TABLE 3

Differences in the mid-parent and offspring means in various crosses

Type of cross	Value of x_1	y_1	x_2	y_2	Difference
Selfing or intercrossing					
$A_1A_2 \times A_1A_2$	0	0	0	0	$\frac{1}{2}h_1$
$A_2A_3 \times A_2A_3$	−1	1	−1	1	$\frac{1}{2}h_2$
$A_1A_3 \times A_1A_3$	0	1	0	1	$\frac{1}{2}h_3$
Backcrossing					
$A_1A_2 \times A_1A_1$	0	0	1	0	0
$A_1A_2 \times A_2A_2$	0	0	−1	0	0
$A_2A_3 \times A_2A_2$	−1	1	−1	0	0
$A_2A_3 \times A_3A_3$	−1	1	−1	2	0
$A_1A_3 \times A_3A_3$	0	1	−1	2	0
$A_1A_3 \times A_1A_1$	0	1	1	0	0

$$
\begin{aligned}
V\{M(x,y)\} = {} & \tfrac{1}{4}\, d^2_1\{2(1+x_1x_2)-(x_1+x_2)^2\} \\
& + \tfrac{1}{4}\, d^2_2\{(1+2y_1y_2)-(y_1+y_2-1)^2\} \\
& + 1/16\, h^2_1\{4-[(y_1+y_2)+(z_1x_2+z_2x_1)]^2\} \\
& + 1/16\, h^2_2\{(y_1+y_2)-(z_1y_2+z_2y_1)\}\{4-(y_1+y_2)+(z_1y_2+z_2y_1)\} \\
& + 1/16\, h^2_3\{y_1(1+x_2)+y_2(1+x_1)\}\{4-y_1(1+x_2)-y_2(1+x_1)\} \\
& - \tfrac{1}{2}\, d_1d_2\{y_1(1+x_1)+y_2(1+x_2)\} \\
& + \tfrac{1}{4}\, d_1h_1(x_1+x_2)\{(z_1x_2+z_2x_1)+(y_1+y_2-2)\} \\
& + \tfrac{1}{4}\, d_1h_2(x_1+x_2+2)\{(z_1y_2+z_2y_1)-(y_1+y_2)\} \\
& - \tfrac{1}{4}\, d_1h_3(x_1+x_2)\{y_1(1+x_2)+y_2(1+x_1)\} \\
& + \tfrac{1}{4}\, d_2h_1(y_1+y_2)\{(z_1x_2+z_2x_1)+(y_1+y_2-2)\} \\
& + \tfrac{1}{4}\, d_2h_2(y_1+y_2-2)\{(z_1y_2+z_2y_1)-(y_1+y_2)\} \\
& - \tfrac{1}{4}\, d_2h_3(y_1+y_2-2)\{y_1(1+x_2)+y_2(1+x_1)\} \\
& + \tfrac{1}{8}\, h_1h_2\{(z_1y_2+z_2y_1)-(y_1+y_2)\}\{(z_1x_2+z_2x_1)+(y_1+y_2-2)\} \\
& + \tfrac{1}{8}\, h_1h_3\{y_1(1+x_2)+y_2(1+x_1)\}\{(z_1x_2+z_2x_1)+(y_1+y_2-2)\} \\
& + \tfrac{1}{8}\, h_2h_3\{y_1(1+x_2)+y_2(1+x_1)\}\{(z_1y_2+z_2y_1)-(y_1+y_2)\} \qquad (11)
\end{aligned}
$$

In particular, the variances of the families obtained by selfing the heterozygotes can be obtained by substituting the particular set of values of x_1, y_1, x_2, y_2 in (11). The expressions obtained are in perfect agreement with those which can be obtained from first principles. This is demonstrated in Table 4.

When k loci are considered simultaneously and linkage and epistasis are assumed to be absent, the mean and variance of the offspring generation from the cross $G_1 \times G_2$ are given by $\sum_a E\{M(x_{1a},y_{1a})\}$ and $\sum_a V\{M(x_{1a},y_{1a})\}$ respectively.

Randomly breeding populations: Consider a random mating population with gene frequencies p_a, q_a, r_a for the alleles A_1, A_2, A_3 respectively at a given locus. With the help of metric values given in Table 2 and the genotypic frequencies $(p^2_a, 2p_aq_a, q^2_a, 2q_ar_a, r^2_a, 2p_ar_a)$ the contribution of this locus to the average of the random mating population is given by

$$d_a = (p_a-q_a-r_a)d_{1a} + 2r_ad_{2a} - (1-2p_aq_a)h_{1a} + 2q_ar_ah_{2a} + 2p_ar_ah_{3a} \qquad (12)$$

The metric values can then be represented, alternatively, in terms of d_a and other quantities leading to symmetric expressions as given in Table 5.

The average of the random mating population is $\sum_a d_a$ where d_a is given by (12).

The contribution of this locus to the genetic variance of the random mating population is given by

$$
\begin{aligned}
& p^2_a(d_{1a}-h_{1a})^2+q^2_a(d_{1a}+h_{1a})^2+2q_ar_a(d_{1a}-d_{2a}+h_{1a}-h_{2a})^2 \\
& \quad +r^2_a(d_{1a}-2d_{2a}+h_{1a})^2+2p_ar_a(d_{2a}-h_{1a}+h_{3a})^2-d^2_a \qquad (13)
\end{aligned}
$$

TABLE 4

Variances in various types of selfed families

Selfing of	Value of x_1	y_1	x_2	y_2	$V\{M(x,y)\}$
A_1A_2	0	0	0	0	$\frac{1}{2}\, d_1^2 + \frac{1}{4}\, h_1^2$
A_2A_3	−1	1	−1	1	$\frac{1}{2}\, d_2^2 + \frac{1}{4}\, h_2^2$
A_1A_3	0	1	0	1	$\frac{1}{2}(d_1-d_2)^2 + \frac{1}{4}\, h_3^2$

TABLE 5

Metric values of various genotypes in a random mating population

Genotype	(x, y)	$M(x, y)$
A_1A_1	(1,0)	$d_a+2(1-p_a)d_{1a}-2r_ad_{2a}-2p_aq_ah_{1a}-2q_ar_ah_{2a}-2p_ar_ah_{3a}$
A_1A_2	(0,0)	$d_a+(1-2p_a)d_{1a}-2r_ad_{2a}+(1-2p_aq_a)h_{1a}-2q_ar_ah_{2a}-2p_ar_ah_{3a}$
A_2A_2	(−1,0)	$d_a-2p_ad_{1a}-2r_ad_{2a}-2p_aq_ah_{1a}-2q_ar_ah_{2a}-2p_ar_ah_{3a}$
A_2A_3	(−1,1)	$d_a-2p_ad_{1a}+(1-2r_a)d_{2a}-2p_aq_ah_{1a}+(1-2q_ar_a)h_{2a}-2p_ar_ah_{3a}$
A_3A_3	(−1,2)	$d_a-2p_ad_{1a}+2(1-r)d_{2a}-2p_aq_ah_{1a}-2q_ar_ah_{2a}-2p_ar_ah_{3a}$
A_1A_3	(0,1)	$d_a+(1-2p_a)d_{1a}+(1-2r_a)d_{2a}-2p_aq_ah_{1a}-2q_ar_ah_{2a}+(1-2p_ar_a)h_{3a}$

When the genes are independent in action and uncorrelated in distribution, the total genetic variation can be obtained by summing contributions given by (13) over all loci. This can finally be given by

$$_nV_R = \tfrac{1}{2}D + \tfrac{1}{4}H \qquad (14)$$

where

$$D = \sum_a [4p_aq_a\{d_{1a}+(2q_a-1)h_{1a}+r_a(h_{1a}-h_{2a}+h_{3a})\}^2 + 4q_ar_a\{d_{2a}+(2q_a-1)h_{2a}+p_a(h_{2a}-h_{1a}+h_{3a})\}^2 + 4r_ap_a\{(d_{1a}-d_{2a})+(2r_a-1)h_{3a}+q_a(h_{1a}-h_{2a}+h_{3a})\}^2]$$

and

$$H = \sum_a [16\{p_aq_ah_{1a}+q_ar_ah_{2a}+r_ap_ah_{3a}\}^2 + 8p_aq_ar_a\{(h_{1a}-h_{2a})^2+h_{3a}(h_{3a}-2h_{1a}-2h_{2a})\}]$$

It is interesting to compare the definitions of D and H given under (14) with those given by MATHER (1949, p. 75) for randomly breeding populations. The definitions are the same as given by MATHER (1949) when either $p_a=0$, or $q_a=0$ or $r_a=0$ i.e. when the third allele at each locus is assumed to be nonexistent. When the two main effects d_{1a} and d_{2a} are identical, and equal to m_a say, the third main effect is zero. Also let $h_{1a}=h_{2a}=h_a$ and h_{3a} be zero. With these conditions, the definitions would become

$$D = \sum_a 4q_a(p_a+r_a)\{m_a+(2q_a-1)h_a\}^2$$

$$\text{and} \quad H = \sum_a 16q^2{}_a(p_a+r_a)^2h^2{}_a \qquad (15)$$

In such a case, therefore, the three alleles at each locus can be regarded as a two-allele system with allelic frequencies as q_a and (p_a+r_a).

The offsprings resulting from a particular cross $\theta_1 \times \theta_2$ in the random mating population form a family of full-sibs. The mean and variance of such a family are given by (8) and (11) respectively. If we take expectation of (8) and substitute $E(x_{1a}) = E(x_{2a}) = (p_a-q_a-r_a)$ and $E(y_{1a}) = E(y_{2a}) = 2r_a$ we get the contribution of the locus to the mean of the offspring generation which is found to be the same as the contribution to the mean of the parental generation given by (12). If we take the expectation of (11) and substitute $E(x_{1a}) = E(x_{2a}) = (p_a-q_a-r_a)$, $E(y_{1a}) = E(y_{2a}) = 2r_a$, $E(x^2{}_{1a}) = E(x^2{}_{2a}) = p^2{}_a+(q_a+r_a)^2$, $E(y^2{}_{1a}) = E(y^2{}_{2a}) = 2r_a(1+r_a)$ and $E(x_{1a}y_{1a}) = E(x_{2a}y_{2a}) = -2r_a(1-p_a)$ we would get the contri-

bution of the locus to the mean variance of full-sib families. The variance of (8) after necessary substitutions would give the contribution to covariance between full-sibs. If expectation of (8) is taken over either $\theta_1=(x_1, y_1)$ or $\theta_2=(x_2, y_2)$ the contribution to the mean of a particular family of half-sibs is obtained. For instance, the contribution of the locus to the mean of half-sib family with parent as θ_2 is given by

$$\{\tfrac{1}{2}\lambda_{1a}(2p_a-1)+(\lambda_{2a}+\lambda_{4a}-\tfrac{1}{2}\lambda_{5a})r_a+\tfrac{1}{2}\lambda_{3a}\}+\tfrac{1}{2}\{\lambda_{1a}+\lambda_{3a}(2p_a-1)+\lambda_{5a}r_a\}x_{2a} + \tfrac{1}{2}\{\lambda_{2a}+\lambda_{4a}(1+2r_a)-\lambda_{5a}(1-p_a)\}y_{2a} \quad (16)$$

Variance of (16) would give the contribution to the covariance between half-sibs.

Effect of multiple allelism in continuous variation: It is well known that the continuous variation observed in the phenotype arises from the discontinuous variation of the genotype when the genes have effects similar to one another, supplementing each other and small in relation to environmental variation. If there are two alleles at each locus with frequencies p and q and there are k loci, the frequency distribution of the phenotypes ignoring environmental effects and assuming no dominance would be described by the various terms in the binomial $(q+p)^{2k}$ which tends to a continuous distribution when k is quite large. When, however, there are more than two alleles at each locus, say for instance three alleles with frequencies p, q, and r, the corresponding distribution is given by the multinomial $(q+p+r)^{2k}$. Now when k tends to become very large, a continuous distribution in two dimensions is approximated. It is, therefore, expected that with multiple allelism, contribution of continuous variation from each dimension is to be added over for the realization of the total variation. When such a contribution is of the same order, whatever may be the dimension under consideration, the existing partition of the total variation is not going to be affected. The case of three alleles, or in fact more than three, can be looked upon as a case of two alleles by dividing the number of alleles into two groups, one consisting of one allele and the other group containing the rest of them and treating the two groups as a two-allele system. Such a situation gives rise to the definitions of D and H as under (15). It, therefore, appears that the nonequality of d_{1a} and d_{2a}, h_{1a} and h_{2a} and the nonvanishing of h_{3a} may reveal some component of variation contributing towards the total variation. Such a component, of course, expected to be existing only in a population where several alleles are operating at each locus, may be described as due to multiple allelism. It may also be looked upon in a manner similar to that of dominance. Dominance is regarded as due to the differences in the effect of gene substitution when the other allele is changed. In a similar way, in a population where several alleles are operating at each locus the effects of gene substitution with respect to any fixed particular allele can be considered. The differences in the effects of gene substitution can be regarded as effects due to multiple allelism. When the effects of gene substitution with respect to the fixed allele are the same, we can say that there is no effect due to the multiple alleles, though there may be several of them operating.

In the first instance, we may regard dominance effects as absent so that h_{1a}, h_{2a} and h_{3a} are zero and the D in (14) reduces to

$$D=\sum_a [4p_aq_ad^2{}_{1a}+4q_ar_ad^2{}_{2a}+4r_ap_a(d_{1a}-d_{2a})^2] \quad (17)$$

Let $d_{1a}=m_a+\alpha_{1a}$ and $d_{2a}=m_a+\alpha_{2a}$, when m_a is the mean of d_{1a} and d_{2a} given by

$$m_a = \frac{p_a d_{1a} + r_a d_{2a}}{p_a + r_a} \tag{18}$$

Here allele A_2 is taken as fixed and A_1 and A_3 with frequencies p_a and r_a are considered with respect to this allele. Since in view of (18), $p_a\alpha_{1a}+r_a\alpha_{2a}=0$, D can be regarded as consisting of two components

$$D_A = \sum_a 4q_a(1-q_a)m^2{}_a$$

$$\text{and} \quad D_B = \sum_a 4(p_a\alpha^2{}_{1a}+r_a\alpha^2{}_{2a}) \tag{19}$$

When there are more than three alleles, say four alleles it is expected that in (19) D_A would still have the same definition but D_B would be given by

$$D_B = \sum_a 4\{p_a\alpha^2{}_{1a} + r_a\alpha^2{}_{2a} + s_a\alpha^2{}_{3a}\} \tag{20}$$

where $q_a+p_a+r_a+s_a=1$, and α_{3a} corresponds to the third independent main effect. Thus it appears that D_B would reflect the contribution of multiple allelic effect to the additive genetic variance.

When there is dominance also present, the nonequality of the dominance effects may be taken into account in the same manner by letting $h_{1a}=h_a+\beta_{1a}$ and $h_{2a}=h_a+\beta_{2a}$ where h_a is given by

$$h_a = \frac{p_a h_{1a} + r_a h_{2a}}{p_a + r_a} \tag{21}$$

Again, since in view of (21), $p_a\beta_{1a} + r_a\beta_{2a}=0$, it can be shown that D and H of (14) take the form

$$\begin{aligned} D = {} & \sum_a [4q_a(1-q_a)\{m_a+(2q_a-1)h_a+\frac{2r_ap_a}{1-q_a}h_{3a}\}^2] \\ & + \sum_a [4p_a\{\alpha_{1a}+q_a\beta_{1a}+r_ah_{3a}\}^2+4r_a\{\alpha_{2a}+q_a\beta_{2a}+p_ah_{3a}\}^2] \\ & - \sum_a \left[\frac{4r^2{}_ap^2{}_a}{1-q_a}\, h^2{}_{3a}\right] \end{aligned} \tag{22}$$

$$\begin{aligned} H = {} & \sum_a [16q^2{}_a(1-q_a)^2\{h_a-\frac{r_ap_a}{(1-q_a)^2}h_{3a}\}^2] \\ & + \sum_a [8q_a(1-q_a)\{p_a(\beta_{1a}-\frac{r_a}{1-q_a}h_{3a})^2+r_a(\beta_{2a}-\frac{p_a}{1-q_a}h_{3a})^2\}] \\ & + \sum_a [\frac{16r^2{}_ap^2{}_a(1-2q_a)}{(1-q_a)^2}h^2{}_{3a}] \end{aligned}$$

Thus the genetic variance of a randomly breeding population can be expressed as

$$_HV_R = \tfrac{1}{2}\, D_A + \tfrac{1}{4}\, H_A + \tfrac{1}{2}\, (D_B+H_B+H_C) \tag{23}$$

where

$$D_A = \sum_a [4q_a(1-q_a)\{m_a+(2q_a-1)h_a+\frac{2r_ap_a}{1-q_a}h_{3a}\}^2]$$

$$D_B = \sum_a [4p_a\{\alpha_{1a}+q_a\beta_{1a}+r_ah_{3a}\}^2+4r_a\{\alpha_{2a}+q_a\beta_{2a}+p_ah_{3a}\}^2]$$

$$H_A = \sum_a [16q_a{}^2(1-q_a)^2\{h_a-\frac{r_ap_a}{(1-q_a)^2}h_{3a}\}^2]$$

$$H_B = \sum_a [4q_a(1-q_a)\{p_a(\beta_{1a} - \frac{r_a}{1-q_a} h_{3a})^2 + r_a(\beta_{2a} - \frac{p_a}{1-q_a} h_{3a})^2\}]$$

and

$$H_C = \sum_a [\frac{4r_a^2 p_a^2(1-3q_a)}{(1-q_a)^2} h_{3a}^2]$$

In (23), D_A and H_A correspond to the components due to additive genetic and dominance effects. D_A, however, includes some effect of the mean dominance h_a taken over the A_1 and A_3 and also of the dominance h_{3a} operating between A_1 and A_3. Similarly H_A includes some effect of the h_{3a}. The remaining three components D_B, H_B and H_C can be regarded as components corresponding to the multiple allelic effects. While D_B includes some effect of the nonequality of the two dominance effects and also of the third dominance effect h_{3a}, H_B includes some effect of the h_{3a} only. The multiple allelic effects would not contribute towards the total variation whenever $d_{1a} = d_{2a}$, $h_{1a} = h_{2a}$ and $h_{3a} = 0$ which correspond to the simultaneously vanishing of D_B, H_B and H_C.

The above investigation shows that in addition to the hitherto described two components of continuous variation in the literature *viz.* additive and dominance, there may be a necessity of describing a third multiple allelic component in genetic populations where several alleles are required to be considered at each of the loci. How this new component having three subcomponents D_B, H_B and H_C would behave in the constitution of various second degree statistics which are utilized for the estimation purposes, remains to be seen in future investigations.

SUMMARY

The effect of three alleles at each locus in continuous variation has been considered. HAYMAN's (1955) device of representing the gene action and interaction by making use of the mathematical representation of laws of genetics can be extended to multiple allelic systems. In a randomly breeding population where several alleles are supposed to be operating at each loci, a new component of genetic variation due to multiple allelic effects is described. This component consists of three subcomponents, and the multiple allelism can be regarded as contributing nothing to the total genetic variation when the three subcomponents vanish simultaneously.

LITERATURE CITED

FISHER, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edin. **52**: 399–433.

HAYMAN, B. I., 1955 Description and analysis of gene action and interaction. Cold Spring Harbor Symp. Quant. Biol. **20**: 79–86.

MATHER, K., 1949 *Biometrical Genetics*. Metheun, London.

NARAIN, P., 1963 On mathematical representation of gene action and interaction. J. Indian Soc. Agri. Stat. **15**: 270.

STANTON, R. G., 1960 Genetic correlations with multiple alleles. Biometrics **16**: 235–244.

Bhubaneswar 1978-79, 71-79.

16

CERTAIN GENERALISATIONS IN RESPONSE TO GENETIC SELECTION

P. Narain

Indian Agricultural Statistics Research Institute, New Delhi.

1. INTRODUCTION

In studies on inheritance of characters exhibiting continuous variation it is usually postulated that a character x is under the control of a large number of genes known as polygenes, each of which has a small effect in comparison with non-heritable factors, the effect being similar but cumulative for the set of genes involved. The effect of a set of such genes confers a genotypic value (G_x) on the individual who is observed for the character x, the observation being designated as the phenotypic value (P_x) of the individual. The difference between P_x and G_x gives environmental deviation (E_x) for the individual on the assumption that there are no interactions between the genotype and the environment. The genotypic value (G_x) is further capable of subdivision into values due to additive genetic effect (A_x) (also called the breeding value), dominance (D_x) and epistatic (I_x) deviaions. The relationship between P_x and A_x is expressed in terms of regression of A_x on P_x ($b_{A_xP_x}$) which in view of A_x being a part of P_x, expresses the fraction of phenotypic variance, $\sigma^2_{P_x}$ due to additive genetic effects. This fraction is known as heritability. If it is assumed that (A_x, P_x), are jointly distributed in a bivariate normal form, the regression of A_x on P_x is necessarily linear and the expectation of A_x for a given individual is $h^2_x P_x$ when the means of the population of indiviauals and their breeding values for the characters are assumed to be zero. In a genetic selection programme, however, superior individuals having phenotypic values greater than or equal to a given value x_o are selected. The expected response to selection is then h^2_x times the mean of the population truncated at x_o which, on the assumption of the normality of the distribution, is $ih^2_x \sigma_{P_x}$ where i, the intensity of selection, is Z/P, Z being the ordinate at the point of truncation for a standard normal distribution with zero mean and unit standard deviation and P is the proportion of population selected. The object of this paper is to present some of the generalised concepts of the relationship between breeding values and phenotypic values and expected response to selection when several correlated characters are considered.

2. COMPONENTS OF DISPERSION

Let the phenotypic measurements ($P_1, P_2, \ldots, P_k$) of an individual in respect of k quantitative characters be expressed in the form of a $k \times 1$ column vector **P** whereas the corresponding conceptual breeding values of the individual ($A_1, A_2, \ldots, A_k$) be expressed by a $k \times 1$ column vector **A**. The non-additive and environmetal effects taken together and represented by ($R_1, R_2, \ldots, R_k$) may be expressed in terms of a $k \times 1$ column vector **R**. The basic vector equation on the assumption of no interaction between genotype and environment is, then given by

$$(2.1) \quad \mathbf{P} = \mathbf{A} + \mathbf{R}$$

where the variables are expressed as deviations from the mean and standardized to have variances as unity and covariances as equal to correlations, so that the correlation matrix of order kxk of the variables **P** is

$$\underset{P}{\Sigma} = \{ \rho_{ij} \} \text{ with } \rho_{ii} = 1 \text{ and } \rho_{ij} = \rho_{ji}, \quad (i, j = 1, 2, \ldots, k).$$

Further let the variance-covariance matrices, each of order kxk, of the variables vectors **A** and **R** be respectively $\underset{A}{\Sigma} = \{ h_{ij} \}$ and $\underset{R}{\Sigma} = \{ e_{ij} \}$ with $h_{ij} = h_{ji}$ and $e_{ij} = e_{ji}$, $(i, j = 1, 2, \ldots, k)$. The h_{ii}'s are the heritabilities of the characters and h_{ij}'s are the covariances of breeding values between the characters. If the genetic correlation between i^{th} and j^{th} character is denoted by r_{ij}, we can express h_{ij} as $r_{ij} \sqrt{h_{ii} h_{jj}}$. Also in view of (2.1),

$$(2.2) \quad \text{Cov.} (\mathbf{A}, \mathbf{P}) = E (\mathbf{A P'}) = \underset{A}{\Sigma}$$

since $\text{Cov.} (A_i P_j) = \text{Cov.} (P_i A_j) = \text{Cov.} (A_i A_j) = h_{ij}$ and $\text{Cov.} (A_i P_i) = \text{Var.} (A_i) = h_{ii}$ $(i, j = 1, 2, \ldots, k)$

From (1), under the assumptions already stated,

$$(2.3) \quad \underset{P}{\Sigma} = \underset{A}{\Sigma} + \underset{R}{\Sigma}$$

so that if the symmetric matrix $\underset{P}{\Sigma}$ is a non-singular one, $| \underset{P}{\Sigma} |$ $\neq 0$ and we have

$$(2.4)\quad I = H + E$$

where

$$(2.5)\quad \begin{cases} H = \underset{A}{\Sigma}\ \underset{P}{\Sigma}^{-1} \\ E = \underset{R}{\Sigma}\ \underset{P}{\Sigma}^{-1} \end{cases}$$

3. CORRELATION BETWEEN BREEDING VECTOR AND PHENOTYPIC VECTOR

In the univariate case, the correlation coefficient between the breeding value and the phenotypic value is square-root of the heritability since the covariance between the two values becomes variance of the breeding value in view of additivity of the breeding value and environmental deviation to produce the phenotypic value. When we consider several correlated characters, both at the genetic as well as at phenotypic levels, the simple concept of correlation between two variables is to be replaced by the concept of correlation between two vector valued variables **A** and **P**. This problem for the set of k characters can be tackled by setting up linear functions

$$(3.1)\quad \begin{cases} A = \sum_{i=1}^{k} a_i A_{xi} = \mathbf{a}' \mathbf{A} \\ P = \sum_{i=1}^{k} a_i P_{xi} = \mathbf{a}' \mathbf{P} \end{cases}$$

and choosing coefficients $\mathbf{a}' = (a_1, a_2, \ldots, a_k)$ such that the correlation between **A** and **P** given by

$$(3.2)\quad h(a) = \sqrt{\frac{\mathbf{a}' \underset{A}{\Sigma}\ \mathbf{a}}{\mathbf{a}' \underset{P}{\Sigma}\ \mathbf{a}}}$$

is maximum. It may be noted that the same coefficients are used in both the linear functions since there is a natural correspondence between the components of **A** and **P**. In view of (2.1), for each i.

$$(3.3)\quad P_{xi} = A_{xi} + R_{xi}$$

so that

$$(3.4)\quad \sum_i a_i P_{xi} = \sum_i a_i A_{xi} + \sum_i a_i R_{xi}$$

giving a relation for the compounded single character as

$$(3.5)\quad P = A + R$$

Taking logarithms in (3.2), differentiating with respect to the elements of **a**, setting the vectors of the derivatives equal to zero, multiplying by $\mathbf{a}' \Sigma \mathbf{a}$ and rearranging we get,

$$(3.6)\quad \left[\sum_A - h^2(a) \sum_P \right] \mathbf{a} = 0$$

In order that there be a non-trivial solution, the matrix on the left must be singular; that is

$$(3.7)\quad \left| \sum_A - h^2(a) \sum_P \right| = 0$$

in view of (2.5), this becomes

$$(3.8)\quad | H - h^2(a) I | = 0$$

This means the desired coefficients are eigenvectors corresponding to eigenvalues of the matrix

$H = \left(\sum_A \sum_P^{-1} \right)$. We get a set of canonical correlations

$$(3.9)\quad h(a_1) \geq h(a_2) \geq \ldots \geq h(a_k)$$

and corresponding pairs of canonical variates

$$(3.10)\quad (\mathbf{a_1}' A, \mathbf{a_1}' P), (\mathbf{a_2}' A, \mathbf{a_2}' P), \ldots, (\mathbf{a_k}' A, \mathbf{a_k}' P)$$

It may be noted that for $i \neq j$,

$$(3.11)\quad \text{Cov.}(\mathbf{a_i}' A, \mathbf{a_j}' A) = \text{Cov.}(\mathbf{a_i}' P, \mathbf{a_j}' P) = 0,$$

but Cov $(\mathbf{a_i}' A, \mathbf{a_i}' P)$ and Cov $(\mathbf{a_j}' A, \mathbf{a_i}' P)$ are not necessarily zero. Thus, we can replace the multiple correlated breeding and phenotypic vectors by $\mathbf{a_1}' A$ and $\mathbf{a_1}' P$ respectively which have the maximum correlation $h(a_1)$. The square of this correlation, $h^2(\mathbf{a_1})$ expresses the fraction of the phenotypic

variance of the compounded characters P which is due to additive genetic effects and can therefore be regarded as a generalised concept of heritability. We can thus term the largest eigenroot of H as the 'generalised heritability'.

For $k = 1$, (3.8) reduces to $(h_{11} - h^2) = 0$, giving the only root as h_{11} which is the usual heritability in the univariate case. For $k = 2$, however, (3.8) reduces to

$$(3.12)\quad \begin{vmatrix} \left[\frac{h_{11} - \rho_{12} h_{12}}{1 - \rho^2_{12}} - h^2(a)\right] & \left[\frac{h_{12} - \rho_{12} h_{11}}{1 - \rho^2_{12}}\right] \\ \left[\frac{h_{12} - \rho_{12} h_{22}}{1 - \rho^2_{12}}\right] & \left[\frac{h_{22} - \rho_{12} h_{12}}{1 - \rho^2_{12}} - h^2(a)\right] \end{vmatrix} = 0$$

The resulting quadratic equation gives two roots, the largest one being,

$$(3.13)\quad h^2(a_1) = \frac{(h_{11} + h_{22} - 2\rho_{12}h_{12}) + \sqrt{(h_{11} - h_{22})^2 + 4(h_{12} - h_{11}\rho_{12})(h_{12} - h_{22}\rho_{12})}}{2(1 - \rho^2_{12})}$$

When $h_{12} = 0$, we get

$$(3.14)\quad h^2(a_1) = [(h_{11} + h_{22}) + \sqrt{(h_{11} - h_{22})^2 + 4h_{11}h_{22}\rho^2_{12}}]/2(1 - \rho^2_{12})$$

When $\rho_{12} = 0$, we get

$$(3.15)\quad h^2(a_1) = [(h_{11} + h_{22}) + \sqrt{(h_{11} - h^2_{22}) + 4h^2_{12}}]/2$$

When both h_{12} and ρ_{12} are zero, we get $h^2(a_1) = h_{11}$. However, when h_{12} and ρ_{12} are such that $h_{12} = h_{11}\rho_{12}$ or $h_{12} = h_{22}\rho_{12}$, we get

$$(3.16)\quad h^2(a_1) = h_{11}$$

in the former case, but

$$(3.17)\quad h^2(a_1) = \frac{h_{11} - h_{22}\rho^2_{12}}{(1 - \rho^2_{12})}$$

in the later case.

4. GENERALISATION OF RESPONSE TO SELECTION UNDER INDEPENDENT CULLING LEVELS

With several characters there are three methods of selection viz. index selection, independent culling level and tandem selection. The method of index selection is well documented (Smith 1936, Hazel 1943) but the method of independent culling levels has only been considered for two characters by Young and Weiller (1960) and also briefly by Finney (1962). As such, we present here a generalisation of this method for k characters.

Let the k-dimensional normal population of the phenotypic vector $\mathbf{P}$, distributed as

$$(4.1) \quad P(d\mathbf{P}) \propto \exp\left[-\tfrac{1}{2}\mathbf{P}' \underset{P}{\Sigma}^{-1} \mathbf{P}\right] d\mathbf{P}$$

be truncated rectangularly by

$$(4.2) \quad S_k(a_j) = \{\mathbf{P}: P_1 \geq a_1, \ldots\ldots, P_k \geq a_k\}.$$

This amounts to selecting a proportion $P_k(a_j; \underset{P}{\Sigma})$ of the population distributed as in (4.1) and consisting of superior individuals which satisfy (4.2) where

$$(4.3) \quad p_k(a_j; \underset{P}{\Sigma}) = P[P_1 \geq a_1, \ldots\ldots, P_k \geq a_k]$$

$$= \int_{a_1}^{\infty} \ldots \int_{a_k}^{\infty} \left(2\pi\right)^{-\frac{k}{2}} |\underset{P}{\Sigma}|^{-\frac{1}{2}}$$

$$\cdot \exp\{-\tfrac{1}{2}\mathbf{P}' \underset{P}{\Sigma}^{-1} \mathbf{P}\}\, dP_1 \ldots dP_k$$

$$= \int_{a_j}^{(k)} \phi_k(P_j; \underset{P}{\Sigma})$$

and is, therefore, the selection by the method of independent culling levels. The vector of the expected response due to this selection can be investigated as below.

It can be shown that the conditional distribution of the breeding vector, given the set of phenotypic values is multivariate normal with mean vector as $\mathbf{HP}$. The vector of the expected response when $P_j > a_j$ for $j = 1, 2, \ldots, k$ can therefore be obtained by performing matrix integration of $\mathbf{HP}$ with respect to the distribution of the vector $\mathbf{P}$ over the range of values of P_j's. If we denote the vector of response by a $k \times 1$ column vector $\mathbf{As}$ then

$$(4.4)\quad E(\mathbf{As}) = \left\{ \frac{\int_{a_1}^{\infty} \cdots \int_{a_k}^{\infty} \left(\sum_{j=1}^{k} h_{1j} P_j \right) P(d\mathbf{P})}{\int_{a_1}^{\infty} \cdots \int_{a_k}^{\infty} \left(\sum_{j=1}^{k} h_{kj} P_j \right) P(d\mathbf{P})} \right\}$$

$$= \left\{ \frac{\sum_{j=1}^{k} h_{1j} \left\{ \int_{a_1}^{\infty} \cdots \int_{a_k}^{\infty} P_j \, \phi_k \left(P_j ; \sum_{\mathbf{P}} \right) \prod_{i=1}^{k} dP_i \right\}}{\sum_{j=1}^{k} h_{kj} \left\{ \int_{a_1}^{\infty} \cdots \int_{a_k}^{\infty} P_j \, \phi_k \left(P_j ; \sum_{\mathbf{P}} \right) \prod_{i=1}^{k} dP_i \right\}} \right\}$$

$$= \mathbf{H} E[\mathbf{P}; S_k(a_j)]$$

where

$$(4.5)\quad E[\mathbf{P}; S_k(a_j)] = \begin{bmatrix} E(P_1 ; a_j) \\ \vdots \\ E(P_k : a_j) \end{bmatrix}$$

is the mean vector of the population (4.1) truncated by (4.2) Birnbaum and Meyer (1953) has given the formulae for the components of this vector using direct method of evaluating multiple integrals, where as Tallis (1961) has first derived the moment generating function of the truncated multinormal distribution and then obtained the various formulae. In the notations of the latter author,

$$(4.6)\quad E[\mathbf{P}; S_k(a_j)] = \sum_{\mathbf{P}} \mathbf{N}$$

where $\mathbf{N}$ is a $k \times 1$ column vector with i th element as

$$(4.7)\quad P_k\left(a_j ; \underset{P}{\Sigma}\right) N_i$$

$$= \int_{a_j}^{(k-1)} \phi_k \left(P_j , P_s = a_s ; \underset{P}{\Sigma} \right) d P_j \; (k \; (j \neq s)$$

$$= \phi (a_s) P_{k-1} \left(A_{sj} ; \underset{P_s}{\Sigma} \right)$$

where

$$(4.8)\quad A_{sj} = (a_j - \rho_{js} a_s) \mid \sqrt{(1 - \rho^2_{js})}$$

and $\underset{P_s}{\Sigma}$ is the matrix of the first order partial correlation coefficients of P_j for $j \neq s$.

We can, therefore, express $E(A_s)$ as

$$(4.9)\quad E(A_s) = H \underset{P}{\Sigma} N$$

This expression reveals the multivariate analogoue of the expected response to selection for single character given by $h^2_x \sigma_{P_x} i$.

It can be observed from the theoritical derivations given above that the result for expected response to selection valid for a single character is having analogues generalizations in the multi-character case, though the interpretations in the two cases need not necessarily be the same. The response which is expected by the method of independent culling levels depends on the nature of the matrix H, the matrix $\underset{P}{\Sigma}$ and the column vector **N**. The elements of **N** are, however, to be evaluated by (4.7) as suggested by Tallis (1961). In case of selecting for single character, on the other hand, no such problem of evaluation of **N** is involved. The advantage of the generalised case, however, is that when reduced to the case of selection for a single character say r^{th} by letting $a_1 = a_2 = \ldots\ldots = -\alpha$ except a_r, it gives direct response for r^{th} charac ter and also simultaneously correlated responses for the rest of the other characters, which are not under selection. Further, if the relative importance of the set of k-characters under study is known and can be put in the form of weights $w_1, w_2, \ldots, w_k$ expressed as $k \times 1$ column vector **w**, it is possible to obtain the expected response in the overall merit of the individuals due to the selection by independent culling levels by finding $\mathbf{w}' E(A_s)$.

5. SUMMARY

The concepts of heritability and expected response due to selection for a single character have been generalised for a multi-character case when selection is by the method of independent culling levels. A new concept of 'generalized heritability' has been introduced. Theoritical derivations have been presented for the components of dispersion, correlation between the two vectors and for the vector of the expected response to selection by independent culling levels.

REFERENCES

[1] BIRNBAUM, W. Z. and MEYER, P. L. (1953)—On the effect of truncation in some or all co-ordinates of a multi-normal population. J. Indian Soc. Agri. Statist , **5** : 17-28.

[2] FINNEY, D. J. (1962)—Genetic gains under three methods of selection. Genet. Res. (Camb.) **3** : 417-423.

[3] HAZEL, L. N. (1943)—The genetic basis for constructing selection indexes. Genetics **28** : 476-490.

[4] SMITH, H. FAIRFIELD (1936)—A discriminant function for plant selection, Ann. Eugen. (Lond.) **7** : 240-250.

[5] TALLIS, G. M. (1961)—The moment generating function of the truncated multi-normal distribution. J. Roy. Statist. Soc. B. **23** : 223-229.

[6] YOUNG, S. S. Y. and WEILER, H. (1960)—Selection for two correlated traits by independent culling levels. J. Genetics **57** : 329-338.

Indian J. Anim. Sci. 48 (10) : 708–711, October 1978

17

Optimum designs for progeny testing with minimum costs

PREM NARAIN[1]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received : 25 August 1977

ABSTRACT

A general procedure is discussed for the problem of optimum group size in a progeny-testing programme when the aim is to minimize the cost of the programme at a given rate of genetic improvement. The optimum structure of the population is dependent on W, the expected genetic superiority of the selected sires expressed in units of additive genetic standard deviation, the heritability h^2 and the cost ratio (r), the ratio of the cost involved in securing a daughter with at least first lactation record to the cost involved in maintaining a sire till he is progeny tested.

For sex-limited traits like milk production, the breeding value of a sire is assessed with the help of the phenotypic values of its progenies. The correlation between the two tends to unity as the number of progenies tends to infinity. This forms the basis for progeny testing in a dairy-cattle-breeding programme. A basic problem requiring statistical considerations is then how many daughters are needed to test a sire adequately. The larger the number of daughters, the more accurate would be the breeding value of the sire. However, the resources available to a breeder limit the number of daughters to be raised per sire. This calls for an optimum design for progeny testing.

In the context of a progeny-testing programme as a technique for genetic improvement rather than as a method for evaluating sires, the optimum strategy rests on the equation predicting the response to selection based on the progeny-tested bulls. Robertson (1957) worked out the optimum number of daughters per sire such that the expected genetic superiority of the sires is maximized for a given amount of resources. He showed that the optimum number is a function of the heritability (h^2) of the trait, and the testing ratio (K) as the total number of daughters (N) which can be measured for the first lactation yield each generation divided by the number of sires (S) to be selected each generation. However, if the breeder is interested in minimizing the cost of running such a programme for a given rate of genetic improvement, the optimum strategy would have to be worked out differently. Narain (1970, 1971) argued that in such a case, the optimum group size would depend on the fixed rate of genetic improvement per year h^2, and a cost ratio (r) defined as the ratio of the cost (C_1) involved in securing a daughter having at least first lactation record to the cost (C_2) involved in maintaining a bull till he is progeny tested. The objective of this paper is therefore to discuss such a strategy and present a general procedure for determining the optimum design for progeny testing with minimum costs.

Robertson and Rendel (1950) observed that the milk yield of a population of dairy cattle can be increased through genetic improvement with increasing size of the breeding unit. In a herd of less than 100 cows, progeny testing is less efficient than the use of sires selected on the basis of their dams' production. In contrast, if the breeding unit consists of 10,000 cows, it is possible to achieve a genetic gain in milk yield of about 2.5%/year by efficient organization of milk recording,

[1]Present address : Joint Director.

progeny testing and selection of bulls. We shall, therefore, assume that the size of the breeding unit, i.e. the total number of milk-recorded cows (N), is considerably large. To bring about improvement through artificial insemination each of the following 4 paths, from parents to offspring through which the genetic improvement can be affected, needs to be considered. These are (a) sires to breed young bulls, (b) dams to breed young bulls, (c) sires to breed replacement cows and (d) dams to breed replacement cows. Skjervold and Langholz (1964) observed that about 40 to 60% of the total genetic improvement were through breeding of new young bulls from proven sires. Hence it is sufficient to consider, in this discussion, only the genetic gain due to selection among the progeny-tested bulls.

Usually 3 parameters are involved in designing a progeny-testing programme. The first is the size of the breeding unit or testing capacity (N). The second is the number of best bulls (S) out of the tested ones added to the stud each year for use in AI, and the third is the number of young bulls (B) to be progeny tested in each cycle. Two more parameters need to be introduced to take into account the costs involved in running the programme. These are first, C_1, the cost involved in securing a daughter having at least first lactation milk record, and secondly C_2, the cost involved in maintaining a bull till he is progeny tested. The total cost C of running the programme would then be

$$C = NC_1 + BC_2 \qquad (1)$$

Defining the testing ratio $K = N/S$ and the proportion of selected sires $p = S/B$, this cost can be expressed as

$$C = b_1 K + b_2/p \qquad (2)$$

where $b_1 = C_1 S$ and $b_2 = C_2 S$. It shows that the cost function is linear in K but inversely proportional to 'p'. Defining the cost ratio $r = C_1/C_2 = b_1/b_2$, we can express the total cost in units of 'b_2', i.e.

$$C^* = C/b_2 = Kr + 1/p \qquad (3)$$

If the selection intensity in standard deviation units is 'i', r_{AI} is the correlation between the breeding value of the sire and progeny average I, and the additive genetic variance is σ_A^2, the expected genetic superiority (ΔG) of the selected sires can be expressed in units of additive genetic standard deviation as :

$$W = \Delta G/\sigma_A = i.\ r_{AI} \qquad (4)$$

A fixed rate of genetic improvement W can then be obtained by various choices of 'i' and r_{AI}. The problem is then to determine the group size $n = N/B$ such that C* is minimized for a fixed value of W. If we assume that we are selecting from a large sample of sires, $i = Z/p$, where Z is the ordinate of the normal curve at the point where the area cut off is p. Also,

$$r_{AI} = \sqrt{p/(p+a/K)} \qquad (5)$$

where $a = (4-h^2)/h^2$. Since 'a' is fixed for a given value of h^2, we find that a given value of W can be realized for pairs of values of 'p' and K. However, we must ensure that 'i' is greater than or equal to W since r_{AI}, being the heritability of the progeny test, is necessarily less than or equal to 1. Now C* is also a function of 'p' and K for a fixed value of 'r'. We are thus concerned to find the pair of values of 'p' and K which minimizes C* for fixed values of W and 'r'.

Substituting K from the expression for W in the expression for C*, we get

$$C^* = (ar\ W^2 + i^2 - W^2)/p(i^2 - W^2) \qquad (6)$$

Differentiating C* with respect to 'p' and noting that

$$(di/dp) = (x-i)/p \qquad (7)$$

where $x = dZ/dp$, we get

$$(dC^*/dp) = (1/p^2) - ar\ W^2\ (i^2 - 2ix + W^2)/p^2\ (i^2 - w^2)^2 = 0, \text{ at the minimum.} \qquad (8)$$

This gives

$$ar = (i^2 - W^2)^2 / W^2\ (i^2 - 2ix + W^2) \qquad (9)$$

This shows that the optimum value of 'p' depends on the values of W and 'ar'. It decreases as 'ar' increases at a fixed value of W but increases as W decreases at a fixed value of 'ar'. Since 'ar' is a

positive finite quantity, we must have $(i^2-2ix+W^2)$ greater than zero. We have, thus, two conditions on the optimum proportion p given by

$$i \geqslant W \quad (10)$$

$$i^2-2ix + W^2 > 0 \quad (11)$$

The range of permissible values of W goes on, therefore, decreasing with the increase in the value of 'ar'. As 'ar' declines to zero, optimum 'p' approaches the value corresponding to i=W at a fixed value of W. For instance with W=0.5, it approaches 0.70 so that for optimum running of the scheme the intensity of selection between tested sires must be at least 3 in 4.

The minimum value of C* is obtained by substituting the optimum value of 'p' in its expression. It depends also on W and 'ar'. Similarly, the optimum value of K is obtained from

$$K/a = W^2/p(i^2-W^2) \quad (12)$$

which is a function of W and 'p'. The optimum value of 'n' is given by

$$n=pK = aW^2/(i^2-W^2) \quad (13)$$

which shows that it depends on 'a' also, in addition to 'p' and W. However, if 'n' is expressed in units of 'a' so that

$$n/a = W^2/(i^2-W^2) \quad (14)$$

the dependence reduces to W and 'ar' only since 'p' depends on W and 'ar'. We can thus determine the optimum values of 'p' as well as of n/a in terms of W and 'ar' and therefore also the minimum value of C*.

As might be expected, $C^*_{min.}$ increases as W increases at a fixed value of 'ar'. It also increases as 'ar' increases at a fixed value of W. For finite values of 'ar', the value of $C^*_{min.}$ approaches 1 as W tends to zero, irrespective of the value of 'ar'. This is algebraically seen from (3), (4) and (12) since in such a case 'p' and K tend to 1 and 0 respectively. As 'ar' becomes infinitely large, the value of W tends to concentrate only at 0 with $C^*_{min.}$ at 1. At a fixed value of W, as 'ar' declines to zero, $C^*_{min.}$ approaches a finite value of 1/p where p corresponds to i = W. This is seen from (6) and shows that, in (3), although K/a tends to infinitely large value and 'ar' tends to zero, their product Kr tends to 0. For instance, with W = 0.5, $C^*_{min.}$ approaches 1/0.7 = 1.43. This implies that when the cost of maintaining a daughter with at least first lactation record is negligible in relation to the cost of maintaining a bull till he is progeny tested, the optimum design for progeny testing requires considerably large number of daughters per sire under test. But the total cost of running such an optimum programme will be minimal, being inversely proportional to the proportion of tested sires selected for a given rate of genetic improvement.

The optimum values of n/a for values of W between 0.1 and 2.0, and values of 'ar' between 0.1 and 50.0 are presented in Table 1.

Table 1. Optimum values of n/a for different values of W and 'ar'

W	ar							
	0.1	0.5	1.0	2.5	5.0	10.0	20.0	50.0
0.10	0.44	0.18	0.12	0.07	0.05	0.04	0.03	0.02
0.15	0.60	0.25	0.17	0.11	0.08	0.06	0.04	0.03
0.20	0.74	0.33	0.25	0.14	0.10	0.08	0.06	0.04
0.25	0.86	0.39	0.26	0.18	0.13	0.10	0.08	0.06
0.50	1.57	0.73	0.54	0.37	0.30	0.25	0.21	*
0.75	2.29	1.11	0.86	0.63	0.52	0.45	0.40	*
1.00	2.89	1.55	1.26	0.92	0.75	*	*	*
2.00	6.01	3.61	*	*	*	*	*	*

*Optimum 'n' cannot be determined as the corresponding value of W is not permissible.

Table 1 provides with the optimum number of daughters such that the total cost of running the progeny-testing programme is minimum at a fixed rate of genetic improvement provided we know the value of heritability and the cost ratio (r). For example, if the heritability is 0.25, so that 'a' is 15 and the cost ratio (r) is about 1/3, we can get the optimum 'n' by multiplying the values in the table under the column for ar = 5 by 15 for different rates of genetic improvement. For W = 1.0, about 11 daughters per sire will be sufficient. Of course, we cannot fix a higher rate of genetic improvement than W = 1.0 in this case due to the restrictions under (10) and (11). However, if the cost ratio is as low as 1/30, a reference to column with ar = 0.5 and row with W = 1.0, gives the optimum value of 'n' as 23. In this case, however, we can fix still higher rate of genetic improvement for optimizing the group size. With W = 2.0 for example, as many as 54 daughters per sire would now be required. For the situation when the heritability is as low as 0.01 so that 'a' is 399 and if the cost ratio (r) is about 1/8, we have to consult the last column with ar = 50.0. The maximum possible value of W that we can fix now is 0.25 only and the optimum group size turns out to be about 24. This number gets practically doubled if we reduce the cost ratio to 1/80 but fix W at the same level of 0.25. In this case, W can be fixed at still a higher level but the optimum group size gets considerably increased, being about 300 with W = 1.0. It is thus found that at a fixed value of heritability, the optimum group size increases with the decrease in the cost ratio at a given value of W. It also increases with the increase in the value of W at a fixed value of the cost ratio. However, with a decrease in the value of heritability, the optimum 'n' decreases if 'r' is fixed and admissible values of W are chosen.

In the above discussion we have considered the group size which will involve minimum cost for a given rate of genetic improvement. It may, however, be useful to determine the limits of variation in group size which can be tolerated before the cost of programme goes up markedly. This was done empirically by studying the curves of C^* against n/a for different values of W and r at a given value of h^2. It is found that tolerable values of 'n' depend much more on 'W' than on 'r' unless 'r' is very small. For example, Table 2 presents the useful ranges of 'n' in trems of W and 'r' which will involve cost within 10% of the minimum value when $h^2 = 0.25$. The preferred values within the range will increase as h^2 decreases.

Table 2. Range of group size involving cost of the programme within 10% of the minimum value when $h^2 = 0.25$

r	W	
	0.5	1.0
0.01	15–30	21–45
0.20	5–9	11–21
0.40	3–6	8–18
0.60	3–6	8–15

ACKNOWLEDGEMENT

The author thanks Shri P. S. Rana, Statistical Assistant of this Institute, for computational help.

REFERENCES

NARAIN, P. 1970. Influence of breeding on the efficiency of livestock production. *proc. of Seminar on "Methodology of studying productivity in Indian Agriculture"*. I.A.R.S., New Delhi, pp. 348–70.

NARAIN, P. 1971. Experimental designs for selection studies in dairy cattle. *Proc. of Seminar on "Dairy Cattle Breeding"*, NDRI (UNDP), Karnal, pp. 1–7.

ROBERTSON, A. 1957. Optimum group size in progeny testing and family selection. *Biometrics* **13** : 442–50.

ROBERTSON, A. and RENDEL, J. M. 1950. The use of progeny testing with artificial insemination in dairy cattle. *J. Genet.* **50** : 21–31.

SKJERVOLD, H. and LANGHOLZ, H. J. 1964. Factors affecting the optimum structure of AI breeding in dairy cattle *Z. Tierzücht. Zücht. Biol.* **80** : 25–40.

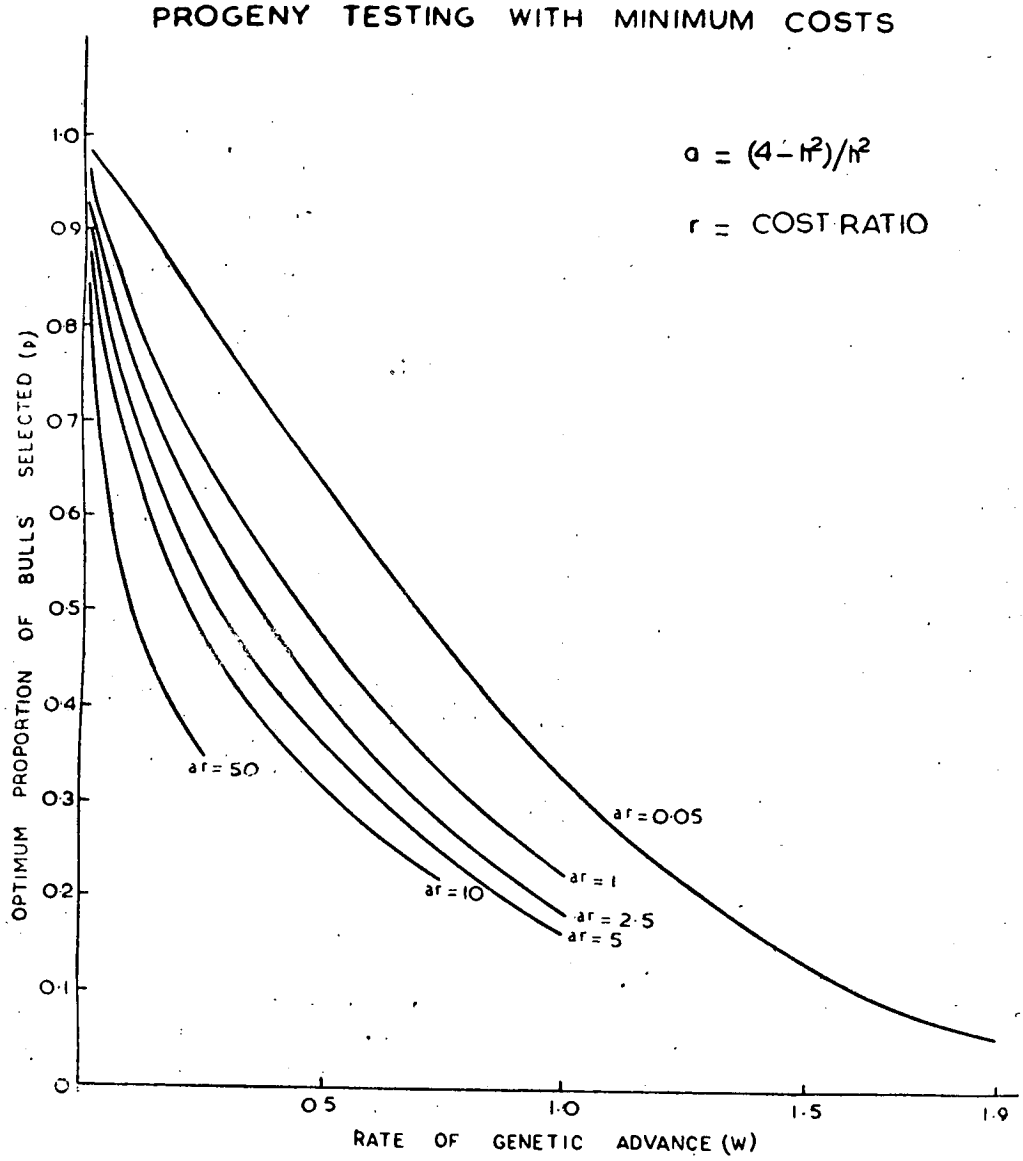

FIG I OPTIMUM 'P' FOR GIVEN 'W' AND 'ar'

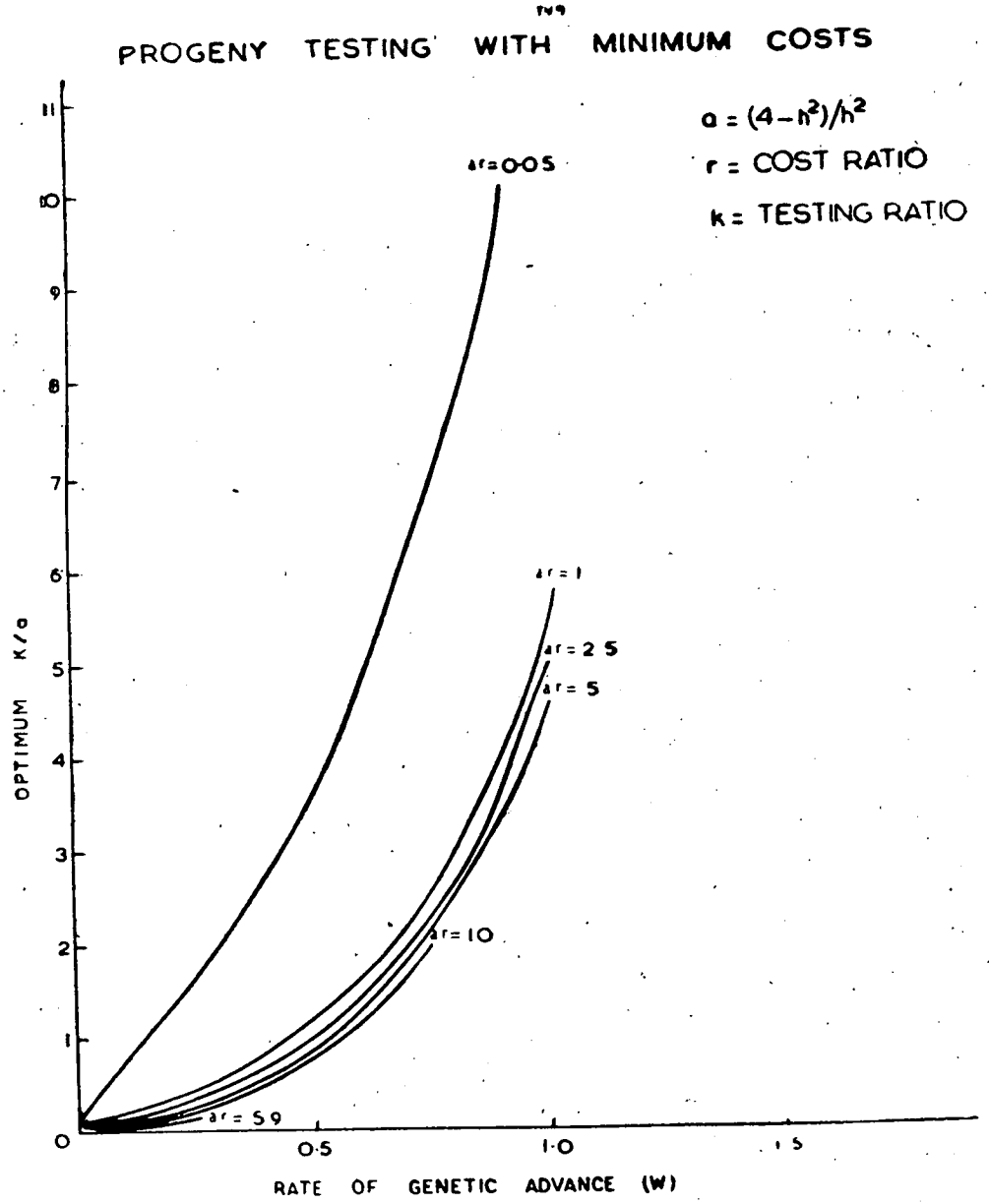

FIG 2 OPTIMUM 'K/a' FOR GIVEN 'W' AND ar

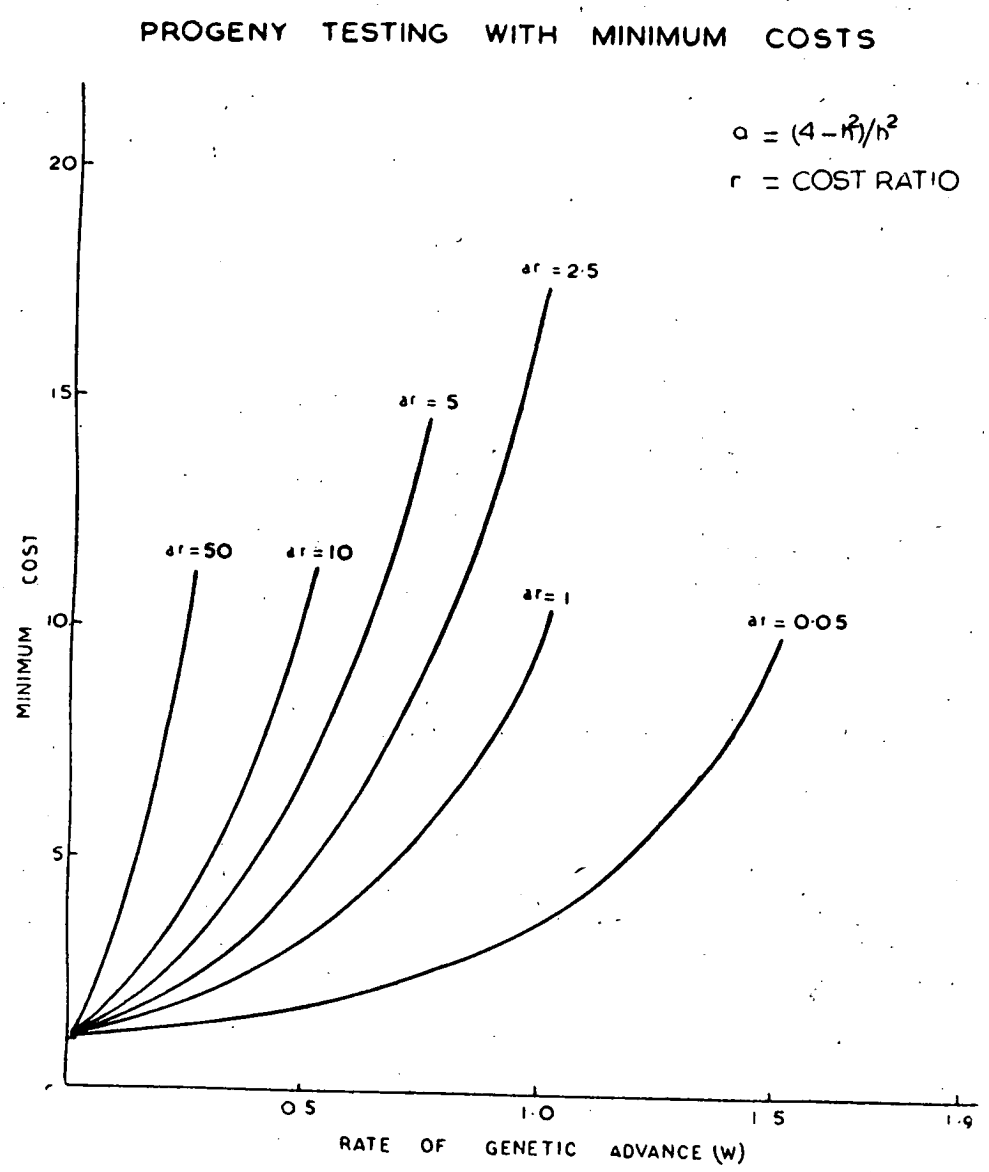

FIG 3 MINIMUM COST FOR GIVEN 'W' AND ar

18

C.P. 42.4

STATISTICAL ASPECTS IN OPTIMISING LIVESTOCK BREEDING PROGRAMMES

by

P. Narain

Contributed Paper
44th Session of the
International Statistical Institute

Madrid - España
September 12th - September 22nd 1983

STATISTICAL ASPECTS IN OPTIMISING LIVESTOCK BREEDING PROGRAMMES

Prem Narain

Indian Agricultural Statistics Research Institute,
New Delhi - 110012

1. INTRODUCTION

The genetic improvement of economic characters in livestock is normally affected by selecting genetically superior animals for the given trait for breeding the next generation. The rate of genetic improvement per generation due to such selection in a population with a given amount of variability depends upon the intensity with which the selection is applied as well as the accuracy of selection as measured by the correlation between the breeding value of the individual under selection and the criterion of selection. Generally, there is a conflict between these two aspects as due to limited resources and limited size of family, increasing the accuracy of selection results in decreasing the intensity of selection. Breeding programmes are therefore optimised by choosing either a strategy which maximises the genetic gain for fixed resources (Robertson, 1957) or a strategy which minimises the cost of running the programme for a fixed rate of genetic improvement (Narain, 1978). However, breeding strategies could be devised which maximises the total returns on expenditure particularly when such returns tend to accumulate over long period of time as in the case of progeny testing programmes. In this paper, statistical aspects in such optimisation problems are considered with particular reference to progeny testing in dairy cattle.

2. THEORY AND DISCUSSION

Consider a progeny testing programme in dairy herds where the pattern of breeding in successive rounds, consists of dividing the total number (N) of female population into two groups, one consisting of elite cows, (1-P) fraction of the total, to be mated to a given number (S) of proven bulls to secure future young males and replacement cows and the other, the remaining female population to be mated to a certain number (B) of young bulls (which are sons of the best proven bulls of

an earlier set) for testing (Narain, 1977). In such a case, we consider only two paths of genetic improvement, from sire to sons and from sire to daughters which together constitute as much as 60 to 70 per cent of total expected genetic improvement. The optimisation problem is then to choose the number of daughters (n) per bull under test such that the profit (π) accruing from the programme in terms of the present value of all future returns is maximised. Since $n = pK$ with $p = S/B$, the proportion of proven bulls selected and $K = N/S$, the testing ratio (Robertson, 1957), the problem boils down to optimising p for given values of K. The profit function is found to be of the form

$$\pi = \alpha(z/p)\phi - \beta(C_o + 1/p) \qquad (1)$$

where z is the ordinate of the normal curve at the point where the area cut off is p, C_o is the component of total cost which is independent of p and

$$\phi = \{ p/(p + a/K)\}^{\frac{1}{2}} \qquad (2)$$

$$a = (4 - h^2)/h^2 \qquad (3)$$

$$\alpha = (2 - P)M\, \sigma_g v/2mR(1 + R)^{y-2} \qquad (4)$$

$$\beta = (1 + R)/R \qquad (5)$$

where M is the number of lactations in which the response is expressed, depending on the replacement rate and average herd life, σ_g is the genetic standard deviation and h^2 the heritability of the milk yield, v is the monetary return for 1 per cent increase the average milk production (m), y is the number of generations before returns start accruing and R is the interest rate per generation.

Setting $\partial\pi/\partial p = 0$ for maximum profit, we get a cubic equation in ϕ as

$$z\phi^3 - (2px - z)\phi - 2\beta/\alpha = 0 \qquad (6)$$

where $x = dz/dp$. This gives three roots for ϕ, of which the positive real root is taken as the solution. From this the optimum value of p for a given value of (K/a) is obtained.

It is interesting to find that the relation between optimum P and (K/a) depends on the values of the function given by

$$2\beta/\alpha = 2m(1 + R)^{y-1} / (2 - P)M\sigma_{\xi}v \qquad (7)$$

As this value tends to zero, we obtain the same curve for the relationship between p and (K/a) as that obtained by Robertson (1957) where the cost of the scheme is not taken into account. In such a case, for optimum running of the scheme, the intensity of selection between tested sires must be at least one in four. However, when the value of $(2\beta/\alpha)$ tends to be high, the optimum intensity of selection between tested sires tends to be lower than that obtained by Robertson (1957).

BIBLIOGRAPHY

1. Narain, P. (1977). A plan for evolving a dairy breed utilising crossbred cattle available in military dairy farms. Military Farms Journal. 9(1), 41-47.

2. Narain, P. (1978). Optimum designs for progeny testing with minimum costs. Indian Journal of Animal Science. 48(10), 708-711.

3. Robertson, A. (1957). Optimum group size in progeny testing and family selection. Biometrics. 13, 442-50.

Key-words: Optimum intensity of selection; Livestock breeding programmes,

SUMMARY

The problem of determining the optimum intensity of selection in a progeny testing programme is discussed. The optimisation problem takes into account the costs as well as returns and the optimum is so chosen that the profit accruing from the programme in terms of the present value of all future returns due to increased milk production is maximised.

RESUMÉ

Le problème de la détermination de l'intensité optimum de sélection dans une programme d'examen de progéniture a été étudié. Le problème de l'optimisation prend en consideration les coûts ainsi que les rendements; on choisit l'optimum de telle manière quo le rendement du programme par rapport à la valeur actuelle de tous les rendements futurs dû à la production augmentée du lait est maximisé.

19

Journal of Genetics Volume 62 Number 2 December 1975

EFFICIENCY OF SELECTIVE BREEDING BASED ON A PHENOTYPIC INDEX

PREM NARAIN AND A. K. MISHRA*
Institute of Agricultural Research Statistics
Library Avenue, New Delhi 110 012

Genetic improvement in a quantitative trait in a population is usually achieved by adopting selective breeding on the basis of the phenotypic values of a trait. The rate of this improvement can, however, be increased if the variation in this trait due to auxiliary traits, particularly at the environmental level, are minimised as far as possible. In such a case selection is made on the basis of an index expressed as deviation of the phenotypic value of the trait from its expected value predicted with the help of the auxiliary traits. Such an index may be called 'phenotypic index' to distinguish it from 'selection index' introduced by Hazel (1943). The genetic improvement expected on the basis of such a selection procedure depends on the number of auxiliary traits, the heritabilities of the traits and the genetic as well as phenotypic correlations between pairs of traits. However, unlike the case of selection index, a knowledge of the estimates of genetic parameters is not necessary for constructing the phenotypic index. As such it is easier to adopt this procedure. If we take the expected genetic improvement in the main trait without the use of any auxiliary traits as a standard of comparison, the efficiency of selective breeding based on the phenotypic index can be expressed as the ratio of the two expected genetic improvements in the main trait. Selection on the basis of phenotypic index is then useful whenever this ratio is expected to be greater than one. This idea of increasing the rate of genetic improvement was first initiated by Rendel (1954) who found that the efficiency of selective breeding for a trait of incomplete heritability may be increased by basing selection on an index which corrects the variation of the main trait for measurable variation introduced by other traits at the environmental level. Osborne (1957) gave a revised estimate for the efficiency of selective breeding for the case when the traits are also genetically correlated. Purser (1960) and Searle (1965) further considered this technique. However, these studies considered only one auxiliary trait. No attempt has so far been made to include more than one auxiliary trait in this method of selection. One can use the technique of partial regression for correcting the variation in the main trait due to several auxiliary traits and investigate the conditions under which this efficiency is increased. This article therefore deals with a study of a 'phenotypic index' based on several auxiliary traits.

* Indian Institute of Management, Ahmedabad

THEORY

Consider n auxiliary traits x_k (k= 1, 2, . . . , n) related to the main trait y. Let the phenotypic value and breeding value of x_k and y, expressed as deviations from the population means, be denoted by $P(x_k)$, $P(y)$ and $A(x_k)$, $A(y)$ respectively. Also, let the phenotypic values be standardised to have unit variances so that the heritabilities of the traits, $h^2(x_k)$ and $h^2(y)$ are the same as the respective genetic variances. Since the regression coefficient of A(y) on P(y) is $h^2(y)$, the expected genetic gain in y due to selection made on the basis of y itself is

$$\Delta G = i\, h^2(y)$$

where i is the intensity of selection.

Consider now selection, with the same intensity, made on the basis of a 'phenotypic index' given by

$$I_P = P(y) - \sum_{k=1}^{n} b_k P(x_k) \tag{2}$$

where b_k is partial regression coefficient of P(y) on $P(x_k)$. The expected genetic gain in y is now

$$\Delta G^* = i\, b_{A(y)I_P} \sigma(I_P) \tag{3}$$

where $b_{A(y)I_P}$ is the regression coefficient of A(y) on I_P and $\sigma(I_P)$ is the phenotypic standard deviation of I_P. The regression coefficient of A(y) on I_P is, in view of (2), equivalent to partial regression coefficient of A(y) on P(y) when $x_1, x_2, \ldots, x_n$ are held constant. Similarly $\sigma(I_P)$ is the standard deviation of P(y) eliminating the effects of the auxiliary traits.

In order to obtain an expression for $b_{A(y)I_P}$ we set up the relationship,

$$E[A(y)] = a_o P(y) + \sum_{k=1}^{n} a_k P(x_k) \tag{4}$$

and evaluate a_o with the help of the resulting normal equations:

$$\begin{aligned} a_o + \underline{R}_o' \underline{a} &= h^2(y) \\ a_o \underline{R}_o + \underline{\underline{R}}\, \underline{a} &= h^2(y)\, \underline{C} \end{aligned} \tag{5}$$

where $\underline{R}_o' = (R_{o1}, R_{o2}, \ldots, R_{on})$, R_{ok} being the phenotypic correlation coefficient between y and x_k, $\underline{\underline{R}}$ is the n x n correlation matrix of phenotypic correlation coefficients R_{kl} between the auxiliary traits and $\underline{C}' = (C_1, C_2, \ldots C_n)$, C_k being $r_{ok}\, h(x_k)/h(y)$ where r_{ok} is the genetic correlation coefficient

between y and x_k. It may be noted that C_k is the relative efficiency of indirect selection based on x_k as discussed by Searle (1965). We then get

$$b_{A(y)I_P} = a_o = h^2(y)(1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{C})(1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o)^{-1} \quad (6)$$

Similarly, to obtain an expression for $\sigma(I_P)$, we set up the relationship,

$$E[P(y)] = \sum_{k=1}^{n} b_k P(x_k) , \quad (7)$$

and evaluate b_k's with the help of

$$\underline{\underline{R}}\, \underline{b} = \underline{R}_o \quad (8)$$

where $\underline{b}' = (b_1, b_2, \ldots, b_n)$. This gives

$$\sigma^2(I_P) = 1 - \sum_{k=1}^{n} b_k R_{ok}$$

$$= 1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o \quad (9)$$

Denoting the efficiency of selection by phenotypic index relative to individual selection by E_P and using (1), (3), (7) and (10), we get

$$E_P = \Delta G^*/\Delta G$$

$$= (1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{C})(1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o)^{-1/2} \quad (10)$$

However, if we consider selection, with the same intensity, made on the basis of the usual selection index of Hazel (1943) given by

$$I_S = P(y) - \sum_{k=1}^{n} w_k P(x_k) , \quad (11)$$

we have to choose optimum values of $\underline{w}' = (w_1, w_2, \ldots, w_n)$. For this we maximise the ratio of genetic gain in y due to selection on the basis of I_S to the gain by direct selection on y. This ratio is efficiency of selection based on I_S given by

$$E_S = \Delta G^{**}/\Delta G$$

$$= b_{A(y)I_S} \sigma(I_S)/h^2(y)$$

$$= (1-\underline{w}' \underline{C})(1-2\underline{w}' \underline{R}_o + \underline{w}' \underline{\underline{R}} \underline{w})^{-1/2} \quad (12)$$

It is found that E_S is maximum when

$$\underline{w} = R^{-1}(\underline{R}_o - \underline{C}K) \quad (13)$$

and the maximum value of E_S is given by

$$E_S = (1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{C} + K \underline{C}' \underline{\underline{R}}^{-1} \underline{C})(1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o + K^2 \underline{C}' R^{-1} \underline{C})^{-1/2} \qquad (14)$$

where

$$K = (1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o)(1-R_o' \underline{\underline{R}}^{-1} \underline{C})^{-1} \qquad (15)$$

If the genetic correlation coefficients r_{ok}'s are all zero, i.e. $\underline{C} = \underline{O}$, then $\underline{w}$ reduces to

$$\underline{w} = \underline{\underline{R}}^{-1} \underline{R}_o$$

$$= \underline{b} \qquad (16)$$

in view of (8) and I_S reduces to I_P. This means when all the auxiliary characters are related to the main character only at the environmental level, the phenotypic index is optimal with efficiency.

$$E_S = E_P = (1-\underline{R}_o' \underline{\underline{R}}^{-1} \underline{R}_o)^{-1/2} \qquad (17)$$

However, even if all the r_{ok}'s are not zero, the phenotypic index could be used though its efficiency would then be less than maximal. This is seen from the relationship between the two efficiencies, given by

$$E_P = (E_S^2 - \underline{C}' \underline{\underline{R}}^{-1} \underline{C})^{-1/2} \qquad (18)$$

CONDITIONS FOR THE USE OF PHENOTYPIC INDEX

It is apparent from the theory given above that if we choose auxiliary characters which have no genetic correlation with the main character, the use of phenotypic index for selection is optimal and is expected to result in maximum genetic improvement in the character. However, if we happen to choose auxiliary characters which are genetically related to the main character, the use of phenotypic index may still result in more genetic improvement in the character than that expected on directly selecting for it, provided certain conditions are satisfied. We therefore investigate below the conditions under which E_P is greater than one.

The relation (10) shows that E_P is always more than unity whenever the corresponding elements of the vectors $\underline{R}_o$ and $\underline{C}$ are of opposite signs. But when this is not so, E_P cannot exceed unity if $\underline{C} \geqslant \underline{e}$ where $\underline{e}$ is a vector with unit elements. However, if $\underline{C} < \underline{e}$, the efficiency may or may not be greater than unity. In such cases, in order that $E_P > 1$, we must have

$$\left(\underline{R}_o' - \frac{2\,\underline{C}'}{1 + \underline{C}'\,\underline{\underline{R}}^{-1}\,\underline{C}} \right) \underline{\underline{R}}^{-1}\,\underline{C} > 0 \qquad (19)$$

Thus when the corresponding elements of $\underline{R}$ and $\underline{C}$ are of the same signs, E_P is greater than one provided $\underline{C} < \underline{e}$ and (19) is satisfied. For example, when $n = 1$, we have only two parameters $C_1 = C$ and $R_{o1} = R$ affecting the efficiency. It is always more than one whenever R and C are of opposite signs. However, when R and C are either both positive or both negative, the efficiency exceeds unit only when $|C| < 1$ and $|R| < 2|C|/(1+C^2)$.

It is further seen from (10) that for given values of $\underline{\underline{R}}^{-1}$ and $\underline{C}$, the efficiency is a function of $\underline{R}'$. So differentiating (10) with respect to $\underline{R}_o'$ and equating it to zero, we find that the efficiency is maximum or minimum when $\underline{R}_o = \underline{C}$. The second differential of E_P at $\underline{R}_o = \underline{C}$ is, however, positive so that the efficiency is minimum when the corresponding elements of $\underline{R}_o$ and $\underline{C}$ are equal in magnitude and possess the same sign. It becomes one when either $\underline{R}_o = \underline{o}$ or $\underline{R}_o = 2\underline{C}/(1+\underline{C}'\,\underline{\underline{R}}^{-1}\,\underline{C})$. For one auxiliary character, when R and $\underline{C}$ are equal and of the same sign, the efficiency is less than one and possesses the minimum value of $(1-R^2)^{1/2}$. It becomes one when either $R = o$ or $R = 2C/(1+C^2)$.

We now assume that the auxiliary traits are uncorrelated so that $\underline{R}$ is an identity matrix. We then have

$$\begin{aligned} E_P &= (1 - \underline{R}_o'\,\underline{C})\,(1 - \underline{R}_o'\,\underline{R}_o)^{-1/2} \\ &= \left(1 - \sum_{k=1}^{n} R_{ok} C_k\right)\left(1 - \sum_{k=1}^{n} R_{ok}^2\right)^{-1/2} \end{aligned} \qquad (20)$$

Now the efficiency depends on whether the sum of products of R_{ok} and C_k over all the auxiliary characters is positive or negative. It may happen that for some of the characters R_{ok} and C_k may have opposite signs and for others they may have the same sign. But if the sum of their products happen to be negative, the efficiency will be greater than unity. On the contrary, it will be less than one. The least value of E_P will be $\left(1 - \sum_{k=1}^{n} R_{ok}^2\right)^{1/2}$ and less than one when $R_{ok} = C_k$ for each k. The value of E_P will be $\left(1 + \sum_{k=1}^{n} R_{ok}^2\right)\left(1 - \sum_{k=1}^{n} R_{ok}^2\right)^{-1/2}$ and greater than one when $R_{ok} = -C_k$ for each k. If, however, all C_k's are equal to C but R_{ok}'s are not equal, the efficiency reduces to $\left(1 - C\sum_{k=1}^{n} R_{ok}\right)\left(1 - \sum_{k=1}^{n} R_{ok}^2\right)^{-1/2}$. Now the efficiency will be greater than one if C and the sum of R_{ok} values over the auxiliary characters are of opposite signs. But if all R_{ok}'s are equal to R but C_k's are not equal, the efficiency becomes $\left(1 - R\sum_{k=1}^{n} C_k\right)(1 - nR^2)^{-1/2}$ which would be greater than unity if R and

the sum of C_k values over the auxiliary characters are of opposite signs. If, however, all C_k's are equal to C and all R_{ok}'s are equal to R, we have

$$E_P = (1-nRC)(1-nR^2)^{-1/2} \quad (21)$$

In this form, the effect of the number of auxiliary characters on the efficiency can be seen. The results are presented in Table 1.

Table 1. Effect of the number of traits and the phenotypic correlation on the efficiency of phenotypic index

C	n \ R	-0.20	-0.10	0.00	0.10	0.20
	1	1.061	1.025	1.00	0.985	0.980
+0.2	5	1.348	1.123	1.00	0.923	0.899
	10	1.809	1.264	1.00	0.843	0.775
	15	2.532	1.410	1.00	0.759	0.633
	20	4.027	1.573	1.00	0.674	0.477
-0.2	1	0.980	0.985	1.00	1.025	1.061
	5	0.899	0.923	1.00	1.123	1.348
	10	0.775	0.843	1.00	1.264	1.809
	15	0.633	0.759	1.00	1.410	2.532
	20	0.447	0.674	1.00	1.573	4.027

It is apparent that the efficiency increases with the number of auxiliary traits when R and C are having opposite signs and it decreases when R and C are having the same signs. For given n, the efficiency decreases as R increases for positive values of C. This relation is, however, reversed for negative values of C. Furthermore, we see that when n is small, the changes in the values of efficiency for different values of R keeping C as fixed is very small but when n is large there is a rapid change in the values of the efficiency. For example, when n= 20, C= +0.2, the efficiency varies from 0.447 to 4.027 but when n= 1, it varies only from 0.980 to 1.061.

APPLICATION TO DAIRY CATTLE BREEDING

In order to demonstrate the practical relevance of the technique of phenotypic index developed in this paper, breeding data on cattle of Kankrej breed collected from an organised herd at Anand (Gujarat), India were used. Records were available for the period 1945 to 1963 in respect of milk yield in first lactation (y), age at first calving (x_1) and body weight of calf at birth at

first calving (x_2) for 180 daughters from 13 sires. Table 2 shows the heritabilities, phenotypic and genetic correlations used in the calculation of three different phenotypic indices. The genetic parameters were estimated on an intra-sire basis. The table also shows the three phenotypic indices as well as their estimated relative efficiencies.

Table 2. Efficiency of Phenotypic index for milk yield.

Trait	Heritability		Correlation between traits	
			Phenotypic	Genetic
Milk yield in first lactation (y)	0.415	$y-x_1$	+0.195	-0.048
Age at first calving (x_1)	0.504	$y-x_2$	-0.184	+0.213
Body weight of calf at first calving (x_2)	0.555	x_1-x_2	+0.122	-0.350

Phenotypic indices	Relative Efficiency %
$I_1 = y-1.34\ x_1$	103
$I_2 = y+30.56\ x_2$	106
$I_3 = y-1.51\ x_1+35.00\ x_2$	111
Selection on y alone	100

It is apparent from this table that for improving the milk yield in first lactation, the use of age at first calving as an auxiliary trait results in an increase in the efficiency of selective breeding by about 3%. This increase in efficiency rises to 6% if the auxiliary trait happens to be the body weight of calf at birth in first calving. It is interesting to find that when both of these auxiliary traits are used simultaneously to correct for variations in the milk yield in the first lactation, the relative efficiency increases by as much as 11%.

In the above illustration, the estimates of heritabilities and genetic correlations were used to work out the relative efficiencies of the phenotypic index but not the index itself. However, we may use them to obtain the relative efficiencies of the selection index in the three cases. It is found that with x_1, the efficiency is 104, with x_2 it is 110 and with both x_1 and x_2, it is 113. It is apparent that use of selection index instead of phenotypic index would result in higher relative efficiency but the increase is only marginal. At the same

time, unlike phenotypic index, the working out of the selection index would require the estimation of genetic parameters. In view of the cost involved in meeting this requirement, the use of phenotypic index may be preferable even though some efficiency is sacrificed, provided it is ensured from the past experience, that the phenotypic and genetic correlations for either of the two auxiliary traits are expected to have opposite signs.

SUMMARY

This article discusses the efficiency of selective breeding based on 'phenotypic index' which is defined as the deviation of the phenotypic value of the trait from its expected value predicted with the help of one or more auxiliary traits. The conditions under which the efficiency of such a procedure is greater than one have been theoretically studied. The practical relevance of this technique has also been demonstrated by applying it to breeding data on cattle.

REFERENCES

HAZEL, L.N. (1943). The genetic basis for constructing selection indexes. Genetics 28, 476-490.

OSBORNE, R. (1957). Correction for regression on a secondary trait as a method of increasing the efficiency of selective breeding. Australian Journal of Biological Sciences, 10, 365-366.

PURSER, A.F. (1960). The use of correction for regression on a second character to increase the efficiency of selection. Biometrical Genetics, Edited by O. Kempthorne, 234 pp. Pergamon Press, London.

RENDEL, J.M. (1954). The use of regression to increase heritability. Australian Journal of Biological Sciences, 7, 368-378.

SEARLE, S.R. (1965). The value of indirect selection. I. Mass Selection. Biometrics 21, 682-707.

The Indian Journal of Animal Genetics and Breeding 1(1) : 20-22, *January* 1979

20

A new sire index for milk production corrected for an auxiliary trait

PREM NARAIN

Indian Agricultural Statistics Research Institute, New Delhi 110 012

ABSTRACT

A new corrected daughter average index for selecting dairy sires using correction for an auxiliary trait (x) has been presented. Phenotypic values for the character of interest (y) are expressed as deviations from its expected value predicted with the help of the correlated auxiliary character.

The proposed index is more efficiency than the corrected daughter average index whenever the phenotypic and genetic correlations are in the opposite direction. When the correlations are in the same direction, the new index is more efficient provided $c<1$ and $r_p > 2c/1+c^2$ where c equals $r_G h_x/h_y$. Using data from three herds the superiority of the new index is illustrated.

Dairy sires are selected on the basis of either daughter-dam comparison or comparison of daughters with their contemporaries. In the former category can be listed such indices as 'simple daughter average index', 'intermediate index' and 'corrected daughter average index'. The last-mentioned index is based on correcting the daughters' average on the basis of the regression of daughters' performances on those of dams' for the unequal production levels of the dams mated to different sires. This regression measures half the heritability of the character under consideration.

Narain and Mishra (1975) showed that the efficiency of selection for a given character can be increased if it is based on an index expressing the phenotypic value of this character as a deviation from its expected value predicted with the help of one or more correlated auxiliary characters. Narain (1975, 1976) used this approach in further correcting the daughters' average for the given character, and proposed a sire index which made use of the information on one auxiliary character. It was shown to be an improvement over the corrected daughter average index. A similar approach was earlier used by Sukhatme (1944). He had suggested that correction for the inequality in lactation period could be made by using the regression technique with lactation period as the auxiliary variable. However, he did not take into account the possible genetic relationship between the auxiliary variable and the character under improvement.

In this paper, the newly proposed sire index and its superiority over the corrected daughter average index have been discussed. The regression of the performance of future daughter on a finite number 'n' of daughters under test has also been taken into account. The theory developed has been applied to data on 3 dairy cattle herds, viz. Red Sindhi herd at Bangalore, Red Sindhi herd at Hosur and Tharparkar herd at Patna.

THE NEW SIRE INDEX

The proposed sire index, S_I is given by

$$S_I = A_I + \frac{2nW}{(n + a_y)} \left[(\bar{D}_I - A_I) - \tfrac{1}{2} h_I^2 (\bar{M}_I - A_I) \right]$$

In this index, $\bar{D}_I$ is the average performance of n daughters for the main character y after correcting for the auxiliary character x; $\bar{M}_I$ is the average performance of the corresponding dams for y after correction for x; and A_I is the herd average for character y corrected for x. Further, $h^2{}_I$ is the heritability of y corrected for x, i.e. of the 'phenotypic index'.

$I = P_y - bP_x$

where P_y and P_x are the phenotypic values of y and x expressed as deviations from the mean and b is the regression coefficient of y on x. As shown by Narain and Mishra (1975), h_I^2 is given by

$$h_I^2 = h_y^2 (1 - RC)^2/(1 - R^2)$$

where

$R = bs_x/s_y$

$C = r\, h_x/h_y$

W in the equation for S_I is given by

$$W = \frac{I - r^2 (n+a_{xy})/(n+a_x)}{1 - r^2 (n+a_{xy})^2/(n+a_x)(n+a_y)}$$

$$a_y = (4 - h_y^2)/h_y^2$$

$$a_x = (4 - h_x^2)/h_x^2$$

$$a_{xy} = (4R - r\, h_x h_y)/r\, h_x h_y$$

In the above expressions, h_x^2 and h_y^2 are heritabilities of x and y respectively; R and r are phenotypic and genetic correlations between x and y, respectively, and s_x and s_y are phenotypic standard deviations of x and y respectively.

In terms of the above notations, the corrected daughter average index for y, S_y, making allowance for a finite number 'n' of daughters, is given by

$$S_y = A_y + \frac{2n}{(n+a_y)} \left[(\overline{D}_y - A_y) - \tfrac{1}{2} h_y^2 (\overline{M}_y - A_y) \right]$$

THE RELATIVE EFFICIENCY OF THE NEW INDEX

Assuming the daughters and dams to have equal variabilities for each of the two characters and equal co-variability between the two characters, and neglecting the sampling variances of h_y^2 as well as h_I^2, it was found that

$$V(S_y) = \left[\frac{2n}{(n+a_y)} \right]^2 s_{D_y}^2 (1 - \tfrac{1}{4} h_y^4)/n$$

$$V(S_I) = \left[\frac{2nW}{n+a_y} \right]^2 s_{D_y}^2 (1 - R^2)(1 - \tfrac{1}{4} h_I^4)/n$$

where $s_{D_y}^2$ is sampling variance for daughters for the main character y.

The efficiency of the new index S_I relative to the corrected daughter average index S_y was, therefore, given by

$$\frac{V(S_y)}{V(S_I)} = \frac{(1 - \tfrac{1}{4} h_y^4)}{W^2 (1 - \tfrac{1}{4} h_y^4 E^2 - R^2)}$$

where

$$E = (1 - RC) / \sqrt{(1 - R^2)}$$

the efficiency of the 'phenotypic index' selection relative to individual selection as obtained by Osborne (1957) and Narain and Mishra (1975). If $r = 0$ then $W = 1$ and $C = 0$; the ratio is clearly greater than 1. When n, the number of daughters per sire is adequately large, W tends to be 1 so that the ratio is greater than 1 whenever E is greater than 1. This happens whenever the phenotypic and genetic correlations are of opposite signs. When they are of the same sign, E is greater than 1 if C is less than 1 and R is greater than $2C/(1 + C^2)$.

APPLICATION TO DAIRY CATTLE HERDS

Sachdeva (1972) carried out a detailed analysis of data pertaining to 3 dairy herds for judging whether the proposed index was more efficient relative to corrected daughter average. He chose as many as 9 characters relating to yield as main character. The choice for auxiliary character was dependent on the availability of the estimate of its heritability and its genetic correlation with the main character, the estimate being based on dam-daughter regression. The proposed index was more efficient than the corrected daughter average index in most of the cases. For the sake of illustration only an overall picture is given in this paper. The percentage of sires showing superiority of the proposed index over the corrected

Table 1. Percentage of sires indicating the superiority of the proposed index over the corrected daughter average index

Main character	Herds		
	Red Sindhi (Bangalore)	Red Sindhi (Hosur)	Tharparkar (Patna)
Milk yield in first lactation	75	21	41
Average milk yield per day in first lactation	50	95	84
Average milk yield per day of calving interval	75	50	57

daughter average index for each of the 3 main characters, viz. milk yield in first lactation, average milk yield per day in first lactation and average milk yield per day of calving interval, in the 3 herds are presented in Table 1. The auxiliary characters considered were age at first calving, calving interval and length of first lactation. The proposed index was, in general, more efficient than the corrected daughter average index.

REFERENCES

NARAIN, P. 1975. A new sire index with multiple characters. *Proceedings of Seminar on "Current Researches in Animal Breeding". National Dairy Research Institute, Karnal, Haryana*, pp. 1–7.

NARAIN, P. 1976. Efficiency of sire evaluation methods. *Proceedings of Summer Institute on "Methodology in Dairy Cattle Breeding and Production", National Dairy Research Institute, Karnal, Haryana*, pp. 1–9.

NARAIN, P. and MISHRA, A. K. 1975. Efficiency of selective breeding based on a phenotypic index. *Journal of Genetics* **62** (2) : 69–76.

OSBORNE, P. 1957. Correction for regression on a secondary trait as a method of increasing the efficiency of selective breeding. *Australian Journal of Biological Sciences* **10** : 365–66.

SACHDEVA, K. S. 1972. *Sire evaluation in cattle breeding*. Diploma Thesis, Institute of Agricultural Research Statistics, New Delhi.

SUKHATME, P. V. 1944. Note on the method of calculating sire index for milk production in cattle. *The Indian Journal of Veterinary Science and Animal Husbandry* 14 : 119 23.

(Received ; 28-12-1978)

21

Indian J. Poult. Sci. Vol. 18, 37-47 (1983)

Reprint

The Use of Auxiliary Traits in Combined Selection for Poultry Improvement

P. Narain, P. K. Malhotra and S. D. Wahi
Agricultural Statistics Research Institute, Library Avenue, New Delhi-110 012

Abstract : The genetic improvement of production traits is normally achieved by selecting genetically superior individuals for the given trait and mating them to produce the next generation. The genetic superiority of the individuals can, however, be determined either on the basis of their own performance or the performance of their relatives such as full-sibs and half-sibs or else an optimum combination of several such information all for the same trait. The latter selection scheme involves construction of a selection index and is known as combined selection. In this paper a new selection index which combines information on several auxiliary traits with the information on the trait under improvement for the individual as well as its relatives, such as full-sibs, half-sibs and dams, has been developed. Its efficiency has been found to be more than the efficiency of an index without any auxiliary trait, the increase in efficiency depending on several parameters The theoretical results have been supplemented with practical results on poultry data analysis involving egg production upto 240 days of age as the main trait and age at first egg and egg weight as the two auxiliary traits. The use of age at first egg as the auxiliary trait results in an increase in the efficiency of the index by about 8 to 9% whereas the use of the egg weight increases it by about 6 to 7%. When both of these traits are used, the efficiency increases by as much as 14 to 16%. The inclusion of dam's performance is found to decrease the efficiency of the new index by about 1%.

Introduction

The results of the analysis of data collected during the operation of the Poultry Breeding Programme for evolving suitable strains of egg type chicken at M. P. Government Regional Poultry Farm, Bhopal were reported in Narain *et al.* (1979). It was found that while the average rate of lay improved due to selection, the average egg weight deteriorated. It appeared therefore that although the method of selection based on the index is effective in increasing the rate of lay, it resulted in a correlated decline in the average egg weight. It was, therefore, desirable to take into account the correlation between rate of lay and egg weight, while improving the rate of lay. This raised the statistical problem of divising an index which includes the individual performance for another character such as egg weight in addition to combining the information about rate of lay on the individual bird with those of its full-sibs and half-sibs. Narain *et al.* (1977, 1979) discussed such an index and found that the efficiency of new index is always increased when the genetic and phenotypic correlation between the character under improvement and the auxiliary trait are of opposite signs However, no attempt was made to generalise the index so as to include several auxiliary traits.

In the present paper, therefore, we construct a selection index which combines the information on several auxiliary traits with the information on the trait under improvement for the individual bird as well as its relatives, such as full-sibs, half-sibs and dams. The theoretical results are

Dedicated to the memory of Dr. D. K. Biswas.

supplemented with practical results on poultry data analysis involving egg production upto 240 days of age as the main trait and age at first egg and egg weight as the two auxiliary traits.

Theory

We first consider the index I_{1k} $(x_1, x_2, \ldots x_k)$ which predicts the breeding value (G_y) of the individual for the character (y) under improvement by combining, in an optimal manner, its own performance (P_y) for y, its performance $\underset{\sim}{P_x}=({}^Px_1, {}^Px_2, \ldots {}^Px_k)$ for the set of k correlated characters $\underset{\sim}{x}=(x_1, x_2, \ldots x_k)$, the average $(\bar{H}_y)$ of the phenotypic values of n paternal half-sibs for y and the average $(\bar{F}_y)$ of the phenotypic values of m full-sibs for y.
The index is given by

$$I_{1k}(x_1 \ldots x_k)=\sum_{i=1}^{k} a_i {}^Px_i + b_1 P_y + b_2 \bar{H}_y + b_3 \bar{F}_y \qquad (1)$$

where coefficients $a_1, \ldots a_k$, b_1, b_2 and b_3 are to be worked out in such a way that accuracy in the prediction of G_y on the basis of I_{1k} is maximised i. e. the multiple correlation between G_y and I_{1k} is maximum. We next consider I_{2k} $(x_1, x_2 \ldots, x_k)$ which predicts G_y by combining, in an optimal manner, in addition to the information included in I_{1k} the information supplied by the average performance (D_y) of the individual's dam for y. This index is then given by

$$I_{2k}(x_1, x_2, \ldots x_k)=\sum_{i=1}^{k} a_i' P_{x_i} + b_1' P_y + b_2' \bar{H}_y + b_3' \bar{F}_y + b_4' D_y \qquad (2)$$

where coefficients $a'_1, a'_2, \ldots a'_k$, b'_1, b'_2, b'_3 and b'_4 are different from those of I_{1k} but are similarly obtained by maximising the multiple correlation between G_y and I_{2k}.

For deriving the coefficients, it is necessary to know the biometric correlation coefficients between relatives for given values of heritabilities of the (k+1) characters (h^2x_1, $h^2x_2 \ldots$ h^2x_k, and h^2_y), the genetic as well as phenotypic correlation coefficients between $x_1 \ldots x_k$ and y. The phenotypic correlation coefficients between P_y and ${}^Px_j$ are denoted by R_{oj}, j=1, 2,...k, whereas R_{ij} denotes the phenotypic correlation between ${}^Px_i$ and ${}^Px_j$, i≠j=1, 2,...k. The genetic correlation coefficients between P_y and ${}^Px_j$ are denoted by r_{oj}, j=1, 2,...k. The increase in the accuracy of prediction on the basis of I_{1k} or I_{2k} over that on individual performance for y can then be judged by comparing the correlation coefficient between G_y and the corresponding index i. e. ${}^rGI_{1k}$ and ${}^rGI_{2k}$ with h_y. The corresponding efficiencies E_{1k} and E_{2k} are given by

$$E_{1k}={}^rGI_{1k}/h_y, \quad E_{2k}={}^rGI_{2k}/h_y \qquad (3)$$

For comparing their efficiencies relative to Osborne indices I_{10} and I_{20}, we calculate the ratios ${}^rGI_{1k}/{}^rGI_{10}$ and ${}^rGI_{2k}/{}^rGI_{20}$ where ${}^rGI_{10}$ and ${}^rGI_{20}$ are the correlation coefficients between G_y and I_{10} and I_{20} respectively.

By following the path coefficient approach of Wright (1921) such correlations were derived with the restriction that the averages of full sibs $(\bar{F}_y)$ and half-sibs $(\bar{H}_y)$ do not include the observation on the individual P_y. These are presented in Table 1.

When n=m=1, we get N=M=1 and from Table 1, we get the correlations for the case when only a single individual of each kind is available. Also when n and m are infinitely large, N and M tends to $(4/h^2_y)$ and $(2/h^2_y)$ respectively.

(a) *Index I_{1k}*

The normal equations for the index I_{1k} obtained by the above procedure are as follows :

$$\begin{aligned}
&{}^R\acute{o}a + b_1 + {}^1/_4 b_2 h_y^2 + {}^1/_2 b_3 h_y^2 = h_y^2 \qquad (4)\\
&Ra + b_1 R_o + {}^1/_4 b_2 c + {}^1/_2 b_3 c = h_y^2 c\\
&{}^1/_2 c'a + {}^1/_4 b_1 h_y^2 + {}^1/_N b_2 + {}^1/_4 b_3 h_y^2 = {}^1/_4 h_y^2\\
&{}^1/_2 c'a + {}^1/_2 b_1 h_y^2 + {}^1/_4 b_2 h_y^2 + {}^1/_M b_3 = {}^1/_2 h_y^2
\end{aligned}$$

Table 1. Matrix of biometrical correlations between relatives and auxiliary traits

	P_{X_1}	P_{X_2}	P_{X_k}	G_ν	P_ν	H_ν	F_ν	D_ν
P_{X_1}	1	R_{11}	R_{1k}	$r_{01}h_{X_1}$	R_{01}	$\frac{1}{4}\, r_{01}h_{X_1}h_\nu\sqrt{N}$	$\frac{1}{2}\, r_{01}h_{X_1}h_\nu\sqrt{M}$	$\frac{1}{2}\, r_{01}h_{X_1}h_\nu$
P_{X_2}		1	R_{2k}	$r_{02}h_{X_2}$	R_{02}	$\frac{1}{4}\, r_{02}h_{X_2}h_\nu\sqrt{N}$	$\frac{1}{2}\, r_{02}h_{X_2}h_\nu\sqrt{M}$	$\frac{1}{2}\, r_{02}h_{X_2}h_\nu$
P_{X_k}			1	$r_{0k}h_{X_k}$	R_{0k}	$\frac{1}{4}\, r_{0k}h_{X_k}h_\nu\sqrt{N}$	$\frac{1}{2}\, r_{0k}h_{X_k}h_\nu\sqrt{M}$	$\frac{1}{2}\, r_{0k}h_{X_k}h_\nu$
G_ν				1	h_ν	$\frac{1}{4}\, h_\nu\sqrt{N}$	$\frac{1}{2}\, h_\nu\sqrt{M}$	$\frac{1}{2}\, h_\nu$
P_ν					1	$\frac{1}{4}\, h_\nu^2\sqrt{N}$	$\frac{1}{2}\, h_\nu^2\sqrt{M}$	$\frac{1}{2}\, h_\nu^2$
H_ν						1	$\frac{1}{4}\, h_\nu^2\sqrt{MN}$	0
F_ν							1	$\frac{1}{2}\, h_\nu^2\sqrt{M}$
D_ν								1

In this table, $N = \frac{n}{1+(n-1)h_\nu^2/4}$ and $M = \frac{n}{1+(m-1)h_\nu^2/2}$, where n is the number of half-sibs and m is the number of full-sibs.

where $R_{\acute{o}}=(R_{o1}, R_{o2}...,R_{ok})$, R is a $k \times k$ correlation matrix of R_{ij}'s, $r'_o=(r_{o1}, r_{o2} \quad r_{ok})$, $c'=(c_1, c_2,...c_k)$, c_j being $r_{oj}\ h x_j/h_y$, $a'=(a_1, a_2 \quad ,a_k)$ and $\sigma_y=\sigma x_1=\sigma x_2=...\sigma x_k$.

Solving these, we get

$$a=h_y^2 R^{-1}(c-R_o)S/D \qquad (5)$$
$$b_1=h_y^2(1-R_{\acute{o}}R^{-1}c)S/D_1$$
$$b_2=2Nh_y^2(2-Mh_y^2)T/D_1$$
$$b_3=Mh_y^2(8-Nh_y^2)T/D_1$$

where

$$S=16-4Mh_y^2-Nh_y^2 \qquad (6)$$
$$U=4M+N-MNh_y^2$$
$$T=(1-R_{\acute{o}}R^{-1}R_o)-h_y^2(1+c'R^{-1}c-2c'R^{-1}R_o)$$
$$D_1=S(1-R_{\acute{o}}R^{-1}R_o)+TUh_y^2$$

The efficiency of the index I_{1k} is then given by

$$E_{1k}=\sqrt{\frac{[h_y^2+\alpha(1-h_y^2)](1-R_{\acute{o}}R^{-1}R_o)-h_y^2[(R_o-r_o)+\alpha(C-R_o)]'R^{-1}(C-R_o)}{h_y^2[1+\alpha(1-h_y^2)](1-R_{\acute{o}}R^{-1}R_o)-\alpha h_y^4(C-R_o)'R^{-1}(C-R_o)}} \qquad (7)$$

where

$\alpha=Uh_y^2/S$. When we consider only one auxiliary trait i. e. k=1, the above expressions (4), (5), (6) and (7) reduce to those given in Narain *et al.* (1977). When phenotypic and genetic correlations are zero i.e. $R_o=r_o=0$, C=0, E_{1k} reduces to E_{10}, the corresponding efficiency of Osborne index given by

$$E_{10}=\sqrt{[h_y^2+\alpha(1-h_y^2)]/h_y^2[1+\alpha(1-h_y^2)]} \qquad (8)$$

It is interesting to note that this is also true when $R_o=C$, even when R_o and r_o are not zero. Thus the condition $R_o=C$ introduces a situation as if there are no auxiliary traits. When only genetic correlations are zero i.e. $r_o=0$, C=0, E_{1k} reduces to

$$E_{1k}(r_o=0)=\sqrt{\frac{h_y^{*2}+\alpha(1-h_y^{*2})}{h_y^2[1+\alpha(1-h_y^{*2})]}} \qquad (9)$$

where

$$h_y^{*2}=h_y^2/(1-R_o'R^{-1}R_o) \qquad (10)$$

is the heritability of the trait after corrections for the auxiliary traits as shown in Narain & Mishra (1975). Since h_y^{*2} is necessarily greater than h_y^2, the efficiency is always greater than that for Osbrone's index.

In the limiting case. when n, the number of half-sibs and m, the number of full-sibs approach infinity, U tends to $(4/h_y^2)$ and S tends to 4 so that α tends to unity. The efficiency E_{1k} then takes the limiting value

$$\text{Lim.}\ E_{1k}=\sqrt{\frac{1-h_y^{*2}(C-r_o)'R^{-1}(C-R_o)}{h_y^2[2-h_y^2-h_y^{*2}(C-R_o)'R^{-1}(C-R_o)]}} \qquad (11)$$

When $R_o=r_o=0$ or when $R_o=C$, this limiting value reduces to

$$\text{Lim.}\ E_{10}=\frac{1}{\sqrt{h_y^2(2-h_y^2)}} \qquad (12)$$

of the corresponding Osborne index. However, when only genetic correlations are zero, the limiting value reduces to $1/\sqrt{h^2_y(2-h_y^{*2})}$ as expected.

It is apparent from (7) that the efficiency of the new index depends on a large number of parameters and as such it is not possible to study its behaviour unless we introduce some simplifications. For instance, if we assume that the

auxiliary traits are uncorrelated, this makes R an identity matrix. We may further assume equality of all C_j's, r_{oj}'s and R_{oj}'s. This would mean equality of all h^2_{Xj}'s for the auxiliary characters. The efficiency would still depend upon as many as six parameters, including one relating to k the number of auxiliary traits. However, the behaviour of E_{1k} with variations in h_y^2 for given values of k, m, n, heritabilities of auxiliary traits and their correlations with the main trait can be numerically studied.

(b) *Index I_{2k}*

By the similar procedure, the weights for the index I_{2k} a', b_1', b'_2, b'_3, b'_4 are given by

$$\begin{aligned} a' &= h_y^2 R^{-1}(c - R_o)\ (S - 4h_y^2)/D_2 \\ b'_1 &= h_y^2\ (1 - R_o'\ R^{-1}c)\ (S - 4h_y^2)/D_2 \\ b'_2 &= 2Nh_y^2\ (2 - Mh_y^2)T/D_2 \\ b'_3 &= Mh_y^2[8 - Nh_y^2) - 4h_y^2]T/D_2 \\ b'_4 &= 4h_y^2(2 - Mh_y^2)\ T/D_2 \end{aligned} \qquad (13)$$

where

$$\begin{aligned} D_2 &= (S - 4h_y^2)\ (1 - R_o'R^{-1}R_o) + TVh_y^2 \\ &= D_1 - 4h_y^2[(1 - R_o'R^{-1}R_o) + T(1 - Mh_y^2)] \qquad (14) \\ V &= 4M + N - MNh_y^2 - 4 - 4Mh_y^2 \\ &= U + 4\ (1 - Mh_y^2) \end{aligned}$$

The efficiency of the index I_{2k} is then given by

$$E_{2k} = \sqrt{\frac{[h_y^2 + \beta(1 - h_y^2)]\ (1 - R_o'R^{-1}R_o) - h_y^2\ [(R_o - r_o) + \beta(C - R_o)]'R^{-1}\ (C - R_o)}{h_y^2[1 + \beta(1 - h_y^2)]\ (1 - R_o'R^{-1}R_o) - \beta h_y^4(C - R_o)'R^{-1}(C - R_o)}} \qquad (15)$$

where

$$\begin{aligned} \beta &= Vh_y^2/(s - 4h_y^2) \\ &= [S\alpha + 4h_y^2(1 - Mh_y^2)]/(S - 4h_y^2) \qquad (16) \end{aligned}$$

It is interesting to observe, as expected, the similarity, in form, of the efficiencies of the two indices I_{1k} and I_{2k} given by (7) and (15) respectively, α being replaced by β when we go from I_{1k} to I_{2k}. When k=1, (13), (14) and (15) reduce to the expressions given in Narain *et al.* (1977). When phenotypic and genetic correlations are zero or when $R_o = C$ even when these correlations are not zero, E_{2k} reduces to

$$E_{20} = \sqrt{[h_y^2 + \beta(1 - h_y^2)]/h_y^2[1 + \beta(1 - h_y^2)]} \qquad (17)$$

the corresponding efficiency of Osborne index. When only genetic correlations are zero, the efficiency is given by (9) with α replaced by β. But when we consider the limiting case with n and m tending to infinity, β also tends to unity, as does α. The limting values of I_{2k}, therefore, remain the same as those for I_1, given by (11) & (12). The remark already made on the behaviour of I_{1k}, also apply for I_{2k}.

Numerical Results

The efficiencies of the indices I_{12} and I_{22} for variations in the values of h_y^2 for various combinations of R_0 and r_0 assumed to be the same for both the auxiliary traits which are further assumed to be uncorrelated, are shown in Fig. 1 when $h_X^2 = 0.5$, m=3, n=20 and $\sigma_Y = \sigma_X$. Out of the four curves shown for either of the two indices, one relates to the case when both the genetic as well as phenotypic correlations between the main traits and the two auxiliary traits are zero.

It is apparent from this Figure that the inclusion of two additional traits, on an individual basis, increases the efficiency of the indices at all values of h_y^2. As has been stated earlier, the maximum gain in the efficiency is expected when R_0 and r_0 are of opposite signs. The curves drawn in the figure further show that a negative value of R_0 has greater influence than a negative of r_0. Even when the genetic correlation is zero, there is gain in the efficiency provided the phenotypic correlation is non-zero. A comparison of the efficiencies of these indices with that of I_{10} and I_{20}, the Osborne's indices (which corresponds to the case $R_0 = r_0 = 0$) indicates that while the later are only useful for low values of h_y^2, the new indices developed here could be useful even at higher values of h_y^2.

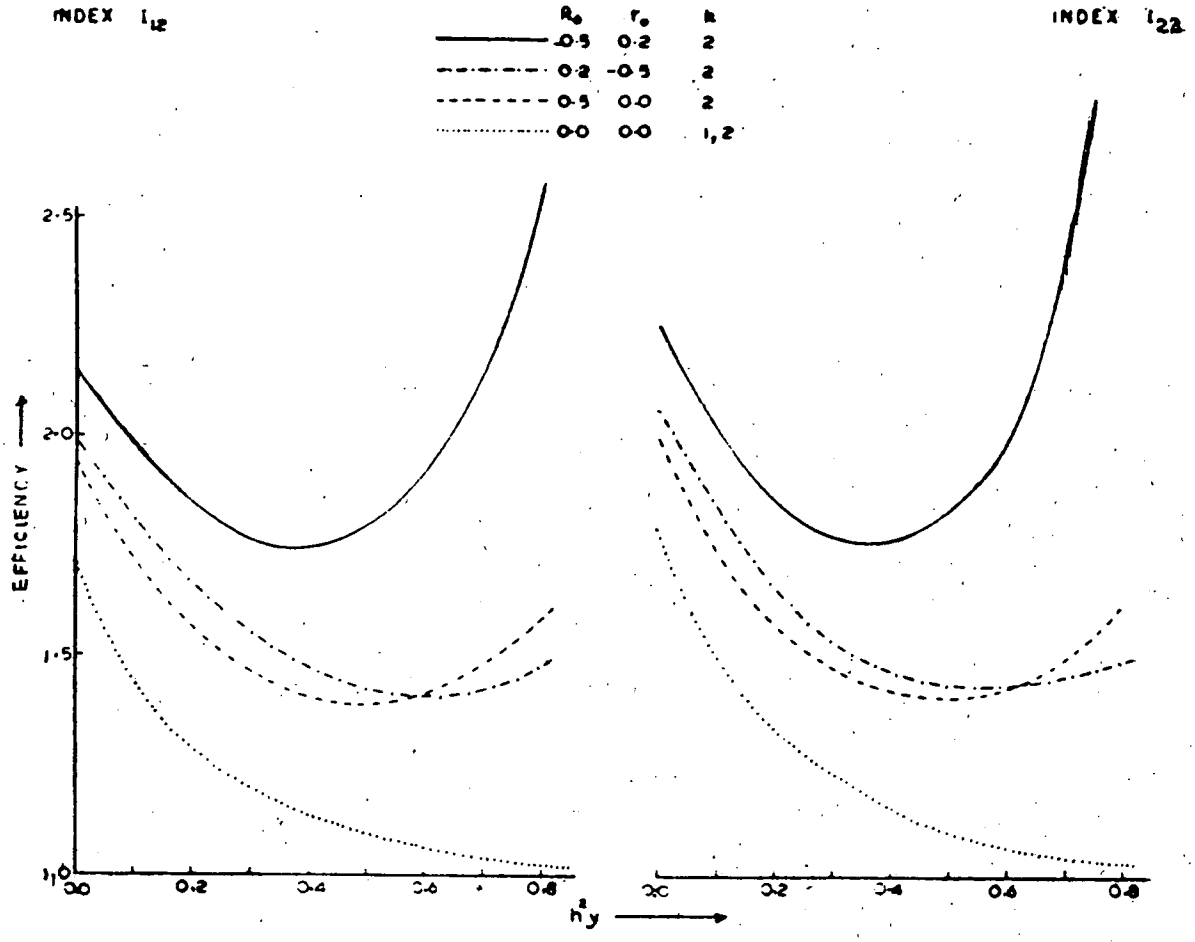

Fig. 1. Efficiency of the selection indices for variations in h_y^2 with $h_x^2=0.5$; m=3; n= 20 and $\sigma_y=\sigma_x$.

In order to compare the effect of including two auxiliary traits as compared to one auxiliary trait, the efficiencies of the two indices for variations in h_y^2 for the combination $R_o=-0.5$ and $r_o=0.2$, assumed to be the same for both the auxiliary traits, uncorrelated among themselves, for k=2 are shown in Fig. 2 when $h_x^2=0.5$, m=3, n=20 and $\sigma_y=\sigma_x$. There are three curves shown, one relating to $R_o=r_o=0$ corresponding to Osborne's index, being shown for comparison purposes.

It is apparent from these figures that inclusion of one more auxiliary trait improves the efficiency. The gain in efficiency appears to be rather high for higher values of h_y^2. The figures also demonstrate the finding, already reported in Narain *et al.* (1977) that the inclusion of the auxiliary trait improves the efficiency for either of the two indices for all value of h_y^2.

In the above cases, the two auxiliary traits are assumed to be mutually uncorrelated. However, the effect of a positive or a negative correlation (P) between the two auxiliary traits on the efficiency has also been studied. The results are graphically presented in Fig. 3 where the variation in the efficiency with the variation in h_y^2 for the combination $R_o=0.2$ and $r_o=-0.5$, assumed to be the same for both the auxiliary traits and $h_x^2=0.5$, m=3, n=20, $\sigma_y=\sigma_x$ has been shown.

There are three figures, besides that for $R_o=r_o=0$, which correspond to the correlations between auxiliary traits of—0.5, 0 and 0.5 respectively. It is quite clear that a positive correlation between the auxiliary traits increases the efficiency whereas a negative correlation decreases it compared to the uncorrelated case for all values of h_y^2 and for either of the two indices. It is further apparent that although the efficiency gets reduced for a negative correlation between the auxiliary traits, it is still higher than efficiency of the Osborne's index for all values of h_y^2. Further, all the three curves appear to show a minimum around a value of $h_y^2=0.6$. This finding is in contrast with Osborne's indices where the efficiency continues to decrease as we move to higher values of h_y^2.

A Practical Example from Poultry

To demonstrate the practical relevance of the theoretical results obtained above, data on poultry involving egg production upto 240 days of age as the main trait and age at first egg and egg weight as the two auxiliary traits, were analysed. The data pertained to the IASRI project entitled "Statistical methodology for developing efficient selection procedures in poultry breeding" already mentioned in the Introduction. The data pertained to 50 sires, each sire being mated to 12 dams and from each mating, 7 to 8 hatches were taken. The hatches were subsequently raised at 6 to 7 locations, of which two were at the Poultry Farm at Bhopal. Observations were made on age at first egg (x_1), average egg weight (x_2) and total egg production upto 240 days of age (y). Since egg production is effected significantly by environmental factors, the data were adjusted for locations and dams effects for each sire separately using the least square technique (Harvey, 1960). The adjusted data were used for estimation of heritability, genetic and phenotypic correlations amongst different characters. The estimates of genetic parameters were obtained by the half-sibs method. The standard errors of the genetic parameters were estimated by the method given by Swinger *et al.* (1964) and quoted in Beckar (1967). The estimates of genetic parameters along with their standard errors for the three traits are given in Table 2. The estimates of heritability for egg production upto 240 days of age is found to be 0.505. The egg production is further found to be negatively correlated, both genetically as well as phenotypically with age at first egg and egg weight.

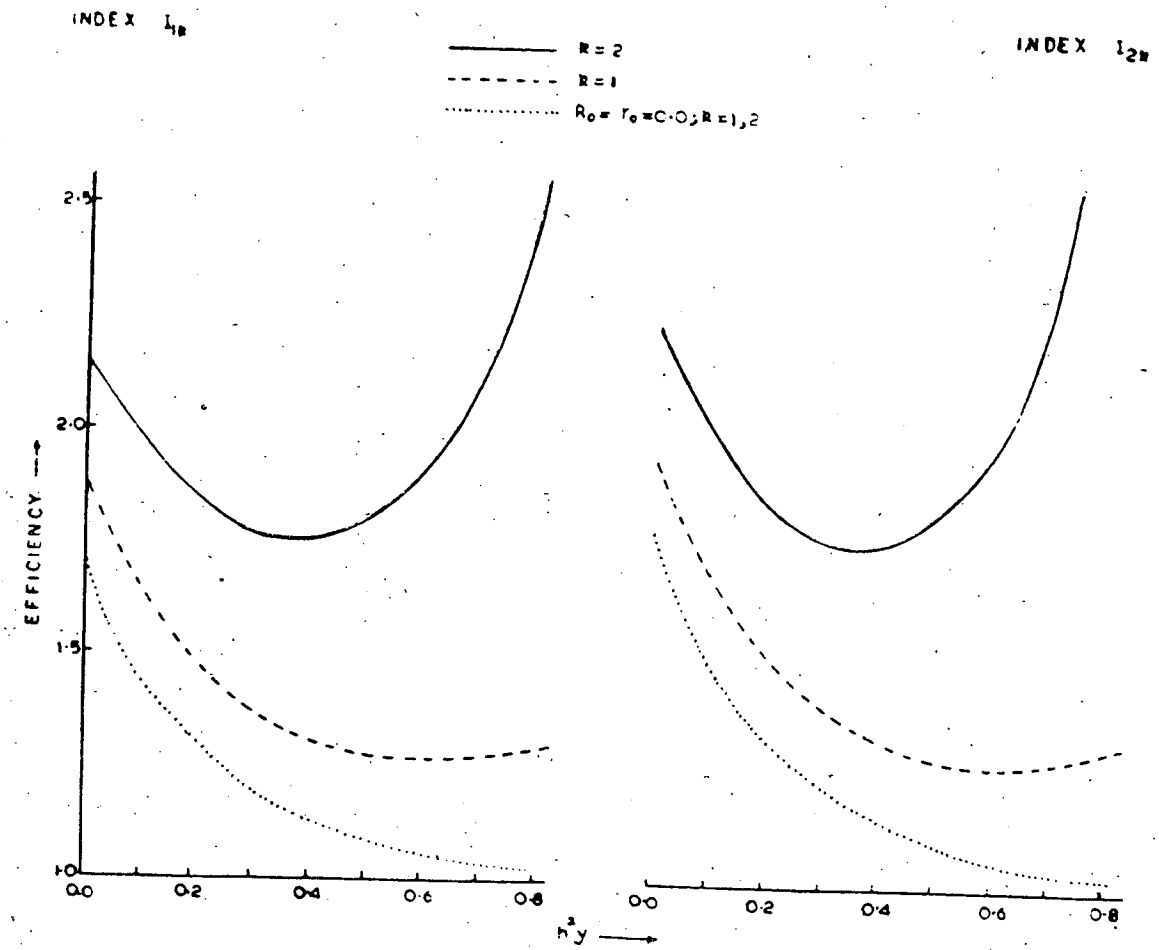

Fig. 2. Effect of the number of auxiliary traits on the efficiency of selection indices with $R_o = -0.5$; $r_o = 0.2$; $h_x^2 = 0.5$; $m = 3$; $n = 20$ and $\sigma_y = \sigma_x$.

Table 2. Estimates of phenotypic and genotypic parameters

	Age at first egg	Egg weight	Egg production up to 240 days of age
	x_1	x_2	y
x_1	0.502	0.383	—0.799
	(0.109)	(0.127)	(0.064)
x_2	0.080	0.833	—0.547
	(0.024)	(0.158)	(0.102)
y	—0.263	—0.117	0.505
	(0.029)	(0.029)	(0.109)

Diagonal values are heritability estimates; values above and below diagonal are genetic (r_o) and phenotypic (R_o) correlations respectively. Figures in parenthesis indicate standard errors of the estimates of phenotypic and genotypic parameters.

Using the above estimates, the following eight selection indices were developed:

(i) Osborne's index (I_{10});

(ii) Osborne's index with dam's performance (I_{20});

(iii) New index with one auxiliary trait I_{11} (x_1);

(iv) New index with one auxiliary trait x_1 including dam's performance I_{21} (x_1);

(v) New index with one auxiliary trait x_2, I_{11} (x_2);

(vi) New index with one auxiliary trait x_2, including dam's performance I_{21} (x_2);

(vii) New index with two auxiliary traits x_1 and x_2, I_{12} (x_1, x_2);

(viii) New index with two auxiliary traits x_1, x_2, including dam's performance I_{22} (x_1, x_2).

The estimates of the coefficients of the eight selection indices are presented in the Table 3 along with their efficiencies. The efficiencies

Table 3. Coefficients of selection indices and their efficiencies over Osborne's indices

Index	x_1	x_2	Y_1	Y_2	Y_3	D_Y	E
I_{10}	—	—	0 4209	0.2124	0.2238	—	1.0000
I_{20}	—	—	0.4021	0.2558	0 1761	0.1056	1.0000
I_{11} (x_1)	—0.2542	—	0.3772	0.1547	0.1651	—	1.0909
I_{21} (x_1)	—0.2458	—	0.3650	0.1884	0.1286	0.0774	1.0825
I_{11} (x_2)	—	—0.2653	0.4164	0.1462	0.1562	—	1.0707
I_{21} (x_2)	—	—0.2599	0.4080	0.1254	0.1338	0.0687	1.0607
I_{12} (x_1.x_2)	—0.2594	—0.2706	0.3781	0.0168	0.0995	—	1.1556
I_{22} (x_1.x_2)	—0.2552	—0.2662	0.3719	0.0815	0.0521	0.0457	1.1401

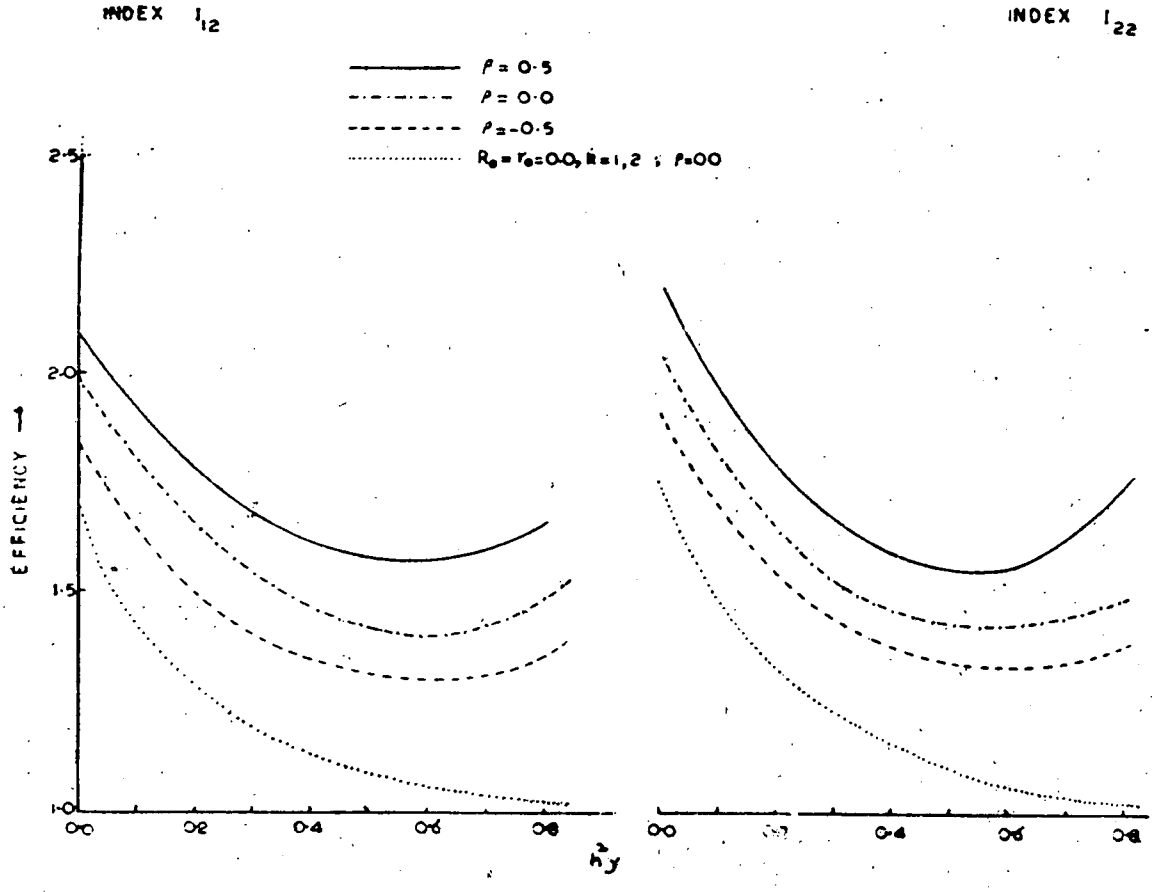

Fig. 3. Effect of the correlation between the auxiliary traits on the efficiency of the selection indices with $R_o = 0.2$; $r_o = -0.5$; $h_x^2 = 0.5$; $m = 3$; $n = 20$ and $\sigma_y = \sigma_x$.

have been worked out taking the corresponding Osborne's index as the standard.

The coefficients of the two auxiliary traits in the index vary only slightly with change in combination of auxiliary traits. The same is also true for egg production upto 240 days based on individual's performance. The inclusion of individual's dams performance has decreased the coefficients for the character under improvement (y_1) and for the full-sib family average (y_3) whereas it has increased for half-sib family averages (y_2) in all the cases under study except for the case when only egg weight is used as auxiliary trait. The efficiencies of the new indices with different combinations of auxiliary traits are more than that of the usual Osborne's indices (I_{10}, I_{20}). The gain in efficiency is of the order of 9% and 8% respectively when x_1 is included as auxiliary trait. When x_2 is included as auxiliary trait, the gain in efficiency is 7% and 6%, respectively. However, when both the auxiliary traits x_1 and x_2 are included in the index, the gain in efficiency is more being respectively 15% and 14%. It is also significant to note that, in all the cases, the inclusion of dam's performance has decreased the efficiency of new index by about 1%.

References

Beckar, W. A. 1967. "Manual of procedures in Quantitative Genetics". Washington State University Press, Washington.

Harvey, W. R. 1960. Least square analysis of data with unequal sub-class numbers. U. S. D. A., A. R. S. *20-8* pp 1-157.

Narain, P. and Mishra, A. K. 1975. Efficiency of selective breeding based on a phenotypic index. J. Genet. *62* (*2*): 69-76.

Narain P., Garg, L. K. and Jain, J. P. 1977. The effect of including individual's egg weight in a selection index for rate of lay in chickens. Indian Poultry Review *8*: 17-23.

Narain P., Garg, L. K., Jain, J. P., Puri, J. C., Lal, P. and Rana, P. S. 1979. "Statistical methodology for developing efficient selection procedures in poultry breeding. Technical Report, I. A. S. R. I., New Delhi.

Osborne, R. 1957. The use of sire and dam family averages in increasing the efficiencies of selective breeding under a hierarchical mating system. Heredity *4*: 93-116.

Swinger, L. A., Harvey, W. R., Everson, D. O. and Gregory, K. E. 1964. The variance of intraclass correlation involving groups with one observation. Biometrics *20*: 818-26.

Wright, S. 1921. Systems of mating. Genetics *6*: 124-143.

22

[REPRINTED FROM THE *Indian Journal of Genetics and Plant Breeding*, Vol. 34, No. 3, Nov. 1974. pp. 309–317]

PARTIAL DIALLEL CROSSES BASED ON EXTENDED TRIANGULAR ASSOCIATION SCHEME.

By PREM NARAIN, C. SUBBA RAO and A. K. NIGAM.

PARTIAL DIALLEL CROSSES BASED ON EXTENDED TRIANGULAR ASSOCIATION SCHEME

PREM NARAIN, C. SUBBA RAO and A. K. NIGAM

Institute of Agricultural Research Statistics, Library Avenue, New Delhi-12

(Accepted: 22-v-74)

IN a diallel cross, with no reciprocal crosses and parental inbreds, there are n(n-1)/2 possible single crosses among a set of n inbred lines. The number of possible crosses increases rapidly with n. If following Kempthorne (1956) only a few lines are sampled, estimates of the variance of g.c.a. effects will be subject to a very large sampling error. It is therefore desirable to perform only a sample of all the possible crosses among a large number of inbred lines. Kempthorne and Curnow (1961), Fyfe and Gilbert (1963) and Curnow (1963) dealt with the advantages of such a sampling in diallel crosses and termed it 'partial diallel'.

There is a one-to-one correspondence between the full diallel crosses and balanced incomplete block (BIB) designs with two plots per block. Similarly, a correspondence exists between the partial diallel crosses and partially balanced incomplete block (PBIB) designs with two plots per block and two associate classes. Various type of partial diallel crosses can, therefore, be constructed with the help of PBIB designs. For instance, Kempthorne and Curnow (1961) gave partial diallel crosses based on the circulant design in which the number of inbred lines and the number of times each line is involved in the sampled crosses could be odd and even or even and odd respectively. Fyfe and Gilbert (1963) constructed such crosses with the help of 'triangular' designs in which the number of lines could be of the form p(p-1)/2 where p is an integer. In the present investigation, partial diallel crosses have been constructed and analysed when the number of inbred lines is of the form p(p-1)(p-2)/6, where p is an integer greater than 3.

CONSTRUCTION

Let the number of parents n be of the form p(p-1)(p-2)/6 where p is an integer greater than 3. Now denote a parent by a triplet abc, where a takes any value from 3 to p, b takes values from 2 to (a-1) and c takes values from 1 to (b-1). All the parents can then be numbered off into (p-2) different triangles, T_1, T_2, ..., $T_{(p-4)}$, $T_{(p-3)}$ and $T_{(p-2)}$ of orders $(p-2)\times(p-2)$, $(p-3)\times(p-3)$, ... 2×2 and 1×1 respectively. For instance, with $n=35$ and $p=7$ we obtain triangles T_1, T_2, T_3, T_4 and T_5 by numbering as shown in Table 1.

TABLE 1

Numbering off the parents into triangles for constructing ETD with $n=35$, $p=7$

Triangle	Order of triangle	Parents (triplets)	Number of triplets
T_1	5 × 5	765, 764, 763, 762, 761 754, 753, 752, 751 743, 742, 741 732, 731 721	15
T_2	4 × 4	654, 653, 652, 651 643, 642, 641 632, 631 621	10
T_3	3 × 3	543, 542, 541 532, 531 521	6
T_4	2 × 2	432, 431 421	3
T_5	1 × 1	321	1
		Total number of parents	35

The number of parents (triplets) in i^{th} triangle is (p-i)(p-i-1)/2, for i=1, 2, ... (p-2), so that adding over the triangles we get the total number of parents. This is easily verified for the given example from Table 1. We can now construct three different types of partial diallel crosses according to the three designs given below.

Design I: We sample all the crosses of type abc × def, where a, b, c, d, e, and f are all distinct. As such the parent p(p-1)(p-2), 765, for example, of triangle T_1, cannot be crossed with any parent lying in triangles T_1, T_2 and T_3. We can cross it with all the parents which are in the remaining triangles i.e. T_4, T_5, ... and $T_{(p-2)}$. Hence the number of times this parent is involved in crossing with other parents, denoted by s_1, is given by

$$s_1 = \sum_{i=4}^{(p-2)} (p-i)(p-i-1)/2 = (p-3)(p-4)(p-5) \quad /6$$

It can be easily seen that every other parent would be involved in crossing with other parents exactly s_1 times. The resulting sample would consist of $ns_1/2$ crosses. In the example cited above, it is easy to see that $s_1=4$. The sampling plan of this design is given in Table 2.

TABLE 2

Crosses to be sampled for ETD, Design I with $n=35$, $s=4$

1 ×	32,	33,	34,	35
2 ×	29,	30,	31,	35
3 ×	27,	28,	31,	34
4 ×	26,	28,	30,	33
5 ×	26,	27,	29,	32
6 ×	23,	24,	25,	35
7 ×	21,	22,	25,	34
8 ×	20,	22,	24,	33
9 ×	20,	21,	23,	32
10 ×	18,	19,	25,	31
11 ×	17,	19,	24,	30
12 ×	17,	18,	23,	29
13 ×	16,	19,	22,	28
14 ×	16,	18,	21,	27
15 ×	16,	17,	20,	26
16 ×	35			
17 ×	34			
18 ×	33			
19 ×	32			
20 ×	31			
21 ×	30			
22 ×	29			
23 ×	28			
24 ×	27			
25 ×	26			

The above procedure corresponds to picking up the third associates of each treatment in the extended form of triangular association scheme given by John (1966) and pairing the treatment with each member of the associate class. The number of third associates is s_i whereas the number of secondary parameters p^3_{jk} (j, k=1, 2, 3) are exhibited in the form of a symmetric matrix $\mathbf{P}_3$ given by

$$\mathbf{P}_3 = \begin{pmatrix} 0 & 9 & 3(p-6) \\ & 9(p-6) & 3(p-6)(p-7)/2 \\ & & (p-6)(p-7)(p-8)/6 \end{pmatrix}$$

For the example considered with p=7, the only non-zero elements of this matrix are $p^3_{12}=p^3_{21}=p^3_{22}=9$ and $p^3_{13}=p^3_{31}=3$.

Design II: We sample all the crosses of type abc × def, where one of the letters (a, b, c, d, e and f) is common. There will then be there categories of crosses: (i) abc × aef, abc × daf, abc × dea (ii) abc × bef abc × dbf abc × deb (iii) abc × cef, abc × dcf, abc × dec.

Since crosses involve only one letter in common, the parent p(p-1) (p-2)

cannot be crossed with parents which are in the first and second rows of T_1 and in the first row of T_2. We can therefore cross the parent p(p-1)(p-2) with (i) the remaining parents of T_1 which is equal to the number of parents in T_3 (ii) the remaining parents of T_2 which is also equal to the number of parents in T_3, and (iii) all the parents in T_3. Hence the number of times each parent is involved in crossing with the other parents, denoted by s_2, is given by

$$s_2 = 3(p-3)(p-4)/2$$

The resulting sample would consist of $ns_2/2$ crosses. In the example given above, $s_2 = 18$.

In terms of the association scheme, s_2 represents the number of second associates of each treatment whereas the symmetric matrix $\mathbf{P}_2$ giving the secondary parameters p^2_{jk} (j, k = 1, 2, 3) is represented by

$$\mathbf{P}_2 = \begin{pmatrix} 4 & 2(p-4) & (p-5) \\ & (p-5)(p+2)/2 & (p-5)(p-6) \\ & & (p-5)(p-6)(p-7)/2 \end{pmatrix}$$

In the example taken for illustration, the non-zero elements are $p^2_{11} = 4$, $p^2_{12} = p^2_{21} = 6$, $p^2_{13} = p^2_{31} = p^2_{23} = p^2_{32} = 2$ and $p^2_{22} = 9$.

Design III: We sample all the crosses of type abc × def where two of the letters a, b, c, d, e and f are in common. There will then be three categories of crosses given by (i) abc × abf, abc × aeb, abc × dab. (ii) abc × bcf, abc × bec, abc × dbc. (iii) abc × acf, abc × aec, abc × dac.

Since we exclude the crosses of type abc × abc i.e. selfs, and since two letters are distinct, the parent p(p-1)(p-2) can be crossed with (i) the parents which are in the first and second rows of T_1 and (ii) the parents which are in the first row of T_2. Hence, in this design, the number of times each parent is involved in crossing with the other parents, would be $s_3 = 3(p-3)$ and the number of crosses sampled would be $ns_3/2$. In the example chosen for illustration, $s_3 = 12$.

Referring to the association scheme, it is found that each parent gets crossed with its s_3 first associates. The symmetric matric $\mathbf{P}_1$ of secondary parameters p'_{jk} (j, k = 1, 2, 3) is given by

$$\mathbf{P}_1 = \begin{pmatrix} (p-2) & 2(p-4) & 0 \\ & (p-4)^2 & (p-4)(p-5)/2 \\ & & (p-4)(p-5)(p-6)/6 \end{pmatrix}$$

For the example cited, the non-zero elements of the matrix are $p'_{11} = 5$, $p'_{12} = p'_{21} = 6$, $p'_{22} = 9$, $p'_{23} = p'_{32} = 3$ and $p'_{33} = 1$.

It is apparent, from the above construction, that s as well as the number of secondary parameters are functions of p in all the three designs. These three designs when put together represent the complete set of diallel crosses with no reciprocals and selfs.

ANALYSIS

The analysis of partial diallel crosses constructed above follows the

pattern of the analysis of three-associate PBIB designs given by Rao (1947) and subsequently used by Das and Sivaram (1968).

The mean yield of the cross between i^{th} and j^{th} parent is expressed as

$$\bar{Y}_{ij}=\mu+t_i+t_j+s_{ij}+\bar{e}_{ij}$$

where μ is the effect due to overall mean, t_i and t_j are the g.c.a. effects due to i^{th} and j^{th} parents respectively, s_{ij} is the s.c.a. effect due to the cross ixj and $\bar{e}_{ij}$ is the random error. We assume that $\sum_{i=1}^{n} t_i=0$, $\sum_{j=1}^{n} s_{ij}=0$ for each i and that t_j, s_{ij} and $\bar{e}_{ij}$ are each independently normally distributed with zero means and variances, σ^2_g, σ^2_s, σ^2_e respectively.

Design I: Since the third associate classes are sampled to provide the plan for the partial diallel cross, the normal equations, on the above model, for estimating the g.c.a. effects would be for i=1, 2,....n

$$s_1\mu+s_1t_i+S_3(t_i)=T_i$$
$$s^2_1\mu+s_1t_i+s_1S_3(t_i)+p^3_{33}S_3(t_i)+p^2_{33}S_2(t_i)+p^1_{33}S_1(t_i)=S_3(T_i)$$
$$s_1s_2\mu+s_1S_3(t_i)+p^3_{32}S_3(t_i)+p^2_{32}S_2(t_i)+p^1_{32}S_1(t_i)=S_2(T_i)$$

where $S_3(t_i)$ is the sum of the g.c.a. effects of parents (third associates) with which the i^{th} parent is crossed, $S_2(t_i)$ and $S_1(t_i)$ are the sums of the g.c.a. effects of parents (second and first associates respectively) not crossed with the i^{th} parent. Also T_i is the total yield of all the crosses involving the i^{th} parent, $S_3(T_i)$ is the sum of the totals T_i's of those parents (third associates) with which the i^{th} parent is crossed and $S_2(T_i)$ is the sum of the totals T_i's of parents (second associates) not crossed with the i^{th} parent.

The least squares estimate of g.c.a. effect of i^{th} parent is then given by

$$\hat{t}_i=[(A_2B_3-A_3B_2)T_i-B_3S_3(t_i)+A_3S_2(T_i)-(2G/n)A_2B_3-A_3B_2-s_1B_3+s_1A_3)]/\Delta$$

where G is the grand total of the yield of all the crosses and

$$A_1=(p^2-9p+20)/2=(p-4)(p-5)/2$$
$$A_2=(p^3-18p^2+119p-276)/6$$
$$A_3=-(p^2-11p+30)/2=-(p-5)(p-6)/2$$
$$B_1=-(p^2-9p+20)/2=-(p-4)(p-5)/2$$
$$B_2=(p^2-15p+53)$$
$$B_3=(p^3-9p^2+8p+60)/6$$
$$\Delta=s_1(A_2B_3-A_3B_2)-(A_1B_3-A_3B_1)$$

The sum of squares due to g.c.a. effects of the parents is $\sum_{i=1}^{n} \hat{t}_i T_i$ whereas the sum of squares due to s.c.a. effects of the pair of parents crossed is (sum of squares due to crosses $-\sum_{i=1}^{n} \hat{t}_i T_i$). The significance of g.c.a. effects is found by

testing the mean sum of squares due to g.c.a. effects against the error mean sum of squares obtained in a replicated experiment. In the model given above, $\bar{e}_{ij}$ is having expectation zero and variance σ^2_e/r, where r is the number of replicates in the experiment. $\bar{Y}_{ij}$ is then an unbiased estimate of $\mu+t_i+t_j$ with variance $\sigma^2=(\sigma^2_s+\sigma^2_e/r)$.

The variance of the difference between g.c.a. effects of two parents would be of three types:

$V_1=V(t_i-t_j)=2\sigma^2\ (A_2B_3-A_3B_2+B_3)/\Delta$, where j^{th} parent is crossed with i^{th} parent, $V_2=V(t_i-t_k)=2^2\sigma\ (A_2B_3-A_3B_2+A_3)/\Delta$, where j^{th} parent is not crossed with i^{th} parent but is crossed with a parent with which i^{th} parent is crossed. and $V_3=V(t_i-t_l)=2^2\sigma\ (A_2B_3-A_3B_2)/\Delta$ where l^{th} parent is neither crossed with i^{th} parent nor with a parent with which i^{th} parent is crossed.

The average variance of the difference between the g.c.a. effects of any two parents is then given by

$$D_1=\frac{s_1V_1+s_2V_2+s_3V_3}{s_1+s_2+s_3}=\frac{2\sigma^2}{\Delta}\left[(A_2B_3-A_3B_2)+\frac{s_1B_3+s_2A_3}{s_1+s_2+s_3}\right]$$

Design II: A similar procedure as described in Design I is followed to obtain the least squares estimates of g.c.a. effects of the parents when second associate classes are sampled to provide the partial diallel crosses. The average variance of the difference between g.c.a. effects of any two parents is now given by

$$D_2=\frac{2\sigma^2}{\Delta'}\left[(A'_2B'_3-A'_3B'_2)+\frac{s_1B_3'+s_2A_3'}{{}_1+s_2+s_3}\right]$$

where $A'_1=3\ (p^2-13p+48)/2$, $A'_2=(2p^2-21p+67)$, $A'_3=(p^2-17p+70)$, $B_1=-9$, $B_2=2p-17$, $B_3=(3p^2-17p+2)/2$, $\Delta'=s_2\ (A'_2B'_3-A'_3B'_2)-(A'_1B'_3-A'_3B'_1)$

Design III: A similar procedure gives the average variance of the difference between g.c.a. effects of any two parents as

$$D_3=\frac{2\sigma^2}{\Delta''}\left[(A''_2B''_3-A''_3{}''B''_2)+\frac{s_1B_3''+s_2A_3''}{s_1+s_2+s_3}\right]$$

where $A''_1=(3p-9)$; $A''_2=(4p-11)$; $A''_3=4$; $B''_1=-9$; $B''_2=2p-17$; $B''_3=5p-26$; $\Delta''=s_1(A''_2B''_3-A''_3B''_2)-(A''_1B''_3-A''_3B''_1)$.

EFFICIENCY

With the help of the expressions in previous section the average variances of the difference between the g.c.a. effects of any two parents can be worked out for the three designs. Since these expressions are ultimately reducible as functions of p, an idea of the average variance given by the three-variance samples described in this paper can be obtained by giving different values to p. We give values to p ranging between 6 to 11. The average variances for the three designs for p lying between this range are shown in Table 3.

TABLE 3

Average variance of the difference between two g.c.a. effects per unit σ^2 for partial diallel crosses based on ETD

n	p	Design I		Design II		Design III	
		s_1	D_1	s_2	D_2	s_3	D_3
20	6	1	—	9	0·263	9	0·277
35	7	4	0·725	18	0·116	12	0·185
56	8	10	0·222	30	0·068	15	0·137
84	9	20	0·107	45	0·045	18	0·117
120	10	35	0·058	63	0·032	21	0·100
165	11	56	0·036	84	0·024	24	0·087

The designs considered in this paper are such that each of the n parents are involved in the same number of crosses viz. either s_1 or s_2 or s_3 giving the total number of crosses sampled as $ns_1/2$ or $ns_2/2$ or $ns_3/2$. This shows that s_1 (or s_2 or s_3) must be at least 2 and n and s_1 (or s_2 or s_3) cannot both be odd.

This is satisfied for the cases shown in Table 3 except for n=20 with Design I giving $s_1=1$ which is, therefore, not admissible. We do not, therefore, give a value for the average variance in this case.

It is apparent from Table 3 that the average variance decreases with increase in the value of s_1 (or s_2 or s_3). The dependence of the average variance on **s** is so much that even for different values of n and irrespective of the design, its value remains same if s does not vary. For instance, for $s_2=18$, $n=35$, the value is almost the same as that for $s_3=18$, $n=84$. Of course, the number of crosses to be sampled in the latter case will be much more than in the former case.

In order to judge the efficiency of extended triangular designs (ETD) discussed in this paper vis-a-vis CD of Kempthorne and Curnow (1961) we can compare, for the same number of crosses sampled, the average variances in the two cases.

For the specific cases considered in Table 4, the efficiencies of ETD relative to CD were worked out with the help of the values of D_1, D_2 and D_3 taken from Table 3.

It can be seen that the efficiency of ETD is always greater than one and ETD Design I is definitely much more efficient than the other two ETD designs. For n=35, ETD corresponding to Design III is 14% more efficient but that corresponding to Deisgn II is only 5% more efficient than CD.

Curnow (1963) has enumerated in Table 1 of his paper, the two-variance samples from the diallel cross by reference to tabulation on Clatworthy

TABLE 4

Efficiency of Extended Triangular Designs (ETD) relative to circulant Design (CD) in estimating g.c.a. effects

			CD	ETD	E
n=20	s_1	1	—	—	—
	s_2	9	0·522	0·263	1·986
	s_3	9	0·522	0·277	1·886
n=35	s_1	4	2·101	0·725	2·897
	s_2	18	0·121	0·116	1·046
	s_3	12	0·210	0·185	1·136

(1955) of the corresponding PBIB designs. For n=56 and s=10, there exists a two-variance sample listed at 5·8 in this table with average variance as 0·2273 σ^2. This can therefore be compared with the ETD corresponding to Design I with n=56, p=8 and $s_1=10$ given in Table 3. The average variance is now 0·222σ^2. The three-variance sample provided by ETD is therefore 2% more efficient than the two-variance sample given by Curnow (1963). Although CD of Kempthorne and Curnow (1961) does not exist for n=56, we can consider the case nearest to it i.e. with n=51, s=10 given in his paper. This gives an average variance of 0·3276σ^2, about 68% greater than 0·222σ^2.

From the above considerations, we conclude that partial diallel crosses based on triangular association scheme and involving crossing of third associates (Design I of this paper) are more efficient than the partial diallel crosses based on circulant design of Kempthorne and Curnow (1961). Of course, such partial diallel crosses are not available for all values of n. These are normally, available for larger values of n such as 20, 35, 56, 84, 120, 165 etc. as given in column under Design I of Table 3. It is obvious also from this table that the larger the value of n, the smaller is the average variance of the comparison between two g.c.a.'s. This is in contrast with the CD of Kempthorne and Curnow (1961) where with increase in the value of n, the average variance of the comparison between two g.c.a.'s increases. Partial diallel crosses based on triangular designs given by Fyfe and Gilbert (1963) are suited only for smaller values of n such as 28, 21, 15, 10 etc. but for larger values it would always pay to take recourse to extended triangular designs discussed in this paper.

SUMMARY

Procedures of constructing and analysing partial diallel crosses based on extended triangular association scheme (ETD) where the number of parental

lines is of the form $p(p-1)(p-2)/6$ with p as an integer greater than 3 have been discussed. It has been found that these plans for partial diallel crosses are clearly more efficient than those based on circulant Designs (CD) for the same number of crosses sampled.

REFERENCES

Clatworthy, W. H. (1955). Partially balanced incomplete block designs with two associate classes and two treatments per block. *J. Res. nat. Bur. Standards*, **54**: 177–90.

Curnow, R. N. (1963). Sampling the diallel crosses. *Biometrics*, **19**: 287–306.

Das, M. N. and Sivaram, K. (1968). Partial diallel crosses and incomplete block designs. Contributions in Statistics and Agri. Sciences. Vol. presented to Dr. V. G. Panse at his 62nd birthday: 49–59.

Fyfe, J. L. and Gilbert, N. E. G. (1963). Partial diallel crosses. *Biometrics*, **19**: 278–86.

John, P. W. M. (1966). An extension of the triangular association scheme to three associate classes. *J. roy. Stat. Soc.*, **B28**: 361–65.

Kempthorne, O. and Curnow, R. N. (1961). The partial diallel crosses. *Biometrics*, **17**: 229–50.

Rao, C. R. (1947). General methods of analysis for incomplete block designs. *J Amer. Stat. Assoc.*, **42**: 541–61.

Sankhyā : The Indian Journal of Statistics
1981, Volume 43, Series B, Pt. 1, pp. 93–103.

23

TRUNCATED TRIANGULAR ASSOCIATION SCHEME AND RELATED PARTIAL DIALLEL CROSSES

By PREM NARAIN and A. S. ARYA*

Indian Agricultural Statistics Research Institute, New Delhi

SUMMARY. A new association scheme called truncated triangular (TT) with five associate classes when $v = p(p-2)/2$ with p an even positive integer $\geqslant 8$, has been discussed and used to construct partial diallel crosses (PDC). The analysis of the design is based on the characteristic roots and idempotent matrices of the matrix $A = NN'$ where N is the incidence matrix of the design. Of the three designs for PDC constructed with the help of TT association scheme, Design I is found to be consistently more efficient than the other two.

1. Introduction

A considerable amount of work has been done by various workers studying two-class association schemes and related partially balanced incomplete block (PBIB) designs. Schemes of higher classes have also been studied, for example, generalisation of two-associate class group-divisible to m-associate class by Roy (1953-54) and Raghavarao (1960), extended group-divisible PBIB (EGD/m-PBIB) of Hinkelmann (1964), three-class cubic association scheme of Raghavarao and Chandrasekhararao (1964), hyper-cubic association scheme of Kusumoto (1965), extended triangular association scheme of John (1966), m-class triangular association scheme of Ogasawara (1965), right-angular association scheme and generalised right-angular association scheme having four classes of Tharthare (1963, 65) and generalisation of the two-associate cyclical association scheme by Adhikary (1966, 67). The PBIB designs with two-associate classes have been used by workers such as Kempthorne and Curnow (1961), Fyfe and Gilbert (1963), Curnow (1963) and others for construction and analysis of partial diallel crosses (PDC). Narain *et al* (1974) used extended triangular association scheme with three associate class for this purpose. Arya and Narain (1977) have further used group-divisible designs with three-associate classes, simple rectangular lattice with four-associate classes and generalised right angular scheme with four associate classes for this purpose. In this paper, a new scheme called truncated triangular (TT) associate scheme with five associate classes given by Kishen and Shukla (1974) is used for constructing the PDC. We also give the method of characteristic roots and idempotent matrices for its analysis.

*R. K. College, Shamli, U.P.

2. Construction

Consider n inbred lines which have been randomly numbered from 1 to n. Each line will be made into crosses with s other lines. This, then, leads to a partial set of $ns/2$ out of the $n(n-1)/2$ possible single crosses among different lines. Clearly n and s both cannot be odd and s should be at least two. Following Kishen and Shukla (1974), let $n = p(p-2)/2$, where p is an even integer greater than 6. The truncated triangular scheme is as follows :

An association scheme with five associate classes, for $v = p(p-2)/2$ symbols arranged in a square array of p rows and p columns is said to be triangular association scheme if it satisfies the following properties :

(i) The positions in the principal diagonal (running from the upper left hand to the lower right hand corner) as well as in the other diagonal (running from the upper right hand to the lower left hand corner) are left blank.

(ii) The $p(p-2)/2$ positions above the principal diagonal are filled by the numbers 1, 2,... $p(p-2)/2$, corresponding to the symbols.

(iii) The $p(p-2)/2$ positions below the principal diagonal are filled so that the array is symmetrical about the principal diagonal.

(iv) For any symbol i, the first associates are exactly those that occur in the same row or in the same column as i, except those two symbols i' and i'' which occupy the same position as i with respect to the other diagonal (when positions above this diagonal are filled with the symbols and positions below this diagonal are filled so that the array is symmetrical about this diagonal).

(v) The symbols i' and i'' are the second associates of i.

(vi) The symbols occurring in the same column as i' or in the same column as i'' except the common symbol i''' between the two cases are exactly the third associate of i.

(vii) The symnol i''' is the fourth associate of i.

(viii) The remaining symbols are the fifth associates of i.

Clearly, the parameters of this scheme are

$$v = p(p-2)/2, \quad n_1 = 2(p-4),$$

$$n_2 = 2, \quad n_3 = 2(p-4)$$

$$n_4 = 1, \quad n_5 = (p-4)(p-6)/2.$$

Three designs of PDC can now be constructed with the help of the above scheme. In Design I, we grow all the crosses in which a line is crossed with all other lines falling in its first associate class. This means each line will be involved into crosses with $2(p-4)$ other lines. We denote this number by s_1, the suffix of s refers to the design number. The total number of partial crosses raised is, thus $ns_1/2 = p(p-2)(p-4)/2$. In Design II we sample the crosses of the type $i \times j$ where i and j are third associates. Here also $s_2 = 2(p-4)$. In Design III we pick up the crosses of the type involving lines which are fifth associates. For this PDC $s_3 = (p-4)(p-6)/2$ and p should exceed 8. Two more designs with $s = 2$ and $s = 1$ corresponding to 2nd and 4th associates are possible but they lead to singular least squares equations and hence are not discussed.

For illustration, we take $n = 24$ with $p = 8$. The 24 numbers are written off as below :

X	1	2	3	4	5	6	X
1	X	7	8	9	10	X	11
2	7	X	12	13	X	14	15
3	8	12	X	X	16	17	18
4	9	13	X	X	19	20	21
5	10	X	16	19	X	22	23
6	X	14	17	20	22	X	24
X	11	15	18	21	23	24	X

Now all the 24 numbers are classified into five associate classes with respect to a given number. For instance, if we take 1 and 16 the corresponding five associate classes are :

associate class	1	16
1st	2, 3, 4, 5, 7, 8, 9, 10	3, 5, 8, 10, 17, 18, 22, 23
2nd	6, 11	12, 19
3rd	14, 15, 17, 18, 20, 21, 22, 23	2, 4, 7, 9, 14, 15, 20, 21
4th	24	13
5th	12, 13, 16, 19	1, 6, 11, 24

Similarly, we get the associate classes for other groups. Clearly then 1 and 16 are crossed with lines in 1st associate group to give Design I. They are crossed with lines in the 3rd associate class for Design II and with lines in 5th associate class for Design III.

3. Analysis

The analysis of PDC's constructed above is carried out by the usual least squares procedure as described in Arya and Narain (1977). In their notations, the normal equations for the estimates g_i's (general combining ability effects) are given by

$$\boldsymbol{A}\hat{\boldsymbol{g}} = \boldsymbol{Q}$$

where $\boldsymbol{A}$ is a $n \times n$ matrix having diagonal elements a_{ii} all equal to s and $a_{ij} = a_{ji} = 1$ if the cross $(i \times j)$ is sampled and 0 otherwise,

$$\hat{\boldsymbol{g}}' = [g_1, g_2, \ldots, g_n],$$
$$\boldsymbol{Q}' = [Q_1, Q_2, \ldots, Q_n],$$

where

$$Q_i = \sum_{j(i)} \bar{y}_{ij} - (2G/n)$$

is the right hand side of the i-th normal equation. The $\sum_{j(i)}$ refers to summing over lines j crossed with line i in the design and G is the total of cross mean yields.

If $\boldsymbol{A}$ is non-singular the estimates are $\hat{\boldsymbol{g}} = \boldsymbol{A}^{-1}\boldsymbol{Q}$, the sum of squares due to the estimates is $\boldsymbol{Q}'\boldsymbol{A}^{-1}\boldsymbol{Q}$, and the dispersion matrix of the estimates is $\boldsymbol{A}^{-1}\sigma^2$, with σ^2 being the variance of $\bar{Y}_{ij}$. The analysis of variance for a replicated partial diallel cross is given in Table 1.

TABLE 1. ANALYSIS OF VARIANCE FOR PDC

source of variation	D.F.	S.S.	expected values of MS(=S.S./D.F)
replicates	$(r-1)$	$\{\sum_l y^2_{..l}/(ns/2)\} - CT$	
g.c.a.	$(n-1)$	$\boldsymbol{Q}'\boldsymbol{A}^{-1}\boldsymbol{Q}$	$\sigma_e^2 + r\sigma_s^2 + [rs(n-2)/(n-1)]\sigma_g$
s.c.a.	$n(s/2-1)$	*	$\sigma_e^2 + r\sigma_s^2$
crosses	$(ns/2-1)$	$\{\sum_{i,j} Y^2_{ij.}/r\} - CT$	
replicate × crosses	$(r-1)(ns/2-1)$	**	σ_e^2
total	$rns/2-1$	$\sum_{i,j,l} Y^2_{ijl} - CT$	

Y_{ijl}, is the yield of (ij)-th cross in l-th replicate ;
CT, the correction term $= G^2/(rns/2)$; $Y_{..l} = \sum_{ij} Y_{ijl}$; $Y_{ij.} = \sum_{l} Y_{ijl}$
*by difference, i.e. (cross SS $-$ g.c.a. SS) ; ** by difference, i.e. (total SS-replicate SS-cross SS),
σ_g^2 is the variance of g.c.a. and σ_s^2 is the variance of s.c.a.

4. Inversion of A

Since the PDC's under discussion are associated with two-plot block PBIB design with TT association scheme with five associate classes where $\lambda_1, \lambda_2, \ldots, \lambda_5$ take values either zero or one, it will be seen that $A = NN'$, where N is the incidence matrix of the corresponding PBIB design. As such, the number of distinct elements in A^{-1} or in the associated idempotent matrices L_i's cannot exceed 6. The same is true with the number of distinct latent roots of $A(= NN')$. The latent roots and idempotent matrices of A for the three designs were worked out in accordance with a general procedure given in Appendix. The latent roots along with multiplicities are presented in Table 2.

TABLE 2. LATENT ROOTS OF A FOR THREE PDCs

Latent root	Design I	Design II	Design III	Multiplicity
θ_0	$4(p-4)$	$4(p-4)$	$(p-4)(p-6)$	1
θ_1	$2(p-6)$	$2(p-6)$	$(p^2-10p+32)/2$	$p(p-6)/8$
θ_2	$2(p-5)$	$2(p-3)$	$(p-4)(p-6)/2$	$p(p-4)/4$
θ_3	$2(p-4)$	$2(p-4)$	$(p-4)(p-6)/2$	$p(p-2)/8$
θ_4	$(3p-16)$	$(3p-16)$	$(p-6)(p-8)/2$	$(p-2)/2$
θ_5	$3(p-4)$	$(p-4)$	$(p-4)(p-6)/2$	$p/2$

Let a^{ij} be the element in the (ij)-th position in A^{-1}. Then $a^{ij} = a^k$ if the lines i and j are k-th associates ($k = 0, 1, 2, 3, 4, 5$), a^0 being in the diagonal position. For the corresponding elements of A, a_k ($k = 0, 1, \ldots, 5$), the idempotent matrices L_i's (L_i being the idempotent matrix corresponding to root θ_i of A) are given in Table 3.

TABLE 3. DISTINCT ELEMENTS OF L_i's

Element	L_0	L_1	L_2	L_3	L_4	L_5
a_0	1	$(p-4)(p-6)$	$(p-4)$	1	$2(p-4)$	2
a_1	1	$-2(p-6)$	-1	0	$(p-8)$	1
a_2	1	$(p-4)(p-6)$	0	-1	$2(p-4)$	0
a_3	1	8	0	0	-8	0
a_4	1	$-2(p-6)$	1	0	$(p-8)$	-1
a_5	1	$(p-4)(p-6)$	$-(p-4)$	1	$2(p-4)$	-2
Divisor	$\frac{p(p-2)}{2}$	$4(p-2)(p-4)$	$2(p-2)$	4	$2p(p-4)$	$2(p-2)$

The elements of $\boldsymbol{A}^{-1}$ are worked out by substituting the corresponding elements of $\boldsymbol{L}_i$'s in the spectral decomposition i.e. $\boldsymbol{A} = \sum_{i=0}^{5} \frac{1}{\theta_i} \boldsymbol{L}_i$. These elements are given in Table 4 for the three designs of PDC.

TABLE 4. DISTINCT ELEMENTS OF A^{-1} FOR THE THREE PDC'S

Element	Design I	Design II	Design III
a^0	$1/s_1+(6p-31)/D$	$1/s_2+(4p-19)/D$	$1/s_3+2p/D$
a^1	$-(6p-31)/D$	$(2p-13)/D$	$2(p-8)/D$
a^2	$3(p-5)/D$	$(p-3)/D$	$2p/D$
a^3	$6(p-5)/(p-6)D$	$2(p-3)/(p-6)D$	$-2p/D$
a^4	$-1/D$	$-(4p-19)/D$	$2(p-8)/D$
a^5	$1/D$	$-(2p-13)/D$	$2p/D$
s	$2(p-4)$	$2(p-4)$	$(p-4)(p-6)/2$
D	$12(p-4)(p-5)(3p-16)$	$4(p-3)(p-4)(3p-16)$	$(p-4)(p-6)(p-8)$ $\times(p^2-10p+32)$

5. Comparison of the estimates

The estimates $\hat{g}_i$'s are most conveniently worked out by the relation $\hat{\boldsymbol{g}} = \boldsymbol{A}^{-1}\boldsymbol{Q}$ by entering a^k from Table 4 in (ij)-th position in $\boldsymbol{A}^{-1}$ if the lines i and j are k-th associates as already pointed out. For the estimated differences $(\hat{g}_i-\hat{g}_j)$ there will, in general, be 5 different variances according as the lines i and j are 1st or 2nd ..., or 5th associates. For this purpose one has to write down the association relationship among the n symbols representing the n lines. Let us denote the appropriate variances for the difference $(\hat{g}_i-\hat{g}_j)$ by V_k if the two lines correspond to k-th associates. Then

$$V_k = 2\sigma^2(a^0-a^k), \quad k = 1, 2, 3, 4, 5.$$

The average variance $(\overline{V})$, which can be taken as a criterion for comparing the efficiency of one design relative to the other, for the same values of n and s, is

$$\overline{V} = \sum_{k=1}^{5} n_k V_k \Big/ \sum_{k=1}^{5} n_k,$$

For the three PDC's these are respectively:

$$\bar{V}_{D_1} = \frac{(18p^4-216p^3+791p^2-712p-480)}{12(n-1)(p-4)(p-5)(3p-16)}\sigma^2$$

$$\bar{V}_{D_2} = \frac{(6p^4-58p^3+163p^2-104p-96)}{4(n-1)(p-3)(p-4)(3p-16)}\sigma^2$$

$$\bar{V}_{D_3} = \frac{2(p^5-20p^4+148p^3-464p^2+400p+256)}{(n-1)(p-4)(p-6)(p-8)(p^2-10p+32)}\sigma^2.$$

It is found that Design I and Design II present a five-variance sample of PDC whereas Design III leads to a three-variance sample. The average variances in σ^2 units for these designs are given in Table 5 for admissible values of n between 24 and 180.

It is apparent from Table 5 that the average variance decreases with increase in the value of s. The average variance does not change much, if s does not change irrespective of design and value of n. For instance, with

TABLE 5. AVERAGE VARIANCE OF THE DIFFERENCE BETWEEN THE g.c.a. EFFECTS PER UNIT σ^2 FOR THE THREE PDC's

No. of lines		Design I		Design II		Design III	
n	p	s_1	$\bar{V}_{D_1}$	s_2	$\bar{V}_{D_2}$	s_3	$\bar{V}_{D_3}$
24	8	8	0.2862	8	0.2978	4	SING
40	10	12	0.1806	12	0.1871	12	0.1955
60	12	16	0.1323	16	0.1363	24	0.0886
84	14	20	0.1044	20	0.1071	40	0.0516
112	16	24	0.0863	24	0.0883	60	0.0343
144	18	28	0.0736	28	0.0750	84	0.0241
180	20	32	0.0641	32	0.0652	112	0.0180

N.B. SING. indicates that the least squares equations under the design are singular.

$s = 24$, the average variances are found to be $0{\cdot}0886\sigma^2$, $0{\cdot}0863\sigma^2$ and $0{\cdot}0883\sigma^2$ respective for n equal to 60 (Design III), 112 (Design I) and 112 (Design II). A comparison of average variances indicates that Design I is consistently more efficient than Design II since for the same values of n and s, the former always has a lower average variance than does the latter.

Further, when $n = 40$ and $s = 12$, Design I has a lower average variance than do Designs II and III. A comparison of Design I with $n = 24$ and $s = 8$ can be made with PDC Design (I) based on Group-Divisible (GD) association scheme for $n = 24$, $s = 9$ given in Arya and Narain (1977). The former with an average variance of $0{\cdot}286\sigma^2$ is less efficient than the latter with the corresponding figure of $0{\cdot}246\sigma^2$.

It may be mentioned that circulant designs of Kempthorne and Curnow (1961) are not defined for even values of n and s whereas the present ones are. These therefore fill an important gap for some of the non-available PDCs.

Appendix

Method for obtaining the latent roots and idempotent matrices of A

The special feature of the matrix $\boldsymbol{A}$ $(= \boldsymbol{NN}')$ or any other associated matrix such as $\boldsymbol{A}^{-1}$ or its idempotents is that they contain at the most $(m+1)$ distinct elements if the underlying PBIB design has m associate class association scheme. We can therefore work in terms of these elements. The (i, j)-th element of $\boldsymbol{A}$ is denoted by $a_{ij}(\boldsymbol{A}) = a_k(\boldsymbol{A})$ if the treatments i and j are k-th associates in the given association scheme, $k = 0, 1, \ldots, m$. The elements in $\boldsymbol{A}^{-1}$ or any other related matrix can be denoted in a similar way.

Bose and Mesner (1959) showed that the distinct latent roots of $\boldsymbol{A}$ are the same as the distinct latent roots of

$$\boldsymbol{\rho} = s\boldsymbol{I} + \lambda_1\boldsymbol{\rho}_1 + \ldots + \lambda_m\boldsymbol{\rho}_m, \qquad \ldots (1)$$

where λ's are the usual parameters of the PBIB design and s is the diagonal element in $\boldsymbol{A}$. Further $\boldsymbol{\rho}_k = \left(\begin{matrix} j \\ p_{it\neq k} \end{matrix}\right)$ is a matrix of order $(m+1)\times(m+1)$ as referred to in the above mentioned paper where p_{ik}^j $(i, j, k = 0, 1, \ldots, m)$ is the number of associates common to i-th associates of some treatment α and k-th associates of another treatment β where α and β are j-th associates. The ρ_i's matrices of (1) are cumutative and hence the latent roots of $\boldsymbol{\rho}$ can be obtained from those of $\boldsymbol{\rho}_i$'s on the right side when arranged in a SUITABLE ORDER and expressed by same relation. Thus if the latent roots of $\boldsymbol{\rho}_k$ $(k = 1, \ldots, m)$ are θ_{ik} $(i = 0, 1, \ldots, m)$ the roots of $\boldsymbol{\rho}$ will be

$$\theta_i = s + \sum_{k=1}^{m} \lambda_k \theta_{ik}, \quad i = 0, 1, \ldots, m. \qquad \ldots (2)$$

The ordering of the roots is done in such a manner that a particular root has the same multiplicity with respect to the matrix $\boldsymbol{A}$ for all the $\boldsymbol{\rho}_i$ matrices. The multiplicities are obtained as traces of the corresponding idempotent matrices of $\boldsymbol{A}$ which provides an alternative method to those of Bose and Mesner.

The idempotent matrices related with $\boldsymbol{A}$ are worked out by the help of the well known matrix

$$\boldsymbol{C} = \boldsymbol{s}^{\delta} - \boldsymbol{A}/2 \qquad \ldots \quad (3)$$

and its latent roots

$$\phi_i = s - \theta_i/2, \; i = 0, 1, \ldots, m \qquad \ldots \quad (4)$$

where $\mathbf{s}^{\delta}$ is a diagonal matrix with all the diagonal elements equal to s. Also, we consider the associated 2-plot block PBIB design in which only one of the λ's say λ_k is equal to one and the rest of them are zeros. Thus s will be equal to n_k where n_k is the number of k-th associates in the underlying association scheme ($k = 1, 2, \ldots, m$). The distinct elements of $C, C^2, \ldots, C^m$ can be obtained by successive multiplication. For C^2, we multiply the successive rows of C having initial element $a_k(C)$, $k = 0, 1, \ldots, m$ by the first row of C which has $a_0(C)$ as the initial element. The element $a_i(C)a_j(C)$ in this product will appear p_{ij}^k times if i and j are k-th associates ($i, j, k = 0, 1, \ldots, m$). Thus we have, for all k

$$a_k(C^2) = \sum_{i,\, j=0}^{m} p_{ij}^k \, a_i(C)a_j(C) \qquad \ldots \quad (5)$$

$$a_k(C^3) = \sum_{i,\, j=0}^{m} p_{ij}^k a_k(C)a_j(C^2) \qquad \ldots \quad (6)$$

and so on. Such elements are listed upto fifth powers in the Table to follow.

The matrix $\boldsymbol{C}$ of order $n \times n$ having the latent roots θ_i with multiplicities α_i ($i = 0, 1, \ldots, m$) such that $\sum_{i=0}^{m} \alpha_i = n$, will have the idempotent matrices

$$\boldsymbol{L}_i = \prod_{j \neq i} [(\boldsymbol{C} - \theta_j \boldsymbol{I})/(\phi_i - \phi_j)], \; i = 0, 1, \ldots, m.$$

This can be put as

$$D_i \, L_i = C^m - (\sum_{j \neq i} \phi_j) C^{m-1} + \ldots \pm \prod_{j \neq i} \theta_j \qquad \ldots \quad (7)$$

where

$$D_i = \prod_{j \neq i} (\phi_i - \phi_j). \qquad \ldots (8)$$

The last term on the right of (7) vanishes for $i = 1, 2, \ldots, m$ since ϕ_0 is always zero. Further $\boldsymbol{L}_0$ corresponding to the zero latent root of $\boldsymbol{C}$ is known to have all its elements equal to $1/n$. For the distinct elements $a_k(L_i)$, $k = 0, 1, \ldots, m$ of L_i's, we then have

$$D_k a_k(L_i) = a_{ki}(C^m) - \Big(\sum_{j \neq i} \phi_j\Big) a_k(C^{m-1}) + \Big(\sum_{l \neq j \neq i} \phi_l \phi_j\Big) a_k(C^{m-2}) - \qquad \ldots (9)$$

for $i = 1, 2, \ldots, m$. The idempotent matrices related with $\boldsymbol{A}$ which are the same as those related with $\boldsymbol{C}$, are thus known.

TABLE : ELEMENTS OF C, C^2, C^3, C^4 AND C^5 RELATED WITH A 2-PLOT BLOCK PBIB DESIGN WITH PARAMETERS n, $b = s/2$, $r = s$, $K = 2$, $\lambda_k = 1$, $\lambda_{k' \neq k} = 0$ WITH $s = n_k$,

Matrix	a_0	Element $a_i (i \neq k)$	a_k
$2C$	n_k	0	-1
$4C^2$	$n_k(n_k+1)$	p^i_{kk}	$p^k_{kk} - 2n_k$
$8C^3$	$n_k(n_k^2 + 3n_k - p^k_{kk})$	$3n_k 9^i_{kk} - c^i_{kk}$	$3n_k p^k_{kk} - q^k_{kk} - n_k(3n_k+1)$
$16C^4$	$n_k(n_k^3 + 6n_k^2 + n_k - 4n_k p^k_{kk} + q_{kk})$	$n_k(6n_k+1)p^i_{kk} - 4n_k q^i_{kk} + h^i_{kk}$	$n_k(6n_k+1)p^k_{kk} - 4n_k q^k_{kk} + h^k_{kk}$ $- n_k(4n_k^2 + 4n_k - p^k_{kk})$
$32C^5$	$n_k[n_k^4 + 10n_k^3 + 5n_k^2 - 2n_k$ $(5n_k+1)p^k_{kk} + 5n_k q^k_{kk} - h^k_{kk}]$	$n_k(10n_k^2 + 5n_k - p^k_{kk})p^i_{kk} - n_k$ $(10n_k+1)q^i_{kk} + 5n_k h^i_{kk} - l^i_{kk}$	$n_k(10n_k^2 + 5n_k - p^k_{kk})p^k_{kk} - n_k$ $(10n_k+1)q^k_{kk} + 5n_k h^k_{kk} - l^k_{kk}$ $- n_k(5n_k^3 + 10n_k^2$ $+ n_k - 5n_k p^k_{kk} + q^k_{kk})$

$$q^j_{ii} = \sum_{t=1}^{m} p^j_i \; p^t_{ii}, \; h^j_{ii} = \sum_{t=1}^{m} p^j_{it} \; q^t_{ii}, \; l^j_{ii} = \sum_{t=1}^{m} p^j_{it} \; h^t_i$$

References

ADHIKARY, B. (1966) : Some types of m-associate PBIB association scheme. *Cal. Statist. Assn. Bull*, **15**, 47-74.

——— (1967) : A new class of higher associate cyclical association schemes. *Cal. Statist. Assn. Bull*, **16**, 40-44.

ARYA, A. S. and NARAIN, P. (1977) : Partial diallel crosses based on some association schemes with three and four associate classes. *Sankhyā*. **39** (Series B), 394-399.

CURNOW, R. N. (1963) : Sampling the diallel cross. *Biometrics*, **19**, 287-306.

FYFE, J. L. and GILBERT, N. E. G. (1963) : Partial diallel crosses. *Biometrics*, **19**, 278-286.

HINKELMANN, K. (1964) : Extended group divisible partially balanced incomplete block designs *Ann. Math. Statist.*, **35**, 681-695.

JOHN, P. W. M. (1966) : An extension of the triangular association scheme to three associate classes. *J. Roy. statist. Soc.*, B, **28**, 361-365.

KEMPTHORNE, O. and CURNOW, R. N. (1961) : The partial diallel crosses. *Biometrics*, **17**, 229-250.

KISHEN, K. and SHUKLA, A. N. (1974) : Truncated triangular designs and associated asymmetrical factorial designs (Abstract) *J. Indian Soc. Agri. Stat.*, **26**, 169.

KUSUMOTO, K. (1965) : Hyper-cubic designs. *Wakayama Medical Report*, **9**, 123-132.

NARAIN, P., RAO, C. SUBBA and NIGAM, A. K. (1974) : Partial diallel crosses based on extended triangular association scheme. *Indian J. Genet.*, **34**, 309-317.

OGASAWARA, M. (1965) : A necessary condition for the existence of regular and symmetrical PBIB design of T_m type. *Inst. Stat. Mimco. Series* **418** *Chapel Hill, N.C.*

RAGHAVARAO, D. (1960) : A generalisation of group divisible designs. *Ann. Math. Statist.*, **31**, 756-771.

RAGHAVARAO, D. and CHANDRASEKHARARAO, K. (1964) : Cubic designs. *Ann. Math. Statist.*, **35**, 389-397.

ROY, P. M. (1953-54) : Hierarchical group divisible incomplete block designs with m-associate classes. *Science and Culture*, **19**, 210-211.

THARTHARE, S. K. (1963) : Rightangular designs. *Ann. Math. Statist.*, **34**, 1057-1067.

——— (1965) : Generalised rightangular designs. *Ann. Math. Statist.*, **36**, 1535-1553.

Paper received : May, 1979.

Revised : November, 1979.