

**Methodology in medical genetics : an introduction to statistical methods /
Alan E.H. Emery.**

Contributors

Emery, Alan E. H.

Publication/Creation

Edinburgh : Churchill Livingstone, 1986.

Persistent URL

<https://wellcomecollection.org/works/btmz5m4d>

License and attribution

You have permission to make copies of this work under a Creative Commons, Attribution, Non-commercial license.

Non-commercial use includes private study, academic research, teaching, and other activities that are not primarily intended for, or directed towards, commercial advantage or private monetary compensation. See the Legal Code for further information.

Image source should be attributed as specified in the full catalogue record. If no source is given the image should be attributed to Wellcome Collection.



Wellcome Collection
183 Euston Road
London NW1 2BE UK
T +44 (0)20 7611 8722
E library@wellcomecollection.org
<https://wellcomecollection.org>


SECOND EDITION

METHODOLOGY IN MEDICAL GENETICS

AN INTRODUCTION TO
STATISTICAL METHODS

Alan E. H. Emery

M
9432

Churchill Livingstone 

ISBN 0-443-03509-1



9 780443 035098



CONTENTS

CONTENTS

THE EDITORS

DR. J. H. J. VAN OOSTRA, DR. J. H. J. VAN OOSTRA, DR. J. H. J. VAN OOSTRA

Methodology in Medical Genetics

AN INTRODUCTION TO STATISTICAL METHODS

Alan E. H. Emery MD PhD DSc FRCP MFCM FRS(E)

Emeritus Professor of Human Genetics and University Fellow,
The Medical School, University of Edinburgh

SECOND EDITION



CHURCHILL LIVINGSTONE
EDINBURGH LONDON MELBOURNE AND NEW YORK 1986

CHURCHILL LIVINGSTONE
Medical Division of Longman Group Limited

Distributed in the United States of America by
Churchill Livingstone Inc., 1560 Broadway, New York,
N.Y. 10036, and by associated companies, branches
and representatives throughout the world.

© Longman Group Limited 1986

All rights reserved. No part of this publication may be
reproduced, stored in a retrieval system, or transmitted
in any form or by any means, electronic, mechanical,
photocopying, recording or otherwise, without the prior
permission of the publishers (Churchill Livingstone,
Robert Stevenson House, 1-3 Baxter's Place, Leith
Walk, Edinburgh EH1 3AF).

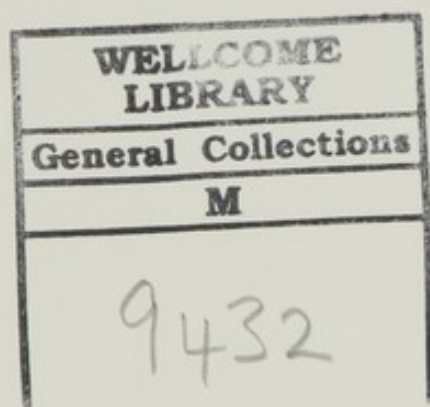
First edition 1976
Italian edition 1986
Second edition 1986

ISBN 0-443-03509-1

British Library Cataloguing in Publication Data

Emery, Alan E.

Methodology in medical genetics:
an introduction to
statistical methods. — 2nd ed.
1 Medical genetics — Statistical methods
I. Title
616.042'072 RB155



Printed in Great Britain by
Butler & Tanner Ltd, Frome and London

Preface to the Second Edition

As in the first edition of this little book the emphasis has remained on its being essentially a practical guide to simple statistical methods which the investigator can easily apply without recourse to anything more than a pocket calculator. Despite the exciting, and often revolutionary, developments in various laboratory disciplines in recent times, many of these statistical methods still remain the cornerstone of much research in medical genetics. In the last few years the use of computer programs has made many computations very much easier—for example, in segregational analysis, linkage studies and risk determination using data from linked DNA probes. However, it would be wrong to apply such programs uncritically without appreciating at least the basic underlying principles involved; in this regard also it is hoped the book may have some value.

The entire text has been revised with an additional chapter on the resolution of genetic heterogeneity, a subject of increasing importance to medical geneticists. Finally, statistical methods involved in the use of DNA probes are also discussed, a field likely to develop considerably in the near future.

Edinburgh/Ibiza
1986

A.E.H.E.

Preface to the First Edition

This is not intended to be a textbook but rather a practical guide to simple statistical methods of use to those with a particular interest in medical genetics. The emphasis throughout is on the solution of practical, rather than theoretical, problems and particularly on problems of medical importance.

It is assumed that the reader has some knowledge of human genetics and an acquaintance with very simple statistics, but a level of mathematical sophistication no greater than simple algebra is required.

An effort has been made to make the book more or less self-contained, with sufficient information, in the form of worked examples and reference tables, to enable the reader to apply the methods to his or her own data. It is hoped that the book will at least encourage, and perhaps help, those who would like to attempt to analyse their own data themselves armed with no more than log tables or a hand calculator.

Edinburgh/Ibiza
1976

A.E.H.E.

Acknowledgements

I should like to thank all those who made many helpful suggestions for the preparation of this second edition. I am especially grateful for the valuable advice of Professor John Edwards (Oxford), and also to Professor Antonio Danieli (Padua) and Dr Jeffrey Sofaer (Edinburgh), as well as Dr J. Clayton (Edinburgh), Dr R.J.M. Gardner (Dunedin), Dr C. Hoff (Mobile) and Dr J. Yates (Glasgow). I must also thank the following authors and publishers for permission to use various tables and figures: Table 4.3 (W.W. Norton Inc., New York), Table 4.5 (Professor C.C. Li and McGraw-Hill Inc., New York), Table 4.6 (Professor C.C. Li and Dr N. Mantel and the editor and publishers of the *American Journal of Human Genetics*), Figure 5.2 (Dr Charles Smith and the editor and publishers of the *Annals of Human Genetics (London)*), Figure 7.4 (Dr R.E. Gaines and Dr R.C. Elston and the editor and publishers of the *American Journal of Human Genetics*), Table 10.4 (Dr J. Sofaer), Table 11.2 (Professor C.A.B Smith and the editor and publishers of the *Annals of Human Genetics (London)*), Table 11.4 (Dr Susan Holloway), Table 12.3 (Dr D. Hewitt and the editor and publishers of the *British Journal of Preventive and Social Medicine*), Table 12.5 (Dr L.S. Freedman and the editor and publishers of the *Journal of Epidemiology and Community Health*), Appendices 1, 2 and 3 (Professor N.T.J. Bailey and the English Universities Press), Appendix 4 (Dr R.R. Sokal and Dr F.J. Rohlf and Freeman Inc., San Francisco), Appendix 5 (Professor D.S. Falconer and the editor and publishers of the *Annals of Human Genetics (London)*), and Appendix 6 (Professor C.A.B. Smith). Finally I should thank the late Mr John Pizer for preparing the illustrations and particularly my secretary, Mrs Isobel Black, for the cheerful and efficient way in which she typed the script.

Contents

1. Introduction	1
2. Hardy-Weinberg equilibrium and the estimation of gene frequencies	3
Hardy-Weinberg equilibrium	3
Estimation of autosomal gene frequencies	4
Determination of the expected frequencies of various matings and the phenotypes of their offspring	8
Estimation of multiple allele frequencies	10
3. Estimation of factors affecting the genetic structure of populations	12
Genetic drift	12
Assortative mating	15
Inbreeding	17
Gene flow	23
Selection	25
Mutation	33
Genetic distance	35
4. Segregation analysis	37
Complete (= truncate) ascertainment	40
Single incomplete ascertainment	46
Multiple incomplete ascertainment	51
X-linked inheritance	53
5. Multifactorial inheritance	55
Tests for multifactorial inheritance	55
Estimation of heritability from family studies	57
Calculation of heritability	59
Estimation of heritability from twin studies	64
6. Genetic linkage	67
Autosomal linkage	67
Prior probabilities of linkage	73
Probability of linkage	73
Probability limits	73
Recombination fraction and map distance	74
X-linkage	75

7. Twin studies, their use and limitations	79
Diagnosis of zygosity	80
The use of twins in genetic analysis	86
Problems and limitations of twin studies	91
8. Estimation of recurrence risks for genetic counselling	93
Unifactorial disorders	93
Linkage and DNA markers	103
Dominant disorders with reduced penetrance	109
Multifactorial disorders	111
9. Disease associations	114
Penrose sib method	114
Woolf's method	116
Smith's method	121
Problems of disease association studies	122
Value of disease association studies	123
10. Resolution of genetic heterogeneity	126
Pedigree studies	127
Analysis of variance	128
Evidence of bimodality	129
Correlations between relatives	133
Cousins and parental consanguinity	135
Disease associations and linkage	139
11. Parental age and birth order	140
Method of Haldane and Smith	141
Choice of controls	145
Method of partial correlations	148
12. Recognition and estimation of changes in disease frequency	154
Incidence and prevalence	154
Comparison of proportions	155
Cumulative sum techniques ('cusums')	156
Cyclical changes	159
Appendices	164
1. Student's <i>t</i> distribution	165
2. χ^2 distribution	166
3. Correlation coefficient	167
4. Transformation of <i>r</i> to <i>z</i>	168
5. Normal distribution for estimation of h^2	170
6. Lod scores	175
References	185
Index	193

Introduction

In the last decade or so, developments in human biochemical genetics and cytogenetics have tended to eclipse quantitative methods in medical genetics. These methods, however, will always provide the basis for much research in the subject. Admittedly some have little practical value, as for example studies of genetic drift and effective population size, assortative mating and inbreeding, gene flow and racial admixture, and natural selection, but clearly the study and measurement of such phenomena are essential for any understanding and appreciation of man's evolution. However developments in recombinant DNA technology (genetic engineering) and the generation of DNA markers in recent years, have lead to the increasing application of linkage studies in genetic counselling and antenatal diagnosis, areas of considerable practical importance.

Several statistical methods are particularly valuable in helping to elucidate the role of environmental factors in congenital malformations of unknown aetiology. Particularly useful in this regard are the techniques for recognizing and measuring changes in disease frequency and cyclical trends, and for estimating parental age and birth order effects.

The study of disease associations has taken a new lease of life with the discovery of strong associations with certain HLA types which may well throw light on the aetiology of those disorders with which they are associated, and though interest in twin studies has somewhat declined in recent years much valuable information concerning the nature versus nurture controversy can still be gained from such studies, particularly in the realm of psychiatric disorders.

Yet other techniques, either directly or indirectly, have yielded information of value in risk prediction for genetic counselling. The estimation of heritability is most valuable as a measure of genetic determination but such information can also be used to predict risks to relatives, and segregation analysis can help establish the mode of inheritance which is obviously important for genetic counselling. Methods for estimating recurrence risks, often employing statistical tools such as Bayes' theorem, have become increasingly important in recent years as the need for genetic counselling has become more widely accepted.

Some of these methods, however, are complicated and have occupied the attention of some of the best intellects in human genetics. For this reason the non-mathematically minded are sometimes discouraged. This book is specially written for those with a level of mathematical sophistication no greater than simple algebra. This of course means that rarely will the derivation and proof of an equation or relationship be given but in all such cases reference is made to where this information can be found. The reader, however, is assumed to have some knowledge of basic genetics and simple statistical methods and so be acquainted with such terms as standard error (SE), statistical significance, correlation coefficient and chi square (χ^2).

The book is intended to be a simple straightforward *practical* guide to methods for analysing human genetic data. Each method is illustrated with worked examples from real data, either published or unpublished, and tables and graphs are included to help the reader with the calculations. The methods described are essentially those which can be applied by the individual investigator armed with no more than log tables or a pocket calculator. Some refined methods, usually requiring a computer for analysis, have therefore been considered beyond the scope of this book; for example, the calculation of the coefficient of inbreeding from marriage distances and computer methods for discriminating between different modes of inheritance. One further point: particular data have been chosen because they illustrate a method of calculation and not because they necessarily (though they often do) represent the best available data on the subject. Since this is more a work book than a text book no serious attempt has been made to assess critically the results of such studies. However the problems and limitations of the various methods are emphasized and discussed, and references are given to original reports so that the interested reader may find more detailed treatment of a particular statistical method. The principal danger is the uncritical application of the methods described. If in doubt the reader should therefore always consult the original reference or an experienced colleague, which will be necessary, in any event, if the data warrant more complex analysis than is covered by this introduction, the aim of which was to deal only with simple basic methods.

It is hoped that the book is more or less self-contained with sufficient information to enable the reader to apply the methods to his or her own data, or at least help the reader to understand and perhaps appreciate more fully the studies of others.

Hardy-Weinberg equilibrium and the estimation of gene frequencies

Hardy-Weinberg equilibrium

Proposed by an English mathematician, G. H. Hardy, and a German physician, W. Weinberg, in 1908, the so-called 'Hardy-Weinberg principle' can be expressed as follows. In a large, randomly mating (= panmixis) population, in which there is no migration, or selection against a particular genotype and the mutation rate remains constant, the proportions of the various genotypes will remain unchanged from one generation to another. An understanding of this principle is essential for much that will follow.

Consider two alleles '*A*' and '*a*' such that the proportion of '*A*' genes is '*p*' and the proportion of '*a*' genes is '*q*', then $p + q = 1$. Throughout, '*q*' will be used to denote the frequency of the recessive allele. Now with random mating the frequencies of the various genotypes will be:

		Male gametes					
		<i>A</i> (<i>p</i>)	<i>a</i> (<i>q</i>)				
Female gametes	{	<i>A</i> (<i>p</i>)	<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding: 5px;"><i>AA</i> (<i>p</i>²)</td> <td style="padding: 5px;"><i>Aa</i> (<i>pq</i>)</td> </tr> <tr> <td style="padding: 5px;"><i>Aa</i> (<i>pq</i>)</td> <td style="padding: 5px;"><i>aa</i> (<i>q</i>²)</td> </tr> </table>	<i>AA</i> (<i>p</i> ²)	<i>Aa</i> (<i>pq</i>)	<i>Aa</i> (<i>pq</i>)	<i>aa</i> (<i>q</i> ²)
	<i>AA</i> (<i>p</i> ²)	<i>Aa</i> (<i>pq</i>)					
<i>Aa</i> (<i>pq</i>)	<i>aa</i> (<i>q</i> ²)						
	}	<i>a</i> (<i>q</i>)					

Thus the frequencies of the various offspring from such matings are $p^2(AA)$, $2pq(Aa)$ and $q^2(aa)$, that is the terms of the expansion $(p + q)^2$.

If these progeny now mate with each other the frequencies of the various matings can be represented as:

		Genotype frequency of male parent		
		AA (p^2)	Aa ($2pq$)	aa (q^2)
Genotype frequency of female parent	AA (p^2)	p^4	$2p^3q$	p^2q^2
	Aa ($2pq$)	$2p^3q$	$4p^2q^2$	$2pq^3$
	aa (q^2)	p^2q^2	$2pq^3$	q^4

Thus, for example, the frequency of matings between persons with the genotypes 'aa' and 'Aa' is $2pq^3 + 2pq^3$ or $4pq^3$. The frequencies of the various offspring from these matings can be represented as:

Mating type	Frequency	Frequency of offspring		
		AA	Aa	aa
$AA \times AA$	p^4	p^4	—	—
$AA \times Aa$	$4p^3q$	$2p^3q$	$2p^3q$	—
$Aa \times Aa$	$4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
$AA \times aa$	$2p^2q^2$	—	$2p^2q^2$	—
$Aa \times aa$	$4pq^3$	—	$2pq^3$	$2pq^3$
$aa \times aa$	q^4	—	—	q^4

Total

$$\begin{aligned}
 &= p^2(p^2 + 2pq + q^2) + 2pq(p^2 + 2pq + q^2) + q^2(p^2 + 2pq + q^2) \\
 &= p^2(p + q)^2 + 2pq(p + q)^2 + q^2(p + q)^2 \\
 &= p^2 + 2pq + q^2 \\
 &= (p + q)^2
 \end{aligned}$$

The proportions of the various genotypes remain the same in the second generation as in the first generation.

Estimation of autosomal gene frequencies

The method of estimation depends upon whether or not the heterozygote is recognizable.

Heterozygote is not recognizable

In this case there is complete dominance and therefore the heterozygote is not

recognizable. Assuming that the genotypes are in equilibrium, then the gene frequencies can be estimated if the frequency of the rare homozygote is known. Thus in alkaptonuria (a recessive disorder) which affects about one child in every million:

$$q^2 = \frac{1}{1\,000\,000}$$

therefore

$$q = \frac{1}{1000}$$

but

$$p + q = 1$$

therefore

$$p \approx 1$$

and the frequency of heterozygous carriers is $2pq$ or $1/500$.

The standard error of the estimation of 'q' (when the estimate of 'q' is based upon the frequency of homozygotes q^2) is $[(1 - q^2)/4N]^{\frac{1}{2}}$ where N is the number of individuals in the sample. Thus Pearn (1973) ascertained 9 cases of Werdnig-Hoffmann disease (a recessive disorder) in a total of 231 370 births in the North-East of England.

Therefore

$$q^2 = \frac{9}{231\,370}$$

$$= 0.000\,039$$

and

$$q = \sqrt{0.000\,039}$$

$$= 0.006\,24$$

and

$$SE = \sqrt{\frac{1 - 0.000\,039}{(4)(231\,370)}}$$

$$= 0.001\,04$$

The 95% confidence limits will therefore be

$$\text{mean} \pm 1.96 \times SE$$

$$= 0.00624 \pm 1.96 (0.001\,04)$$

$$= 0.004\,20 \text{ to } 0.008\,28$$

Heterozygote is recognizable

If a characteristic is suspected of being determined by two codominant alleles, the heterozygote therefore being recognizable, the frequencies of the two genes can be estimated. Since the frequency of heterozygotes (H)

$$= 2pq$$

if the disorder is very rare then

$$q \approx \frac{H}{2}$$

But this is only true when p is almost unity, otherwise

$$\begin{aligned} H &= 2pq \\ &= 2(1 - q)q \\ &= 2q - 2q^2 \\ 1 - (1 - 2q)^2 &= 2H \\ 1 - 2q &= \sqrt{1 - 2H} \\ q &= \frac{1 - \sqrt{1 - 2H}}{2} \end{aligned}$$

and squaring this would give the frequency of affected homozygotes. Thus in parts of Africa where the incidence of carriers of sickle cell anaemia (sickle cell trait) has been found to be as high as 1 in 3,

$$\begin{aligned} q &= \frac{1 - \sqrt{1 - 0.667}}{2} \\ &= 0.211 \end{aligned}$$

and therefore

$$q^2 = 0.044 \text{ or } 1 \text{ in } 23$$

Another approach is illustrated by a study (Kellerman et al, 1973) in which the induction of aryl hydrocarbon hydroxylase in human lymphocytes showed a trimodal distribution in the population and it was suggested that the three phenotypes represented the action of two alleles (A and B). Out of a total of 161 individuals investigated the phenotypic frequencies were:

$$\text{low inducibility} = 86 \text{ (AA)}$$

$$\text{intermediate inducibility} = 59 \text{ (AB)}$$

$$\text{high inducibility} = 16 \text{ (BB)}$$

$$\begin{aligned} \text{Therefore } A \text{ gene frequency} &= \frac{86}{161} + \frac{1}{2} \left(\frac{59}{161} \right) \\ &= 0.717 \end{aligned}$$

$$\begin{aligned} \text{and } B \text{ gene frequency} &= 1 - 0.717 \\ &= 0.283 \end{aligned}$$

Therefore the *expected* phenotype frequencies are:

$$AA = 161 (0.717) (0.717) = 82.8$$

$$AB = 161 (2) (0.717) (0.283) = 65.3$$

$$BB = 161 (0.283) (0.283) = 12.9$$

To determine if the observed (O) and expected (E) results differ significantly we calculate the value of chi square (χ^2) which is equal to the square of the difference between O and E divided by E summed (represented by Σ) for all groups.

Thus:

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(3.2)^2}{82.8} + \frac{(6.3)^2}{65.3} + \frac{(3.1)^2}{12.9} \\ &= 1.48\end{aligned}$$

We next determine the *number of degrees of freedom* (DF). In this sort of test—referred to as a ‘goodness of fit’ test—the number of degrees of freedom

$$= (\text{no. of classes}) - (\text{no. of estimated parameters}) - 1$$

In the above example there are three classes and there was one estimated parameter, namely the gene frequency, upon which the expected values were calculated. Therefore there is *one* degree of freedom. (The reader is referred to one of the standard text books of statistics for a discussion of the number of degrees of freedom in various statistical calculations.) With one degree of freedom, to be significant ($P < 0.05$) the value of χ^2 would have to be greater than 3.84 (Appendix 2, p. 166). In fact the value of χ^2 is only 1.48 and therefore there is no significant difference between the observed and expected numbers of low, intermediate and high inducers if it is assumed that these phenotypes result from the operation of two codominant alleles, though subsequent research has now shown that the genetic control of aryl hydrocarbon hydroxylase inducibility is in fact more complicated than this.

In the case of autosomal dominant disorders with late onset, such as Huntington’s chorea, the frequency of heterozygotes in the general population (H) has to be determined indirectly because some will not yet be affected. A useful method is that proposed by Reed et al (1958):

$$H = \frac{A}{\sum N_x P_x}$$

where

A = number of observed patients in a given area

N_x = total individuals aged x

P_x = proportion of heterozygotes diagnosed by age x

summing over all ages.

The weakness of such estimates however, is that they depend on the completeness of patient ascertainment.

Determination of the expected frequencies of various matings and the phenotypes of their offspring

Autosomal disorders

If it is considered that a certain characteristic could be due to the operation of two alleles, it is possible to determine the expected frequencies of the various types of matings, and the frequencies of the various types of offspring from these matings and to compare these findings with those observed.

For example, Evans et al (1960) showed that it is possible to divide individuals into two classes according to their ability to metabolize the drug isoniazid. These are referred to as 'rapid' and 'slow' inactivators. In order to determine if the slow inactivator phenotype represents the homozygous recessive genotype, Professor Price Evans and colleagues compared the observed and expected mating frequencies and their offspring. Out of a total of 291 individuals investigated the phenotype frequencies were:

$$\text{slow inactivators} = 152$$

$$\text{rapid inactivators} = 139$$

If *slow* inactivation represents the homozygous expression of an autosomal recessive gene (i.e. $I_r I_r$).

$$\text{Then } I_r I_r (q^2) = \frac{152}{291}$$

$$= 0.5223$$

$$\text{therefore } I_r (q) = \sqrt{0.5223}$$

$$= 0.7227$$

$$\text{and } I_R (p) = 1 - 0.7227$$

$$= 0.2773$$

Assuming random mating the number of expected mating types can then be calculated and compared with the observed numbers (Table 2.1).

Table 2.1 Numbers of observed matings compared with those expected if slow inactivation of isoniazid represents the homozygous expression of an autosomal recessive gene (Evans et al, 1960)

Phenotypic matings	Genotypic matings	Expected frequency of matings	Expected occurrence in 53 matings	Observed occurrence
$S \times S$	$I_r I_r \times I_r I_r$	q^4 0.2728	14.46	16
$R \times S$	$I_R I_R \times I_r I_r$	$2p^2 q^2$ 0.0803	26.45	24
	$I_R I_r \times I_r I_r$	$4pq^3$ 0.4187		
$R \times R$	$I_R I_R \times I_R I_R$	p^4 0.0059	12.09	13
	$I_R I_R \times I_R I_r$	$4p^3 q$ 0.0616		
	$I_R I_r \times I_R I_r$	$4p^2 q^2$ 0.1606		

The observed and expected numbers of the different mating types can then be compared in the usual manner (Table 2.2).

Table 2.2 Comparison of the observed and expected numbers of matings in Table 2.1.

Mating	Observed	Expected	$(O - E)^2$	$\frac{(O - E)^2}{E}$
$S \times S$	16	14.46	2.372	0.164
$R \times S$	24	26.45	6.003	0.227
$R \times R$	13	12.09	0.828	0.0685
				$\chi^2 = 0.4595$
				(DF = 1)

The value of χ^2 is 0.4595 which is not significant (Appendix 2, p. 166). Therefore the observed numbers of different mating types do not differ significantly from the expected numbers when it is assumed that slow inactivation represents the homozygous recessive genotype.

A further test of this hypothesis is to compare the expected with the observed numbers of children of each phenotype which result from various matings. Thus in matings between rapid and slow inactivators, assuming slow inactivation represents the homozygous recessive genotype, the expected proportion of slow inactivators ($I_r I_r$) offspring is $2pq^3$ (p. 4), and the proportion among offspring resulting from this particular mating is:

$$\begin{aligned} & \frac{2pq^3}{2pq^3 + 2p^2q^2 + 2pq^3} \\ &= \frac{q}{p + 2q} \\ &= \frac{q}{1 + q} \\ &= \frac{0.7227}{1.7227} \\ &= 0.4195 \end{aligned}$$

Therefore the expected *number* of slow inactivator offspring among 70 offspring of matings between rapid and slow inactivators is (70) (0.4195) or 29.36. Similarly the expected number of children of slow and rapid inactivator phenotype among the offspring of other matings can be determined (Table 2.3).

Table 2.3 Expected numbers of children of each isoniazid inactivator phenotype compared with those observed (Evans et al, 1960)

Phenotypic matings	No. of matings	No. of children	No. of children of each phenotype				χ^2	DF
			Rapid <i>E</i>	<i>O</i>	<i>E</i>	Slow <i>O</i>		
<i>S</i> × <i>S</i>	16	51	0	0	51	51	—	—
<i>R</i> × <i>S</i>	24	70	40.62	42	29.36	28	0.110	1
<i>R</i> × <i>R</i>	13	38	31.30	31	6.68	7	0.018	1
	53	159		73		86	0.128	2

Since there is no significant difference between the observed and expected numbers, the data fit the hypothesis that slow inactivator phenotype represents the genetically homozygous recessive individual.

X-linked disorders

In an X-linked disorder the frequency of the mutant allele ('*q*') is equal to the incidence of the disorder among males. The frequencies of the various types of matings and the proportions of the various types of offspring from these matings can be represented as:

Mating type			Proportion among offspring of a given sex				
			Males		Females		
Male	Female	Frequency	<i>a</i>	<i>A</i>	<i>aa</i>	<i>Aa</i>	<i>AA</i>
<i>a</i>	<i>AA</i>	p^2q	—	1	—	1	—
<i>a</i>	<i>Aa</i>	$2pq^2$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	—
<i>a</i>	<i>aa</i>	q^3	1	—	1	—	—
<i>A</i>	<i>AA</i>	p^3	—	1	—	—	1
<i>A</i>	<i>Aa</i>	$2p^2q$	$\frac{1}{2}$	$\frac{1}{2}$	—	$\frac{1}{2}$	$\frac{1}{2}$
<i>A</i>	<i>aa</i>	pq^2	1	—	—	1	—

As in the above example (p. 8), knowing '*q*' it is possible to calculate:

1. The expected frequencies of various matings and compare these with the observed frequencies.
2. The expected frequencies of different types of offspring from various matings and compare these with the observed frequencies.

Estimation of multiple allele frequencies

When there are three alleles but only certain phenotypes can be recognized, gene frequencies have to be determined indirectly. For example, in the case of the ABO blood groups, if the frequency of individuals with blood group *O* (*OO*) is represented as (\bar{O}), with blood group *A* (*AA* and *AO*) as (\bar{A}) and with

blood group B (BB and BO) as (\bar{B}) then by simple algebra it can be shown that the gene frequencies are respectively:

$$I^A = \sqrt{(\bar{O}) + (\bar{A})} - \sqrt{(\bar{O})}$$

or

$$1 - \sqrt{(\bar{O}) + (\bar{B})}$$

$$I^B = \sqrt{(\bar{O}) + (\bar{B})} - \sqrt{(\bar{O})}$$

or

$$1 - \sqrt{(\bar{O}) + (\bar{A})}$$

$$I^O = \sqrt{(\bar{O})}$$

When calculated in this way the sum of all the gene frequencies may not be equal to 1.00. There will be a deviation from unity referred to as 'D', where

$$D = \sqrt{(\bar{O}) + (\bar{A})} + \sqrt{(\bar{O}) + (\bar{B})} - \sqrt{\bar{O}} - 1$$

An improved estimate of the gene frequencies can be obtained in the following way:

$$I^O = (1 + D/2)(\sqrt{(\bar{O})} + D/2)$$

$$I^A = (1 + D/2)(1 - \sqrt{(\bar{O}) + (\bar{B})})$$

$$I^B = (1 + D/2)(1 - \sqrt{(\bar{O}) + (\bar{A})})$$

This and other methods of estimating blood group gene frequencies are clearly described in Race & Sanger (1975) and Levitan & Montagu (1977). ABO blood group gene frequencies in the United Kingdom are given in Table 2.4.

Table 2.4 Blood group gene frequencies in the United Kingdom (data selected from Mourant et al, 1958)

	<i>A</i>	<i>B</i>	<i>O</i>
England	0.252	0.050	0.698
Scotland	0.210	0.071	0.719
Wales	0.244	0.064	0.692
Northern Ireland	0.210	0.069	0.721
Overall	0.257	0.060	0.683

Estimation of factors affecting the genetic structure of populations

We have seen that according to the Hardy-Weinberg principle it is assumed that the various genotypes in a population are in equilibrium, and their proportions therefore remain constant from one generation to another. However, this is only true in large populations with no *genetic drift*, and where there is random mating (panmixis) with no significant *assortative mating* or *inbreeding*, no *gene flow* from migration or racial admixture, no *selection* against a particular genotype and a constant rate of *mutation*. We shall now discuss how each of these factors can be estimated in a given population.

Genetic drift

In large populations random variations in the number of children produced by individuals with different genotypes has no significant effect on gene frequencies but this is not so in small populations ('*demes*' or '*isolates*') where such variations may have a considerable effect on gene frequencies (Sewall Wright effect). If only a few people carry a particular gene, if such individuals do not have children or they have children but by chance do not transmit this gene to their offspring, then, barring a fresh mutation, the gene in question will completely disappear from the population (Fig. 3.1) and is said to have been '*extinguished*' (gene frequency zero) and its allele to have become '*fixed*' (gene frequency 1.0). The amount of random genetic drift depends on the size of the population being greatest in small populations where oscillations in gene frequencies from one generation to another may be considerable.

Genetic drift is therefore a function of population size although not of *total* population size but rather the numbers of adults of breeding age and their ability to have offspring to contribute to the gene pool of the next generation. This is referred to as the *effective size* of the population or ' N_e '. Significant genetic drift is likely to occur in a given population whenever:

$$\mu, s \text{ or } m < 1/2N_e$$

where μ = mutation rate; s = coefficient of selection; m = migration rate.

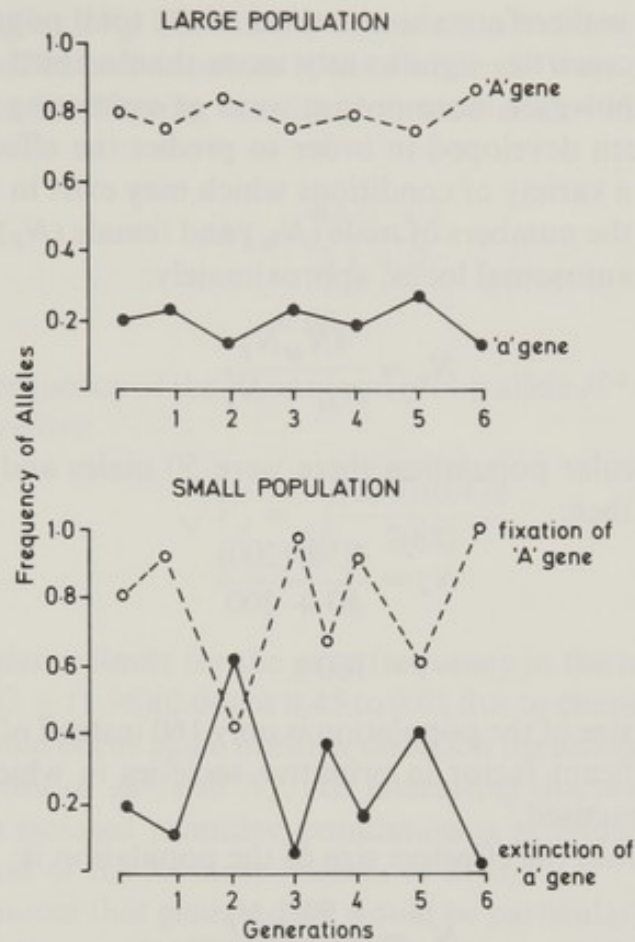


Fig. 3.1 The effects of genetic drift on gene frequencies in large and small populations (diagrammatic)

If the mutation rate is of the order of 10^{-6} , and if selection and migration are extremely low, then random genetic drift can still be important in populations of an effective size of up to 250 000 (Wright, 1948).

In a small population, in the absence of mutation, selection or migration, the percentage of loci which will become fixed or eliminated in each generation is $100/2N_e$. Thus in a religious isolate in Pennsylvania, the so-called 'Old Order "Dunker" (Old German Baptist) brethren', the community consisted of 298 individuals and the effective size of the population was estimated to be about 90 individuals (Glass et al, 1952). Therefore the percentage of loci which might be expected to become fixed or eliminated per generation in this community is $100/180$ or 0.55%, not an inconsiderable number of loci considering the possible size of the human genome.

Estimation of effective population size

In Western countries roughly a third of the population is in the reproductive

age-group and N_e is therefore about a third of the total population size, but in less developed countries significantly more than a third are in this age-group. There are however more precise ways of estimating N_e and various equations have been developed in order to predict the effective size of the population under a variety of conditions which may exist in nature (Kimura & Crow, 1963). If the numbers of male (N_M) and female (N_F) parents are not equal then for an autosomal locus, approximately:

$$N_e = \frac{4N_M N_F}{N_M + N_F}$$

Thus if in a particular population there were 50 males and 200 females of reproductive age, then:

$$\begin{aligned} N_e &= \frac{4(50)(200)}{50 + 200} \\ &= 160 \end{aligned}$$

Thus the effective size of the population is only 160 instead of 250. This could have been a significant factor in primitive societies in which polygyny (or polyandry) was practised.

For X-linked genes the effective size of the population is:

$$N_e = \frac{9N_M N_F}{4N_M + 2N_F}$$

Taking into account variation in number of offspring, and providing the population is fairly stable in size, then:

$$N_e = \frac{4N - 2}{V_o + 2}$$

where N = number of individuals of reproductive age
(say 15 to 45 years)

V_o = variance in number of offspring
(ideally those surviving to reproduction)

If necessary this can be computed independently for male and female parents, and separate estimates for the effective size of the population obtained for the two sexes.

Effective population size and gene frequencies

The variance in gene frequency is:

$$V_q = \frac{pq}{2N_e}$$

and therefore the expected (standard) deviation in one generation due to

chance sampling is $\sqrt{V_q}$. Thus in the Cashinahua Indians, a genetic isolate in Peru, there were 206 individuals in 1966 of whom 87 were of reproductive age and the variance in offspring was 3.1 (Johnston et al, 1969).

Therefore:

$$N_e = \frac{4(87) - 2}{3.1 + 2}$$

$$= 68$$

Now the gene frequency of the Kidd blood group allele Jk^a was 0.53 (Johnston et al, 1968) therefore:

$$\sqrt{V_q} = \sqrt{\frac{(0.53)(0.47)}{2(68)}}$$

$$= 0.04$$

The 95% confidence limits for the gene frequency in the next generation will therefore be $0.53 \pm (1.96)(0.04)$ or 0.45 to 0.61 due to chance alone. After this, genetic drift would occur again in either direction the amount being a function of the new values of Jk^a and N_e . An interesting discussion of the genetic structure of an isolated primitive population is provided by Salzano et al (1967) in the case of the Xavante Indians of Brazil.

It should be noted that genetic drift would be particularly important in the spread of neutral genes. This has been referred to as non-Darwinian evolution in contrast to classical Darwinian evolution in which natural selection plays the major role (Thoday, 1975).

To determine if selection or drift have played the greater role in determining specific gene frequencies, investigators have sometimes employed the so-called Wright's F_{ST} which is:

$$= \frac{V_q}{pq} \left(\text{which is } = \frac{1}{2N_e} \right)$$

high values implying selection, low values implying drift. Thus in Western countries low values have been obtained for the ABO blood groups but relatively high values for Duffy (F_y) blood group (Tills, 1977).

Assortative mating

The Hardy-Weinberg equilibrium only holds true if there is random mating (panmixis). *Assortative mating* and *inbreeding* disturb the equilibrium and result in an increase in the proportion of homozygotes and a decrease in the proportion of heterozygotes.

Assortative mating is usually concerned with resemblance between phenotypic traits such as height, intelligence, skin colouring and general physiognomy which have a multifactorial basis. In order to estimate the contribution of assortative mating to the total variance of a particular trait it is necessary to compute the following (Cavalli-Sforza & Bodmer, 1971):

$$C_1 C_2 = \frac{2r_{P/O}}{1 + r_{SP}}$$

$$\hat{A} = r_{SP}(C_1 C_2)$$

$$C_1 = 4r_{S/S} - C_1 C_2(1 + 2\hat{A})$$

where the correlation between spouses is ' r_{SP} ', between parent and offspring is ' $r_{P/O}$ ' and between sibs is ' $r_{S/S}$ '. This is perhaps a little oversimplified (Vetta, 1976) but is adequate for present purposes.

The total variance of a trait is made up of environmental and genetic factors and (ignoring epistatic effects) the latter is due to the effects of dominant and additive genes (Falconer, 1981). These various components of the total variance can be calculated thus:

1. *Environmental* = $1 - C_1$

2. *Genetic*

- a. Non-additive (due to dominance) = $C_1 - C_1 C_2$

- b. Additive:

- expected under random mating = $C_1 C_2(1 - \hat{A})$

- due to assortative mating = $C_1 C_2 \hat{A}$

Using data on IQ from Burt & Howard (1956) *but purely for illustrative purposes since the actual validity of these data has recently been questioned*:

$$\text{between spouses } r_{SP} = 0.3875$$

$$\text{between parent/offspring } r_{P/O} = 0.4887$$

$$\text{between sibs } r_{S/S} = 0.5069$$

therefore:

$$C_1 C_2 = \frac{2(0.4887)}{1 + 0.3875}$$

$$= 0.7044$$

$$\hat{A} = 0.3875(0.7044)$$

$$= 0.2730$$

$$C_1 = 4(0.5069) - 0.7044(1 + 0.5460)$$

$$= 0.9386$$

Therefore the partition of the total variance is then:

1. *Environmental* = $1 - 0.9386 = 0.0614$

2. *Non-environmental*

- a. Non-additive (due to dominance)

$$= 0.9386 - 0.7044 = 0.2342$$

b. Additive:

expected under random mating

$$= 0.7044(1 - 0.2730) = 0.5121$$

due to assortative mating

$$= 0.7044(0.2730) = 0.1923$$

Thus assortative mating can have a significant effect on the genetic variance. However, as Cavalli-Sforza & Bodmer (1971) point out, the real situation may well be more complicated than such simple models would lead us to believe.

The partition of variance has also been calculated for stature and total dermal ridge count (Table 3.1). In the latter trait, as one would expect, the contribution by assortative mating is very small and not significantly different from zero.

Table 3.1 Partition of variance determined from correlations between relatives for IQ, stature and total dermal ridge count

	Correlations			Non-genetic	Partition of variance (%)		
					Genetic		
	Spouses (r_{SP})	Parent-offspring ($r_{P/O}$)	Sibs ($r_{S/S}$)		Non-additive	Additive	Random mating
IQ	0.39	0.49	0.51	7	23	51	19
Stature	0.28	0.51	0.54	—	21	62	17
Ridge count	0.05	0.48	0.50	—	9	87	4

Roberts (1977) has reviewed correlations between spouses for a vast number of physical characteristics. In general correlations are less than 0.2 for traits such as weight and stature. Social and psychological traits show much higher correlations.

Inbreeding

Two individuals are said to be *consanguineous* if they have at least one ancestor in common and, in practice, this common ancestor is usually considered to be no more remote than a great-great grandparent. The offspring of consanguineous parents are by definition *inbred*.

Determination of the coefficient of inbreeding

The coefficient of inbreeding (F) may be defined as the probability that an individual (say C) will have, at a given locus, two genes identical by descent from a common ancestor. There are a number of ways in which it may be determined, the simplest being by path analysis or isonymy.

Path analysis. If n and n' are the number of generations in the lines of descent from a common ancestor to the *parents* of individual C then

$$F = \sum \left(\frac{1}{2} \right)^{n+n'+1}$$

where summation is for each common ancestor.

Thus in the offspring of first cousins once removed (Fig. 3.2) the number of generations in line of descent from A to the mother is 3 (n), and to the father is 2 (n'). Similarly the number of generations in line of descent from B to the mother is 3 (n) and to the father is 2 (n'). Therefore:

$$\begin{aligned} F &= \left(\frac{1}{2} \right)^6 + \left(\frac{1}{2} \right)^6 \\ &= \frac{1}{32} \end{aligned}$$

In the case of X-linkage (Wright, 1922, 1950/1)

$$F = \sum \left(\frac{1}{2} \right)^{n_f}$$

where n_f = number of females in a line of descent and the summation relates to paths which have no male to male succession. A method for calculating the inbreeding coefficient for X-linked genes has been described in detail by Kudo & Sakaguchi (1963).

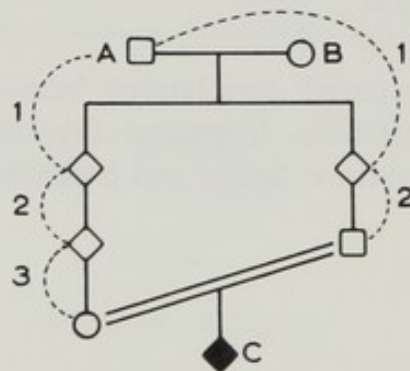


Fig. 3.2 Paths used in calculating ' F ' of a child of parents who are first cousins once removed

Isonymy. The coefficient of inbreeding can also be determined from the frequency of marriages between individuals with identical surnames (Crow & Mange, 1965) the reason being that in many societies the surname is transmitted in a regular pattern which closely corresponds to the biological ancestry. Now the frequency of isonymous pairs divided by four gives the inbreeding coefficient from random mating (F_r). Thus if:

p_i = proportion of males with a certain surname

q_i = proportion of females with a certain surname

$$\text{then } F_r = \frac{\sum p_i q_i}{4}$$

$$\text{and } F = F_n + (1 - F_n)F_r$$

$$\text{and } F_n = \frac{P - \sum p_i q_i}{4(1 - \sum p_i q_i)}$$

where P = observed proportion of isonymous marriages
 F_r = inbreeding coefficient from *random* mating
 F_n = inbreeding coefficient from *non-random* mating
 F = total inbreeding coefficient.

Thus in a study of the Hutterites, a religious isolate in North America, there were 446 marriages between 1940 and 1961 of which 87 were between individuals with the same surname. For example there were 30 males and 33 females with the surname 'Wi' therefore the *expected* number of marriages between these individuals with random mating is:

$$\begin{aligned} & \left(\frac{30}{446} \right) \left(\frac{33}{446} \right) 446 \\ & = 2.22 \end{aligned}$$

whereas 5 such marriages were observed. In this way the total number of expected isonymous marriages with random mating was 79.55 whereas the observed number was 87.

Therefore the expected proportion of isonymous marriages was:

$$\begin{aligned} & \frac{79.55}{446} \\ & = 0.178 \end{aligned}$$

and therefore

$$\begin{aligned} F_r &= \frac{0.178}{4} \\ &= 0.0445 \end{aligned}$$

Now

$$\begin{aligned} P &= \frac{87}{446} \\ &= 0.195 \\ F_n &= \frac{P - \sum p_i q_i}{4(1 - \sum p_i q_i)} \end{aligned}$$

$$\begin{aligned}
 &= \frac{0.195 - 0.178}{4(0.822)} \\
 &= 0.0052
 \end{aligned}$$

Now

$$\begin{aligned}
 F &= F_n + (1 - F_n)F_r \\
 &= 0.0052 + (1 - 0.0052)0.0445 \\
 &= 0.0495
 \end{aligned}$$

Thus the average relationship is equivalent to something between first cousins ($F = 1/16$ or 0.0625) and first cousins once removed ($F = 1/32$ or 0.0313). Almost the entire inbreeding effect is due to random marriages. The component from non-random marriages ($F_n = 0.0052$) is very small and not significantly different from zero.

The attraction of the isonymy method in estimating the coefficient of inbreeding is its simplicity. But the method should not be applied uncritically. For example, there may have been some duplication of surnames at the time the names were first introduced into the population, and assigning a family's name to an adopted child and giving any name other than the father's to an illegitimate child will affect the estimate of the coefficient of inbreeding from isonymy. Overall it seems likely that isonymy will tend to *overestimate* the actual amount of inbreeding in a given population especially when the level of inbreeding is low or the number of surnames is small (Tay & Yip, 1984).

It is of interest that surname has also been used as a 'genetic marker' in some studies (Ashley & Davies, 1966) though its value in this regard has yet to be fully assessed.

The amount of consanguinity in a population is best expressed as the *average inbreeding coefficient*

$$= \sum p_i F_i$$

where p_i is the proportion of marriages with inbreeding coefficient F_i . Thus in a study of consanguinity among French Canadians in the Province of Quebec (Laberge, 1966), out of a total of 96 marriages in the Isle aux Coudres in the Gulf of St Lawrence, 13 were consanguineous (Table 3.2). In this study the overall average inbreeding coefficient in the Province was 0.0014.

Table 3.2 Average inbreeding coefficient in the Isle aux Coudres (Laberge, 1966)

	Total	1st cousin	1st cousin once removed	2nd cousin
No. of marriages	96	1	4	8
Proportion (p_i)	—	0.0104	0.0417	0.0833
Inbreeding (F_i)	—	0.0625	0.0313	0.0156
$p_i F_i$	—	0.00065	0.00131	0.00130
		$\Sigma p_i F_i = 0.0033$		

In most Western societies the average inbreeding coefficient is always less than 0.001 but in some isolated societies it may be greater than 0.04, but this is obviously influenced by marriage customs. Thus the coefficient of inbreeding is high in communities in Southern India because of preferential uncle–niece marriages, but is low in Eskimo communities because of taboos against inbreeding in any form.

The value, in practical terms, of estimating the coefficient of inbreeding is that it allows us to predict:

1. The incidence of a particular recessive disorder in an inbred population since this is equal to

$$Fq + q^2(1 - F)$$

Thus if the incidence of a recessive disorder in a randomly mating population is 1 in 10 000 ($q^2 = 0.0001$) then among marriages in an inbred population in which F is 0.04, the expected incidence (all else being equal) will be

$$\begin{aligned} &(0.04)(0.01) + (0.0001)(0.096) \\ &= 0.000496 \end{aligned}$$

or approximately 1 in 2000.

2. The proportion of heterozygotes for a particular recessive disorder in an inbred population since

$$H_F = (1 - F)H_O$$

where H_F and H_O denote the proportion of heterozygotes in populations with and without inbreeding.

3. The 'genetic load' (defined as the proportion of the population lost by selection), because certain components of the genetic load increase linearly with the coefficient of inbreeding. However in order to calculate the genetic load in this way it is necessary to know not only the value of F but also the fitness (see p. 29) of the various genotypes and apart from one or two disorders this is rarely known with any precision. The subject of genetic load has been interestingly discussed by Fraser & Mayo (1974), Freire-Maia (1976) and Knudson (1979).

The *coefficient of relationship* (R) is a measure of the degree of genetic relationship between two individuals and may be defined as the probability that both possess an identical gene by descent from a common ancestor(s). It is equal to $(\frac{1}{2})$ to the power of the number of generations in the lines of descent from a common ancestor(s) to the individuals whose coefficient of relationship is being determined:

$$R = \sum \left(\frac{1}{2}\right)^{n+n'}$$

or $R = 2F$

That is, the inbreeding coefficient of a child is half the coefficient of

relationship of its parents. Some values of F and R are given in Table 3.3.

Table 3.3 Coefficients of inbreeding (F) and relationship (R) and probability of isonymy (P) for various consanguineous matings

Mating	F	R	P
Sibs	1/4	1/2	1
Uncle-niece, aunt-nephew	1/8	1/4	1/2
1st cousins	1/16	1/8	1/4
1st cousins once removed	1/32	1/16	1/8
2nd cousins	1/64	1/32	1/16
2nd cousins once removed	1/128	1/64	1/32
3rd cousins	1/256	1/128	1/64

Cousin marriages

With rare recessive traits, the parents of affected individuals are often related, the reason being that such individuals are more likely to carry the same genes because they have inherited them from a common ancestor. In fact the chance that first cousins will carry the same gene is 1 in 8. The frequency (C) of first-cousin marriages among the parents of children with any particular autosomal recessive disorder is (Dahlberg, 1947)

$$C = \frac{a(1 + 15q)}{a + 16q - aq}$$

where a = frequency of 1st cousin marriages in the general population.

If ' q ' is very small then approximately

$$C = \frac{a}{a + 16q}$$

Alternatively if the frequency of first-cousin marriages in the general population (a) and among parents of affected children (C) are known then the gene frequency can be estimated since

$$q = \frac{a(1 - C)}{16C - Ca - 15a}$$

Some examples of recessive disorders and the approximate frequencies of consanguinity among the parents are given in Table 3.4, where it is assumed that the frequency of first-cousin marriages in the general population is about 1 in 200. Note that the rarer a recessive disorder the more likely are the parents to be related. An increase in consanguinity among the parents of children with a particular rare disorder may therefore be used as evidence that the disorder is inherited as a recessive trait.

Table 3.4 Prevalence of first-cousin marriages among the parents of individuals with various recessive disorders

Disorder	Frequency of homozygotes (q^2)	Gene frequency (q)	% Consanguinity*	
			(1)	(2)
Alkaptonuria	1/1 000 000	0.0010	24.2	23.8
Cystinuria	1/100 000	0.0032	9.3	8.9
Albinism	1/20 000	0.0071	4.7	4.2
Phenylketonuria	1/15 000	0.0082	4.1	3.7
Cystic fibrosis	1/2000	0.0224	1.8	1.4

* Consanguinity estimated from (1) $\frac{a(1+15q)}{a+16q-aq}$ (2) $\frac{a}{a+16q}$

Gene flow

Another process by which genetic variation is introduced into a population is by gene flow. That is when individuals from outside the population contribute to the gene pool either by *migration* or *racial admixture*. Migration may result in a gene being spread in one direction only and the frequency gradient that may result is referred to as a *cline*. Thus the frequency of the gene for blood group B is very high in Asia (over 25%) but gradually decreases as one travels westward across Europe until in Britain, France and Scandinavia it is less than 10% (Mourant et al, 1976). It has been suggested that this gradient or cline is the consequence of invasions by Mongoloids who pushed westward from about A.D. 500 until A.D. 1500. Miscegenation between the invaders and the native population in which blood group B was rare or absent, led to the diffusion of the B gene across Europe (Candela, 1942). Of course it is equally possible that this gradient in the frequency of blood group B gene might have been the result of some as yet unknown selective force which followed a similar geographic gradient.

Gene flow may also mean *admixture* of two or more genetically dissimilar populations so creating a hybrid group, for example, the racial admixture which resulted in the United States from miscegenation between African Negroes and American whites or in Hawaii between Polynesians, Asiatics and Europeans.

If ' m ' is the proportion of genes at a particular locus in a hybrid population (H) which is derived from a population (P) which has miscegenated with an immigrant population (I), and if q is the gene frequency, then

$$q_H = mq_P + (1 - m)q_I$$

and therefore

$$m = \frac{|q_H - q_I|^*}{|q_P - q_I|^*}$$

*The vertical lines mean the *absolute* values of the differences, that is the differences are always positive.

This is sometimes referred to as *Bernstein's* equation. If the gene in question is absent from the immigrant population then

$$m = \frac{q_H}{q_P}$$

In studying the problem of gene flow in relation to the American Negro, Reed (1969) considered the Duffy blood group ($Fy(a+)$) a good marker in this regard. Thus the mean frequency of Fy^a gene in West Africa (I) is at most 0.030, in American whites (P) is 0.429, and in American Negroes (H), in southern California, is about 0.094, therefore

$$\begin{aligned} m &= \frac{q_H - q_I}{q_P - q_I} \\ &= \frac{0.094 - 0.030}{0.429 - 0.030} \\ &= 0.160 \end{aligned}$$

If however one assumes that Fy^a might well have been absent from the original African population, then

$$\begin{aligned} m &= \frac{q_H}{q_P} \\ &= \frac{0.094}{0.429} \\ &= 0.219 \end{aligned}$$

Thus from the evidence of Fy^a gene of the Duffy blood group system, the proportion of American Negro genes which are of American white origin is between 16 and 22% in southern California. In contrast the proportion is less than 4% in Charleston, South Carolina, which probably reflects cultural barriers to gene flow. However when data for a number of blood groups, enzymes and protein variants are considered, estimates for gene flow vary considerably even within northern states of America, being as high as 64% for G6PD in New York for example (Dyer, 1976).

It should be noted however, that in estimating ' m ' in this way, several assumptions are made (Workman et al, 1963). It is assumed that the deviation of q_H from q_I is solely due to gene flow. It disregards the possible effects of natural selection which can introduce a serious bias. Reed (1969) chose the Duffy blood group system because there was no obvious evidence of strong selection at this locus in Californian Negroes, at least as shown from studies of fetal and infant growth and viability and from adult growth and fertility. It also assumes that there is no assortative or preferential mating between the two populations. Such calculations also depend on the estimation of gene frequencies in the original populations and in the present hybrid population.

Finally it should be noted that gene flow is also related to the effective size of a population (N_e) and the coefficient of inbreeding (F). That is

$$F = \frac{1}{4N_e m + 1}$$

therefore

$$m = \frac{1 - F}{4N_e F}$$

Thus in the Dunkers, a religious isolate in the United States, F was estimated to be 0.0254 and N_e to be 90 (see p. 13), and therefore

$$\begin{aligned} m &= \frac{1 - 0.0254}{4(90)(0.0254)} \\ &= 0.1066 \end{aligned}$$

which represents the gene flow into the isolate each generation (Glass et al, 1952).

Selection

Selection has been studied more than perhaps any other aspect of human population genetics. The subject is discussed in detail elsewhere (for example Fisher, 1930; Spuhler, 1963; Bajema, 1971; Roberts, 1975), and here we shall only be concerned with how selection forces, in relation to human disease, can be measured.

Selection forces may be either natural or artificial. The former occurs under natural conditions without the intervention of man, whereas the latter is a direct consequence of man's intervention by introducing effective treatments for otherwise lethal disorders or limiting the reproduction of persons with hereditary defects. Selection forces operate at all stages of development though in humans this is usually considered in relation to postnatal development and may operate through differential mortality or differential fertility.

Selection forces disturb the Hardy-Weinberg equilibrium by increasing or decreasing fitness. In this sense fitness has a very special meaning and will be discussed later (p. 29).

The *coefficient of selection* (s) may be defined as the proportional reduction in the gametic contribution of a particular genotype to the next generation. If f is fitness

$$s = 1 - f$$

It can be shown that at equilibrium for an autosomal recessive trait

$$s = \frac{\mu}{q^2}$$

for a *rare* autosomal dominant trait

$$s = \frac{\mu}{q}$$

and for an X-linked recessive trait, in males

$$s = \frac{3\mu}{q}$$

where μ = mutation rate.

Heterozygote advantage in recessive disorders

The estimation of 's' has mainly been of interest in autosomal recessive disorders in which the apparent high frequency of affected individuals cannot be accounted for by mutation alone or genetic heterogeneity (due to different loci or multiple alleles at the same locus) and therefore it is postulated that the heterozygote may have some selective advantage. Thus the Hardy-Weinberg equilibrium is modified to

$$f_1p^2 + f_22pq + f_3q^2$$

where f_1 , f_2 and f_3 are the relative fitnesses of the three genotypes.

When a stable equilibrium is maintained by selection such that the heterozygote is 'superior' to either homozygote ($f_2 > f_1$ and f_3) this is sometimes referred to as *over-dominance*.

If the coefficients of selection in the normal homozygote and affected homozygote are s_1 and s_2 respectively then in either the entire population, in the case of an autosomal recessive disorder, or among females in the case of an X-linked recessive disorder:

	Genotypes			Total
	<i>AA</i>	<i>Aa</i>	<i>aa</i>	
Initial population	p^2	$2pq$	q^2	1
After selection	$p^2(1 - s_1)$	$2pq$	$q^2(1 - s_2)$	$1 - p^2s_1 - q^2s_2 = T$
Relative contribution to next generation	$\frac{p^2(1 - s_1)}{T}$	$\frac{2pq}{T}$	$\frac{q^2(1 - s_2)}{T}$	1

Therefore in the next generation

$$\begin{aligned} q_{n+1} &= \frac{1}{2}(Aa) + (aa) \\ &= \frac{pq + q^2(1 - s_2)}{T} \\ &= \frac{q - q^2s_2}{T} \end{aligned}$$

Therefore the change in gene frequency

$$\begin{aligned} &= q_{n+1} - q \\ &= \frac{q - q^2 s_2}{T} - q \\ &= \frac{pq(ps_1 - qs_2)}{T} \end{aligned}$$

But at equilibrium the change in gene frequency from one generation to another is zero. That is when

$$\begin{aligned} ps_1 &= qs_2 \\ (1 - q)s_1 &= qs_2 \\ s_1 - qs_1 &= qs_2 \\ q &= \frac{s_1}{s_1 + s_2} \end{aligned}$$

If, as in many serious recessive disorders, the rare homozygote is so severely affected as not to survive to have children or survives but is infertile then $s_2 = 1$ and

$$q = \frac{s_1}{s_1 + 1}$$

and therefore

$$s_1 = \frac{q}{1 - q}$$

Thus in the case of cystic fibrosis ($q^2 = 1/2000$: $q = 0.022$) $s_1 = 0.022/0.978$ or 0.0225. To maintain the present frequency of this disease the heterozygote must therefore have a fitness of 2.25% greater than the normal homozygote. To demonstrate a relative increase in fitness of this order of magnitude in heterozygotes would be very difficult but attempts have been made. For example in cystic fibrosis (Danks et al, 1965; Knudson et al, 1967), Tay-Sachs disease (Myriantopoulos & Aronson, 1966) and phenylketonuria (Woolf et al, 1975). How this is done will be discussed later (p. 29). The relationship between the frequency of affecteds ($q^2\%$) and heterozygote advantage is given in Figure 3.3.

Heterozygous advantage may well have played a part in determining the relatively high incidences in certain populations of sickle-cell anaemia and perhaps β -thalassaemia, in these cases through an increased resistance to falciparum malaria (Allison, 1964). The high incidence of some other disorders (Table 3.5) might also be due to heterozygous advantage but may be partly related to population size in former times in which mutation could

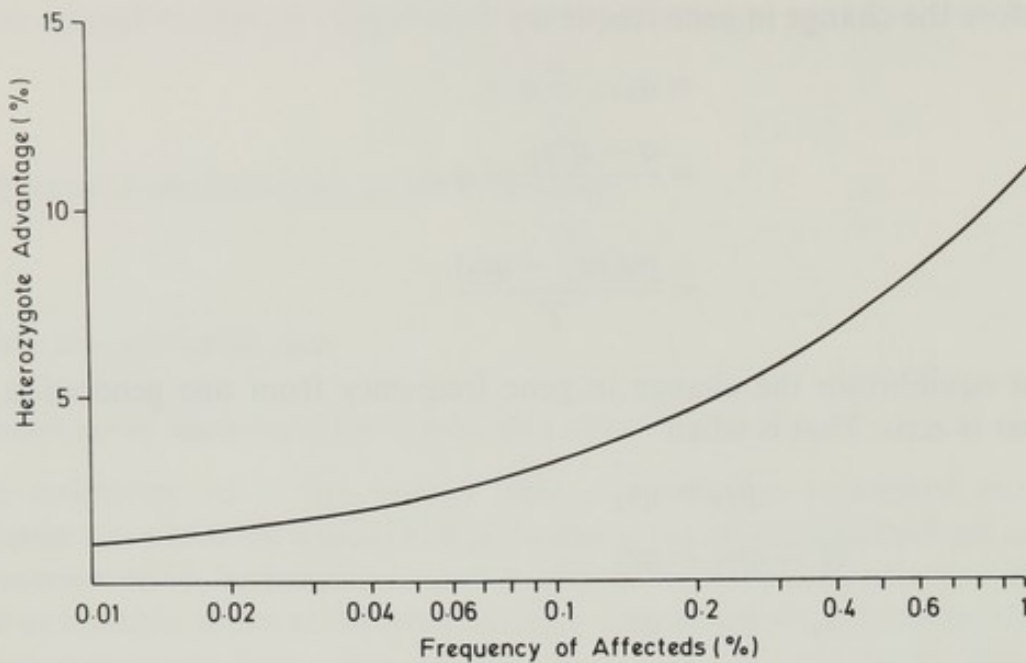


Fig. 3.3 Relationship between the frequency of affecteds ($q^2\%$) and heterozygote advantage (%).

have played a greater role in determining gene frequencies (Mayo, 1970). It is also possible that the so-called *founder effect* may have been important in certain circumstances as in the case of the high incidence of Tay-Sachs disease in a non-Jewish semi-isolate in North America (Kelly et al, 1975). Alternative explanations for some of these high gene frequencies are close linkage to genes whose alleles have been favoured by selection (*hitchhiker effect*) or epistatic interaction with an unlinked gene (Wagener & Cavalli-Sforza, 1975). Reproductive compensation may also be important (Koeslag & Schach, 1984).

Table 3.5 Estimated heterozygous advantage (in the absence of other factors) in maintaining the frequencies of certain recessive disorders in which it is presumed in earlier times, without treatment, most homozygous affecteds would not have survived to have children

Disorder	Location	Frequency of affecteds (q^2)	Gene frequency (q)	Heterozygous advantage (%)
Sickle-cell anaemia	Africa	1/25	0.200	25.0
β -Thalassaemia	Mediterranean	1/200	0.071	7.6
Cystic fibrosis	Europe	1/2000	0.022	2.2
Tay-Sachs disease	Ashkenazi Jews	1/3600	0.017	1.7
Phenylketonuria	Europe	1/15 000	0.008	0.8
	Ireland and West Scotland	1/5000	0.014	1.4

Equations relating gene frequencies to coefficients of selection under various conditions are summarized in Table 3.6.

Table 3.6 Summary of equations relating gene frequencies to coefficients of selection under various circumstances

Selection against	Initial fitnesses			Equilibrium
	<i>AA</i> (p^2)	<i>Aa</i> ($2pq$)	<i>aa</i> (q^2)	
<i>AA</i> and <i>aa</i>	$1 - s_1$	1	$1 - s_2$	$q = \frac{s_1}{s_1 + s_2}$
<i>AA</i> and <i>Aa</i>	$1 - s$ (rare)	$1 - s$	1	$q \approx \frac{\mu}{s}$
<i>aa</i>	1	1	$1 - s$ (rare)	$q = \sqrt{\frac{\mu}{s}}$

Estimation of fitness

We have seen that it is possible to estimate the coefficient of selection from gene frequencies. It can also be estimated directly by determining the fitness of various genotypes.

Biological or Darwinian fitness is a measure of the extent to which an individual with a mutant gene can reproduce so that the gene is maintained in the population. It is not synonymous with fertility per se since any offspring who die before reaching maturity will not contribute at all to the next generation. The subject, and some of the problems involved in estimating fitness have been discussed in non-mathematical terms by Clarke (1959b). Biological fitness has been variously expressed as:

1. The total number of offspring (excluding stillbirths and abortions)
2. The number of offspring who reach reproductive age (say 20)
3. The number of offspring who reach the mean age at which the parents reproduced
4. The number of offspring who complete their reproductive life (say 45).

Method three is the ideal but it is often not easy to determine because in most family studies there is insufficient data. For this reason investigators often rely on the first method.

Comparisons are usually made with comparable data (similar age and sex) from the general population or normal sibs. However, fertility is affected by many factors other than genetic including race, period of time and social class, etc. It is therefore not easy in practice to obtain a value for a general population which is strictly comparable with affected individuals. Statistical procedures have been developed for getting round these problems but they are complex and require extensive demographic data (Reed, 1959; Charlesworth & Charlesworth, 1973).

For this reason some investigators have made comparisons with normal sibs. However, sibs of affected individuals are not always representative of the general population. Further, apparently normal sibs may in fact be carriers

of the gene and subsequently develop the disease if onset is not necessarily in early life. They may also have not completed their reproductive life at the time of the study.

When there is variable age at death in affected individuals whose fitness is to be assessed, a simple approach is to determine the number of offspring per number of reproductive years (say 20 to 45). Thus in a study of benign Becker type X-linked muscular dystrophy the mean number of live births per 100 fertile years was 4.959 for affected males and 7.418 for their unaffected male sibs. The relative fitness is therefore 0.67. It should be noted that in this disease one can be certain that sibs are in fact normal by determining their serum level of creatine kinase.

Another common problem has been to assess the relative fitness of heterozygotes for severe recessive disorders in order to determine if there is any heterozygous advantage (p. 27). In this situation there is the problem that if one considers the offspring of two heterozygous parents, such matings will have been ascertained in the first place because they have produced an affected child and this might well have biased their plans for further children. Also they will have had to have had at least one child.

There are essentially two ways of dealing with these problems. Firstly, one may consider the reproductive performance of the *grandparents* of affected children since at least one maternal and one paternal grandparent will be unsuspecting heterozygotes, but comparisons must be made with comparable controls of the same generation. Such an approach has been made, for example, in the case of cystic fibrosis (Danks et al, 1965; Knudson et al, 1967) and Tay-Sachs disease (Myriantopoulos & Aronson, 1966).

Alternatively one can study the reproductive performance of couples who are both heterozygotes by determining the number and outcome of pregnancies *previous* to the affected child, as was done by Woolf et al (1975) in the case of phenylketonuria, or by making special allowances for the way in which families have been ascertained. Thus of all families in which both parents are heterozygous $(3/4)^s$ will have *no* affected children and $1 - (3/4)^s$ families will have *at least one affected child*, s being the number of children (sibs) in each family. The corrected numbers can therefore be calculated by multiplying by $1/1 - (3/4)^s$ (Table 3.7).

Table 3.7 Observed and corrected sizes of families with an autosomal recessive trait

Size of family s	Observed no.		Corrected no.	
	Families n_s	Individuals	Families*	Individuals**
1	5	5	20.0	20.0
2	3	6	6.9	13.8
3	8	24	13.8	41.4
4	4	16	5.8	23.2
5	2	10	2.6	13.0
6	1	6	1.2	7.2
Total	23	67	50.3	118.6
Mean	—	2.91	—	2.36

* $n_s \{1/1 - (3/4)^s\}$

** $s \cdot n_s \{1/1 - (3/4)^s\}$

By applying this correction for ascertainment, the mean family size is clearly *reduced*. But this assumes that there has been *complete* ascertainment of all cases in a particular population. In some instances this might be impractical and a more sophisticated method which assumes incomplete ascertainment is then necessary (Kate, 1977).

Tanaka (1974) has devised a very simple and effective way of estimating fitness. He has shown that

$$f = \frac{A_p}{A_o}$$

where A_p = frequency of the trait among *parents* of index cases
and A_o = frequency of the trait among *offspring* of index cases

The underlying principle is that, if selection against a specific disorder is sufficiently strong then the frequency of affected individuals among parents of index cases will be lower than among the offspring of index cases, and this reduction is proportional to the intensity of selection. The method is widely applicable but is particularly valuable in autosomal dominant and multifactorial disorders. Tanaka gives worked examples for a number of disorders. For example, in one study of schizophrenia $A_p = 4.38\%$, $A_o = 12.31\%$ and therefore $f = 0.36$; and in one study of neurofibromatosis $A_p = 18.27\%$, $A_o = 35.23\%$ and therefore $f = 0.52$.

In order to obtain relative fitnesses for *each sex separately*, for males this is

$$\frac{A'_p \cdot x' + x''}{A'_o \cdot 2x'}$$

where A'_p and A'_o refer respectively to the frequencies of affected individuals among *fathers* of patients and among offspring of *male* index patients, and x' and x'' are the relative frequencies of male and female patients in the general population. Similarly for females the relative fitness is

$$\frac{A''_p \cdot x' + x''}{A''_o \cdot 2x''}$$

where A''_p and A''_o refer respectively to the frequencies of affected individuals among *mothers* of patients and among offspring of *female* index patients (Tanaka, 1975).

Unfortunately the method is not particularly suitable for rare disorders because A_o is too small to be estimated precisely. Further, the method is only valid if the frequency of the disorder is roughly the same in both the parent and offspring generations. In some disorders, because of changes in accepted diagnostic criteria over the last few decades, this may present an important objection to the method.

Finally, before leaving the subject, it should be noted that fitness may not always be reduced. In Huntington's chorea, for example, evidence suggests that affected individuals are in fact more fecund than their normal sibs and members of the general population of comparable age. In one study fitness in Huntington's chorea was estimated to be 1.14 compared with the general population (Shokeir, 1975). If maintained this would result in a steady increase in the incidence of the disorder in future generations.

Fitness and incidence of X-linked disorders

In X-linked recessive disorders knowing the fitness of affected males (f) and if the fitness of carrier females is 1.0, then it is possible to determine the incidence of affected males (I) and carrier females (H) in terms of the mutation rate* since at equilibrium

$$I = \mu + H/2$$

and
$$H = 2\mu + H/2 + If$$

In a condition such as Duchenne muscular dystrophy (and many other serious X-linked recessive disorders) where $f = 0$ then

$$H = 2\mu + H/2$$

and therefore
$$H = 4\mu$$

and since
$$I = \mu + H/2$$

$$I = 3\mu$$

In Becker type muscular dystrophy an approximate value of f is 0.70. Therefore

$$H = 2\mu + H/2 + I(0.7)$$

but
$$I = \mu + H/2$$

substituting for I ,

then
$$H = 2\mu + H/2 + (\mu + H/2)0.7$$

therefore
$$H = 18\mu$$

and
$$I = 10\mu$$

Values for H and I for some other disorders are given in Table 3.8. These results have important implications in probability calculations for genetic counselling (see p. 99).

*Assuming the mutation rate is the same in males and females.

Table 3.8 The incidence of affected males and carrier females in terms of the mutation rate (μ) in X-linked recessive disorders where the fitness in carrier females is assumed to be 1.0.

Fitness of affected males	Incidence (times μ)		Example
	Affected males	Carrier females	
0.0	3	4	{ Duchenne muscular dystrophy { Lesch-Nyhan syndrome, etc.
0.1	3.4	4.7	—
0.2	3.8	5.5	—
0.3*	4.3	6.6	Neonatal hypoparathyroidism
0.4*	5	8	Vitamin D resistant rickets
0.5	6	10	—
0.6*	7.5	13	Anhidrotic ectodermal dysplasia
0.7	10	18	{ Becker muscular dystrophy { Haemophilia A
0.8*	15	28	Diabetes insipidus, pituitary type
0.9	30	58	—

* From data in Stevenson & Kerr (1967)

Mutation

The Hardy-Weinberg equilibrium depends on a constant rate of mutation, but clearly this may be increased by exposure to X-radiation or mutagenic chemicals. In discussing mutation this usually infers gene or so-called 'point' mutation though it should be remembered that chromosomal rearrangements are also a form of mutation. There are two methods for determining mutation rates: direct and indirect.

Direct method for estimating mutation rates

This method is only applicable to autosomal dominant traits which are rare and always fully penetrant and X-linked recessive disorders when the carrier state is detectable. An example of the former is provided by a study of achondroplasia.

Summarizing the results of several recent newborn surveys, Gardner (1977) found 7 cases of *true* achondroplasia of normal parents out of a total of 242 257 live births. However in each affected child the mutation could have occurred in either the gene supplied by the mother or that supplied by the father. Therefore the mutation rate *per gene*

$$= 7/2(242\,257)$$

$$= 14.4 \times 10^{-6}$$

the standard error of which is

$$= \sqrt{\frac{pq}{2N}}$$

$$= \sqrt{\frac{0.000\,014\,4 \times 0.999\,985\,6}{484\,514}}$$

$$= 5.5 \times 10^{-6}$$

Therefore the mutation rate for achondroplasia may be expressed as $14.4 \pm 5.5 \times 10^{-6}$.

The application of the direct method to an X-linked recessive disorder where it is possible to detect the carrier state, is illustrated in the case of Duchenne muscular dystrophy (DMD) in which a proportion of healthy female carriers have a raised serum level of creatine kinase. In one study (Gardner-Medwin, 1970), 22 out of 35 known carriers had raised serum levels of creatine kinase. Of 56 mothers of sporadic cases, 15 had raised levels. Thus the proportion of new mutations (mothers are non-carriers) among sporadic cases is $[56 - (35/22)15]/56$ or 0.574. Now over a 9-year period (1952–1960) 43 *sporadic* cases were born and therefore the number of new mutations is (43) (0.574) or 24.682. The total number of males born in this period who survived to age 5 (by which time almost all cases of DMD are diagnosed) was 236 200. Thus the mutation rate is $24.682/236\ 200$ or 10.5×10^{-5} . Thus if ' P ' is the proportion of sporadic cases presumed to be due to new mutations and if in a given period ' n ' is the number of sporadic cases and ' N ' the total male births, the mutation rate is equal to

$$Pn/N$$

The direct method of estimating mutation rates is not applicable to recessive traits since a mutation to a recessive gene will go unrecognized if the mutant gene is completely recessive and not manifest in the heterozygote. The method is most useful in relatively severe dominant disorders in which affected individuals often do not have offspring, so that a significant proportion of affected persons are likely to be the result of new mutations. In other situations the so-called *indirect* method is used.

Indirect method for estimating mutation rates

In dominant disorders if μ is the mutation rate (per gene per generation) then the frequency of cases due to fresh mutations is 2μ . If the reproductive fitness is ' f ' and the incidence (p. 154) of the disorder is ' I ' then in each generation the number of cases eliminated is

$$= (1 - f)I$$

In a state of equilibrium where the frequency of the condition does not change from generation to generation, then the number of cases arising as a result of new mutations must be equal to the number being eliminated because of reduced fitness. Thus

$$2\mu = I(1 - f)$$

and therefore

$$\mu = \frac{1}{2}I(1 - f)$$

Similarly it can be shown that for an autosomal recessive trait

$$\mu = I(1 - f)$$

for an X-linked dominant trait

$$\mu = \frac{2}{3}I(1 - f)$$

for an X-linked recessive trait

$$\mu = \frac{1}{3}I'(1 - f)$$

and for an holandric trait

$$\mu = I'(1 - f)$$

In the latter two cases I' represents the incidence of affected males among all males whereas I represents the incidence of affected males and females in the total population. The term $(1 - f)$ is referred to as the coefficient of selection (s) against the gene (p. 25).

It should be remembered that there are a number of problems and sources of error in estimating mutation rates by both the direct or indirect methods. Spuriously high estimates will be obtained if clinically similar but genetically different disorders are lumped together. Another problem is that both methods depend upon the accurate determination of the incidence of the disorder in the general population and this may be difficult to obtain. Further, the incidence of a disorder in a particular population may be affected by factors other than selection and mutation, i.e. by inbreeding, genetic drift, founder effect, etc. Finally, the indirect method depends on the estimation of fitness of affected individuals and this poses special problems (p. 29). The subject of spontaneous mutation in man has been critically and interestingly reviewed by Vogel & Rathenberg (1975) and Vogel (1983) where the reader will find the subject dealt with in detail.

Genetic distance

It is often of interest from a genetic or anthropological point of view to consider to what extent human populations are genetically different as a result of factors such as genetic drift, selection pressures and mutation. In order to make such comparisons, the so-called *genetic distance* between the populations is determined. A number of methods are available for computing genetic distance but perhaps the simplest, and intuitively the easiest to appreciate, depends on estimating what is referred to as *Euclidean distance*. Imagine the distribution of two measured characteristics defined by x and y coordinates in a 2-dimensional plane. If the coordinates for one individual are x_1 and y_1 and for another individual x_2 and y_2 , then the distance between the two points can be calculated according to Pythagoras' theorem:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

However in estimating genetic distances between populations, more than two characteristics per individual are considered, and in general, the greater the

number of characteristics, the more precise the estimate of genetic distance.

If n different characteristics are studied, then each individual will have values

$$x_1, x_2, \dots, x_n$$

For each of these characteristics a mean can be obtained for each population and the differences between the two populations (say A and B) can be calculated

$$\bar{x}_1^A - \bar{x}_1^B = d_1; \bar{x}_2^A - \bar{x}_2^B = d_2; \dots \bar{x}_n^A - \bar{x}_n^B = d_n.$$

Because different units of measurement often have to be used for different characteristics, comparisons are made possible by dividing each by its standard deviation (σ). The genetic distance between any two populations is then given by

$$\sqrt{\sum (d_i/\sigma_i)^2}$$

where d_i = difference between the means for each characteristic

σ_i = standard deviation for each characteristic (assumed to be approximately the same in the two populations)

This approach is useful for simple metric characteristics which are not strongly correlated. For characteristics which are continuously distributed and strongly correlated other methods for computing genetic distance have to be adopted. These include various methods developed by Mahalanobis, Sanghvi, Bhattacharyya and Edwards and Cavalli-Sforza. However the details of these methods are somewhat complex and can be found in Cavalli-Sforza & Bodmer (1971), Weiner & Huizinga (1972) and Smith (1977).

Segregation analysis

To test a particular genetic hypothesis in experimental animals one studies the progeny of controlled matings. This, of course, is not possible in human populations and the geneticist has to approach the problem indirectly by fitting probability models to family data: that is by comparing the observed proportion of affected sibs and offspring with the proportion expected according to a particular genetic hypothesis. This is referred to as *segregation analysis*. The main problems of such studies arise through the different methods of *ascertaining* families and affected individuals, and through pooling data from different families which is necessary because no single family is ever large enough to test a given genetic hypothesis. Various statistical methods have been developed in order to eliminate such biases, but it should be remembered that segregation analysis may also be complicated by factors inherent in the data itself such as incomplete families, inaccurate diagnoses and genetic heterogeneity. It therefore behoves the medical geneticist to consider these possibilities carefully before attempting to combine data from different families and applying the methods of segregation analysis.

The simplest approach to the problem is to compare the observed number of affected individuals in families with that expected assuming a particular mode of inheritance. This can be illustrated in the case of a disorder suspected of being inherited as an autosomal dominant trait, the number of affected offspring of an affected parent with a healthy spouse being compared with the expected number using χ^2 test. Thus in one study of opalescent dentine, out of a total of 112 offspring of affected parents, 52 were similarly affected whereas 56 would have been expected assuming simple dominant inheritance (Neel & Schull, 1954):

Offspring	Normal	Affected	Total
observed	60	52	112
expected	56	56	112
$(O - E)^2$	16	16	—
$\frac{(O - E)^2}{E}$	0.286	0.286	0.572

Thus χ^2 is 0.572. With one degree of freedom (p. 7) to be significant χ^2 should exceed 3.841 (see Appendix 2, p. 166). The value obtained is less than this and therefore there is no significant departure from the expected number of normal and affected offspring assuming autosomal dominant inheritance.

The significance of a departure from an expected ratio of 1 : 1 among the offspring of affected parents can be calculated quite simply from the formula given by Roberts & Pembrey (1978):

$$\chi^2 = \frac{[(A - N) - 1]^2}{A + N}$$

where A = total number of affected offspring
 N = total number of normal offspring

The subtraction of unity from the difference in the numerator is Yates' correction which has to be included when dealing with small numbers. Thus if there were 15 affected and 12 normal offspring in a number of families in each of which one of the parents was affected then

$$\begin{aligned}\chi^2 &= \frac{[(15 - 12) - 1]^2}{27} \\ &= \frac{4}{27} \\ &= 0.148\end{aligned}$$

In this example, therefore, there is no significant departure from the expected 1 : 1 ratio among the offspring in the families studied.

It should be remembered that in testing for autosomal dominant inheritance, families should be ascertained irrespective of the nature of the offspring of affected individuals, i.e. never because there are affected children in the families.

This simple approach though useful ignores two very important problems: at risk matings which by chance do not produce affected children, and the bias which may result from the manner in which families are actually ascertained for study. In the case of autosomal recessive disorders, where most of these problems arise, matings between heterozygous parents are ascertained only because they have produced affected children; however, by chance, some families where both parents are heterozygous will produce only normal children (heterozygotes and normal homozygotes) and will therefore not be detected. By selecting only families which produce affected children a bias is introduced which will result in a spuriously high proportion of affected individuals in such families. Secondly, the more affected children there are in a given sibship, the more likely it is that the sibship will be ascertained. Thus the manner in which families are ascertained is critically important in testing for recessive inheritance and determines the method of analysis to be used.

There are three ways in which families may be ascertained:

1. An exhaustive search may be made to ascertain every affected individual in the community regardless of whether or not there are any affected relatives (= *complete ascertainment*). But since it is impossible to ascertain every 'at risk' family because of the exclusion of those which have not produced affected offspring, complete ascertainment is in effect always *truncate*. Also in practice it is often impossible to ascertain every affected case in the community, in which case ascertainment is referred to as being *incomplete*.

2. Each family may have been ascertained through one, and only one, affected individual irrespective of how many affected children there may be in the family (= *single incomplete ascertainment*).

3. Some families may have been ascertained more than once through different affected sibs (= *multiple incomplete ascertainment*).

Some cases will be ascertained independently of other members of the family (so-called *probands* or index cases), but other cases will be ascertained only through probands, and these are referred to as *secondary* cases. In order to select the appropriate model for ascertainment, the *ascertainment probability* π (Greek pi) is used, which may be defined as the probability of an affected individual being a proband:

$$\pi = \frac{A}{R}$$

where in the population

A = number of affected individuals ascertained independently as probands

R = total number of affected individuals

There are a number of ways of estimating π (Simpson, 1983) perhaps the simplest being

$$\frac{\sum A(A - 1)}{\sum A(R - 1)}$$

where summation is over all families. Thus using the data given in Table 4.7, π is 0.73.

The methods of analysis depending on the mode of ascertainment may be summarised as follows:

1. *Complete ascertainment* (π approaches 1)
 - a. A priori (= direct) method
 - b. Maximum likelihood method
 - c. 'Singles' method
2. *Single incomplete ascertainment* (π approaches 0)
 - 'Sib' method
3. *Multiple incomplete ascertainment* ($0 < \pi < 1$)
 - a. Proband method
 - b. Modified 'singles' method
 - c. Maximum likelihood method.

Thus the investigator should be quite clear how families and affected individuals have been ascertained and then apply the method appropriate to the mode of ascertainment. Detailed discussions of some of these methods together with references to earlier work are given by Steinberg (1959), and Elandt-Johnson (1971).

Complete ascertainment

A priori method (Hogben, 1931, 1946). The actual number of affected individuals is compared with the number expected calculated from the truncate binomial

$$\sum \frac{sp}{1 - q^s} \cdot n_s$$

which has a variance of

$$\sum \left[\frac{spq}{1 - q^s} - \frac{s^2 p^2 q^s}{(1 - q^s)^2} \right] n_s$$

where

$$\begin{aligned} s &= \text{sibship size} \\ n_s &= \text{number of sibships of size } s \\ p &= \text{theoretical proportion, i.e. } 0.25 \\ q &= 1 - p \end{aligned}$$

Values for these two equations for various sibship sizes are given in Table 4.3 (p. 42).

For this method of analysis it is assumed that all cases in a given community have been ascertained. In practice, this is rarely possible, but an exception would be the situation when every case of a rare disorder is studied in a well defined and relatively small community.

An example of this approach would be the reported study of the so-called 'Mast syndrome' (a form of presenile dementia with motor disturbance) among the Amish, a religious isolate in the United States (Cross & McKusick, 1967).

The data from this study are summarized in Table 4.1.

Table 4.1 Sibships in which one or more persons with the Mast syndrome were offspring of unaffected parents. Individuals who died prior to age 12 have been excluded because of uncertainty as to their genotype.

Family	Affected	Normal	Total	No. of 'singles'
1	1	5	6	1
2	1	1	2	1
3	4	3	7	—
4	4	7	11	—
5	2	6	8	—
6	3	4	7	—
7	1	7	8	1
8	2	4	6	—
9	1	5	6	1
Totals	19	42	61	4

To apply the a priori method of analysis the data are set-out as shown in Table 4.2.

Table 4.2 The observed and expected numbers of affected individuals with the Mast syndrome assuming complete ascertainment

Size of sibship <i>s</i>	No. of sibships <i>n_s</i>	Total no. of individuals <i>s · n_s</i>	No. of affected individuals		Variance
			observed	expected	
2	1	2	1	1.1428	0.1224
6	3	18	4	5.4744	2.3278
7	2	14	7	4.0392	1.9405
8	2	16	3	4.4450	2.3448
11	1	11	4	2.8710	1.8053
Totals	9	61	19	17.9724	8.5408
					SE = 2.9225

The expected numbers of affecteds and the variances are calculated in the following manner. In the case of sibships of size 6 (*s* = 6), there are 3 (*n_s* = 3) of these and therefore from the data in Table 4.3 the expected number of affected individuals in these sibships is

$$\begin{aligned} & \frac{sp}{1 - q^s} \cdot n_s \\ &= (1.8248)3 \\ &= 5.4744 \end{aligned}$$

and the variance is

$$\begin{aligned} & (0.77595)3 \\ &= 2.3278 \end{aligned}$$

From the data in Table 4.2 it will be seen that the observed number of

Table 4.3 Values of $sp/1 - q^s$ and its variance for various sibship sizes (s). (From Hogben, 1946.)

s	$p = 1/4$ and $q = 3/4$		$p = 1/2 = q$	
	$\frac{sp}{1 - q^s}$	Variance	$\frac{sp}{1 - q^s}$	Variance
1	1.000	0.0000	1.000	0.000
2	1.1428	0.122 45	1.333	0.2222
3	1.2973	0.262 97	1.715	0.4898
4	1.4628	0.420 05	2.134	0.7822
5	1.6389	0.591 78	2.581	1.082
6	1.8248	0.775 95	3.047	1.379
7	2.0196	0.970 24	3.527	1.667
8	2.2225	1.1724	4.015	1.945
9	2.4328	1.3802	4.509	2.215
10	2.649	1.5917	5.005	2.478
11	2.871	1.8053	5.503	2.737
12	3.098	2.0196	6.001	2.992
13	3.329	2.2335	6.5	3.245
14	3.563	2.4464	7.0	3.497
15	3.801	2.6575	7.5	3.748
16	4.041	2.8667	8.0	3.999
17	4.282	3.0738	8.5	4.249
18	4.525	3.2787	9.0	4.500
19	4.770	3.4814	9.5	4.75
20	5.016	3.6821	10.0	5.00

affected individuals differs from the expected number by 1.0276 which is 0.3516 times the standard error. Thus there is a close agreement between the observed and expected numbers of affected sibs assuming autosomal recessive inheritance ($p = 0.25$). Tables (see p. 44) are also available which give values for $sp/1 - q^s$ for various values of $p = 0.15$ to $p = 0.35$ (Li, 1961).

Maximum likelihood method (Haldane, 1938). In this method no prior assumption is made regarding a value for 'p', i.e. 0.25 if testing for recessive inheritance. Instead the maximum likelihood estimate of 'p' is determined where

$$\frac{R}{P} = \sum \frac{sn_s}{1 - q^s}$$

which has a variance of

$$\frac{pq}{\sum \frac{(1 - q^s - spq^{s-1})sn_s}{(1 - q^s)^2}}$$

where R = number of affected individuals in all sibships
 s = sibship size
 n_s = number of sibships of size 's'

the first equation being solved for 'p' by iteration.

To apply the maximum likelihood method of analysis the data are set out as in Table 4.4.

Table 4.4 The observed and expected numbers of affected individuals with the Mast syndrome when $p = 0.25$ and $p = 0.275$

Size of sibship s	No. of sibships n_s	No. of affected individuals			Reciprocal of variance†	
		observed	expected*		$p = 0.250$	$p = 0.275$
2	1	1	1.143	1.159	3.483	3.371
6	3	4	5.475	5.790	66.213	64.782
7	2	7	4.040	4.304	55.196	53.960
8	2	3	4.446	4.764	66.694	65.096
11	1	4	2.871	3.116	51.350	49.722
Totals	9	19	17.975	19.133	242.936	236.931

* $n_s \times$ values for $p = 0.250$ and $p = 0.275$ in Table 4.5A

† $n_s \times$ values for $p = 0.250$ and $p = 0.275$ in Table 4.5B

A trial value of $p = 0.250$ is first chosen. This gives an expected number of affecteds of 17.975 (slightly different from Table 4.2 because of differences in the values of $sp/1 - q^s$ in Table 4.5A due to rounding-off). This is less than the observed number (i.e. 19). Therefore a greater value of ' p ' is chosen. When $p = 0.275$, the expected number of affecteds is 19.133 which is slightly greater than the observed number. Thus ' p ' must lie somewhere between 0.250 and 0.275. By linear interpolation when there are 19 affected individuals the corresponding value of ' p ' is 0.272. Similarly by linear interpolation when $p = 0.272$ then the reciprocal of the variance is 237.6. The variance is therefore

$$1/237.6 = 0.00421$$

and the standard error is

$$\sqrt{0.00421} = 0.0649$$

The final result may be stated as

$$p = 0.272 \pm 0.065.$$

The reciprocal of the variance is given in Table 4.5B merely for convenience in order to avoid a lot of zeros if the variance itself were used. Thus when $s = 15$ and $p = 0.250$ then the variance is 0.01323. It should be noted that here the variance relates to the estimate of ' p ' and not to the number of affected individuals as in the a priori method.

Table 4.5A Values of $sp/1 - q^s$ for various values of 'p' and sibships of size 's'. (From Li, 1961.)

<i>s</i>	<i>p</i> = 0.15	<i>p</i> = 0.20	<i>p</i> = 0.225	<i>p</i> = 0.25	<i>p</i> = 0.275	<i>p</i> = 0.30	<i>p</i> = 0.35
2	1.081	1.111	1.127	1.143	1.159	1.176	1.212
3	1.166	1.230	1.263	1.297	1.333	1.370	1.448
4	1.255	1.355	1.408	1.463	1.520	1.579	1.704
5	1.348	1.487	1.562	1.639	1.719	1.803	1.980
6	1.445	1.626	1.723	1.825	1.930	2.040	2.271
7	1.545	1.772	1.893	2.020	2.152	2.288	2.576
8	1.649	1.923	2.069	2.223	2.382	2.547	2.892
9	1.757	2.079	2.252	2.433	2.620	2.814	3.217
10	1.868	2.241	2.441	2.649	2.865	3.087	3.548
11	1.982	2.407	2.635	2.871	3.116	3.367	3.884
12	2.098	2.577	2.833	3.098	3.371	3.651	4.224
13	2.218	2.751	3.035	3.329	3.631	3.938	4.567
14	2.341	2.929	3.241	3.563	3.893	4.229	4.912
15	2.465	3.109	3.450	3.801	4.158	4.521	5.258

Table 4.5B Reciprocal of variances for various values of 'p' and sibships of size 's'. (From Li, 1961.)

<i>s</i>	<i>p</i> = 0.15	<i>p</i> = 0.20	<i>p</i> = 0.225	<i>p</i> = 0.25	<i>p</i> = 0.275	<i>p</i> = 0.30	<i>p</i> = 0.35
2	4.583	3.858	3.640	3.483	3.371	3.295	3.229
3	9.600	8.188	7.774	7.480	7.278	7.149	7.061
4	15.038	12.967	12.367	11.948	11.665	11.489	11.386
5	20.882	18.163	17.380	16.833	16.463	16.230	16.077
6	27.113	23.739	22.761	22.071	21.594	21.279	21.008
7	33.708	29.651	28.458	27.598	26.980	26.545	26.069
8	40.641	35.856	34.415	33.347	32.548	31.947	31.172
9	47.885	42.308	40.578	39.258	38.229	37.416	36.257
10	55.412	48.962	46.896	45.274	43.969	42.898	41.282
11	63.191	55.775	53.322	51.350	49.722	48.356	46.228
12	71.193	62.706	59.815	57.445	55.456	53.762	51.088
13	79.389	69.721	66.341	63.531	61.146	59.103	55.865
14	87.749	76.789	72.873	69.585	66.780	64.372	60.566
15	96.247	83.881	79.387	75.592	72.349	69.568	65.203

'Singles' method (Li & Mantel, 1968). This is the simplest method of testing for recessive inheritance when there is complete ascertainment and according to the originators it is just as reliable as more involved methods. The method consists simply of determining the number of sibships in which there is only one affected individual (= 'singles') and

$$P = \frac{R - J}{T - J}$$

where R = number of affected individuals in all sibships
 T = total number of individuals in all sibships
 J = number of 'singles'

Using the data as presented in Table 4.1 (p. 41)

$$p = \frac{19 - 4}{61 - 4}$$

$$= \frac{15}{57}$$

$$= 0.263$$

Note the close agreement with the value obtained by the maximum likelihood method. Unfortunately though this is a very simple method for calculating 'p', the determination of the variance is complicated. Li & Mantel (1968) have shown that the variance is

$$\frac{1}{W}$$

where

$$W = \sum n_s w_s$$

where

$$w_s = \frac{s}{pq} \cdot \frac{(1 - q^{s-1})^2}{(1 - q^s)[1 - q^s + (s - 2)pq^{s-1}]}$$

Fortunately Li & Mantel (1968) have provided tables of 'w' for various 'p' values (Table 4.6). For example in the above example the following values of $w_s n_s$ are obtained:

	Sibship size s	No. of sibships n_s	w_s	$w_s n_s$	
If $p = 0.26$					
	2	1	3.43	3.43	
	6	3	21.18	63.54	
	7	2	26.49	52.98	
	8	2	32.06	64.12	
	11	1	49.72	49.72	$\therefore \sum w_s n_s = 233.79$
If $p = 0.27$					
	2	1	3.39	3.39	
	6	3	21.01	63.03	
	7	2	26.29	52.58	
	8	2	31.81	63.62	
	11	1	49.18	49.18	$\therefore \sum w_s n_s = 231.80$

and by interpolation $\sum w_s n_s = 233.18$. Therefore the variance is $1/233.18$ and the standard error of the estimate is $1/\sqrt{233.18}$ or 0.065.

It should be noted that in assuming complete ascertainment this tends to *overestimate* the value of p .

Single incomplete ascertainment (Fisher, 1934)

The underlying assumption in this method (sometimes referred to as the 'sib' method) is that each affected individual has a very small chance of being ascertained, and therefore there is never more than one proband per family. The probability of ascertaining each family is proportional to the number of affected individuals in the family.

In this case

$$p = \frac{R - N}{T - N}$$

which has a variance of $\frac{pq}{T - N}$

where R = number of affected individuals in all sibships
 T = total number of individuals in all sibships
 N = number of sibships

Table 4.6 'Singles' method of estimating the segregation ratio under complete ascertainment. (From Li & Mantel, 1968.)

$$\text{Values of } w = \frac{s}{pq} \frac{(1 - q^{s-1})^2}{(1 - q^s)[1 - q^s + (s - 2)pq^{s-1}]}$$

<i>s</i>	<i>p</i> = 0.16	<i>p</i> = 0.17	<i>p</i> = 0.18	<i>p</i> = 0.19	<i>p</i> = 0.20	<i>p</i> = 0.21	<i>p</i> = 0.22	<i>p</i> = 0.23	<i>p</i> = 0.24	<i>p</i> = 0.25
2	4.40	4.23	4.09	3.97	3.86	3.76	3.68	3.60	3.54	3.48
3	9.13	8.81	8.54	8.29	8.08	7.90	7.74	7.60	7.48	7.37
4	14.21	13.74	13.34	12.99	12.68	12.41	12.18	11.98	11.81	11.66
5	19.64	19.03	18.50	18.04	17.65	17.30	17.01	16.75	16.53	16.34
6	25.42	24.67	24.02	23.46	22.97	22.55	22.19	21.88	21.61	21.38
7	31.54	30.64	29.88	29.21	28.64	28.14	27.71	27.34	27.01	26.73
8	37.98	36.95	36.06	35.29	34.62	34.04	33.53	33.09	32.70	32.36
9	44.74	43.56	42.54	41.66	40.88	40.21	39.61	39.08	38.62	38.20
10	51.79	50.46	49.30	48.29	47.40	46.61	45.90	45.28	44.71	44.20
11	59.11	57.61	56.30	55.14	54.12	53.20	52.37	51.62	50.93	50.30
12	66.68	65.00	63.51	62.19	61.00	59.93	58.95	58.06	57.23	56.46
13	74.47	72.58	70.90	69.39	68.02	66.77	65.62	64.55	63.56	62.63
14	82.44	80.32	78.42	76.70	75.13	73.68	72.33	71.07	69.89	68.78
15	90.58	88.20	86.06	84.10	82.29	80.62	79.05	77.58	76.20	74.88
16	98.84	96.19	93.77	91.55	89.49	87.56	85.76	84.06	82.45	80.93
17	107.21	104.24	101.53	99.02	96.68	94.49	92.43	90.49	88.65	86.92
18	115.65	112.35	109.31	106.49	103.85	101.38	99.05	96.85	94.78	92.83
19	124.15	120.48	117.09	113.94	110.99	108.22	105.61	103.16	100.85	98.67
20	132.67	128.62	124.86	121.35	118.07	115.00	112.11	109.39	106.84	104.45

Table 4.6—continued

<i>s</i>	<i>p</i> = 0.25	<i>p</i> = 0.26	<i>p</i> = 0.27	<i>p</i> = 0.28	<i>p</i> = 0.29	<i>p</i> = 0.30	<i>p</i> = 0.31	<i>p</i> = 0.32	<i>p</i> = 0.33	<i>p</i> = 0.34
2	3.48	3.43	3.39	3.35	3.32	3.30	3.27	3.26	3.24	3.23
3	7.37	7.28	7.20	7.13	7.08	7.03	7.00	6.97	6.95	6.95
4	11.66	11.53	11.42	11.34	11.26	11.20	11.16	11.13	11.11	11.10
5	16.34	16.18	16.04	15.93	15.84	15.77	15.72	15.68	15.66	15.65
6	21.38	21.18	21.01	20.88	20.76	20.67	20.60	20.55	20.52	20.50
7	26.73	26.49	26.29	26.12	25.97	25.85	25.75	25.67	25.61	25.57
8	32.36	32.06	31.81	31.58	31.38	31.21	31.07	30.94	30.83	30.74
9	38.20	37.83	37.50	37.20	36.94	36.70	36.48	36.29	36.11	35.95
10	44.20	43.73	43.31	42.92	42.56	42.23	41.92	41.64	41.38	41.13
11	50.30	49.72	49.18	48.67	48.20	47.76	47.35	46.96	46.59	46.25
12	56.46	55.73	55.06	54.42	53.82	53.25	52.72	52.21	51.73	51.28
13	62.63	61.75	60.92	60.13	59.39	58.68	58.01	57.38	56.78	56.22
14	68.78	67.72	66.72	65.78	64.88	64.03	63.22	62.46	61.74	61.07
15	74.88	73.64	72.47	71.35	70.29	69.30	68.35	67.46	66.63	65.84
16	80.93	79.50	78.14	76.85	75.63	74.48	73.40	72.39	71.43	70.55
17	86.92	85.28	83.73	82.27	80.89	79.60	78.38	77.24	76.18	75.19
18	92.83	90.99	89.25	87.62	86.09	84.65	83.30	82.05	80.88	79.80
19	98.67	96.63	94.71	92.91	91.22	89.64	88.17	86.81	85.54	84.36
20	104.45	102.20	100.10	98.13	96.30	94.59	93.00	91.53	90.16	88.90

Table 4.6—continued

s	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.20$	$p = 0.25$	$p = 0.30$	$p = 0.35$	$p = 0.40$	$p = 0.45$	$p = 0.50$
2	11.07	6.16	4.58	3.86	3.48	3.30	3.23	3.26	3.36	3.56
3	22.42	12.61	9.50	8.08	7.37	7.03	6.94	7.04	7.29	7.71
4	34.05	19.38	14.76	12.68	11.66	11.20	11.11	11.28	11.68	12.30
5	45.95	26.46	20.36	17.65	16.34	15.77	15.66	15.88	16.36	17.08
6	58.15	33.86	26.30	22.97	21.38	20.67	20.51	20.69	21.16	21.86
7	70.62	41.57	32.58	28.64	26.73	25.85	25.54	25.60	25.93	26.52
8	83.39	49.60	39.18	34.62	32.36	31.21	30.67	30.49	30.61	31.02
9	96.44	57.94	46.11	40.88	38.20	36.70	35.81	35.31	35.16	35.37
10	109.78	66.60	53.33	47.40	44.20	42.23	40.91	40.03	39.58	39.61
11	123.42	75.56	60.84	54.12	50.30	47.76	45.93	44.64	43.89	43.76
12	137.35	84.83	68.61	61.00	56.46	53.25	50.86	49.14	48.12	47.86
13	151.57	94.39	76.62	68.02	62.63	58.68	55.69	53.56	52.29	51.92
14	166.09	104.24	84.83	75.13	68.78	64.03	60.44	57.91	56.42	55.95
15	180.90	114.37	93.23	82.29	74.88	69.30	65.11	62.21	60.51	59.97
16	196.01	124.76	101.79	89.49	80.93	74.48	69.72	66.47	64.59	63.98
17	211.41	135.41	110.48	96.68	86.92	79.60	74.28	70.70	68.65	67.99
18	227.09	146.30	119.27	103.85	92.83	84.65	78.79	74.91	72.70	71.99
19	243.08	157.41	128.15	110.99	98.67	89.64	83.28	79.11	76.75	76.00
20	259.35	168.73	137.08	118.07	104.45	94.59	87.74	83.29	80.80	80.00

Table 4.6—continued

<i>s</i>	<i>p</i> = 0.50	<i>p</i> = 0.55	<i>p</i> = 0.60	<i>p</i> = 0.65	<i>p</i> = 0.70	<i>p</i> = 0.75	<i>p</i> = 0.80	<i>p</i> = 0.85	<i>p</i> = 0.90	<i>p</i> = 0.95
2	3.56	3.84	4.25	4.82	5.64	6.83	8.68	11.86	18.37	38.19
3	7.71	8.31	9.13	10.23	11.74	13.85	17.01	22.21	32.44	62.71
4	12.30	13.14	14.26	15.71	17.66	20.36	24.37	31.01	44.28	84.17
5	17.08	18.04	19.29	20.93	23.14	26.28	31.05	39.13	55.53	105.26
6	21.86	22.81	24.10	25.84	28.28	31.86	37.44	47.04	66.66	126.32
7	26.52	27.40	28.67	30.51	33.21	37.29	43.73	54.90	77.78	147.37
8	31.02	31.80	33.07	35.05	38.05	42.65	50.00	62.74	88.89	168.42
9	35.37	36.06	37.37	39.51	42.84	48.00	56.25	70.59	100.00	189.47
10	39.61	40.24	41.60	43.93	47.61	53.33	62.50	78.43	111.11	210.53
11	43.76	44.35	45.80	48.34	52.38	58.67	68.75	86.27	122.22	231.58
12	47.86	48.44	49.98	52.74	57.14	64.00	75.00	94.12	133.33	252.63
13	51.92	52.50	54.16	57.14	61.90	69.33	81.25	101.96	144.44	273.68
14	55.95	56.55	58.33	61.54	66.67	74.67	87.50	109.80	155.56	294.74
15	59.97	60.60	62.50	65.93	71.43	80.00	93.75	117.65	166.67	315.79
16	63.98	64.64	66.67	70.33	76.19	85.33	100.00	125.49	177.78	336.84
17	67.99	68.68	70.83	74.73	80.95	90.67	106.25	133.33	188.89	357.89
18	71.99	72.73	75.00	79.12	85.71	96.00	112.50	141.18	200.00	378.95
19	76.00	76.77	79.17	83.52	90.48	101.33	118.75	149.02	211.11	400.00
20	80.00	80.81	83.33	87.91	95.24	106.67	125.00	156.86	222.22	421.05

If we apply this approach to the data in Table 4.1 (p. 41) then

$$\begin{aligned} p &= \frac{19 - 9}{61 - 9} \\ &= \frac{10}{52} \\ &= 0.192 \end{aligned}$$

and the standard error of this estimate is

$$\begin{aligned} &\sqrt{\frac{pq}{T - N}} \\ &= \sqrt{\frac{(0.192)(0.808)}{52}} \\ &= 0.055 \end{aligned}$$

The estimate of ' p ' calculated in this way would therefore appear to be less than expected. When inappropriately applied this method in fact tends to *underestimate* the value of ' p '. However in this example the value obtained has wide 95% confidence limits because the numbers are small (i.e. $0.192 \pm (1.96)(0.055)$ or 0.084 to 0.300), which would accommodate a theoretical value for ' p ' of 0.25.

Multiple incomplete ascertainment (Bailey, 1951; Morton 1959)

In practice, ascertainment varies being somewhere between complete and single incomplete in which case the method of multiple incomplete ascertainment is the most appropriate for analysis, particularly of families who present to the clinician. There is often more than one proband per sibship but not all affected individuals are probands. Under such circumstances the simplest method of determining the proportion of affected sibs of probands is to count each sibship once for each time it has been independently ascertained, omitting the proband each time. This is sometimes referred to as *Weinberg's 'proband' method*.

The method is illustrated in data from a random sample of families with phenylketonuria (Table 4.7).

In this case

$$\begin{aligned} p &= \frac{8}{32} \\ &= 0.25 \end{aligned}$$

Table 4.7 Analysis of family data assuming multiple incomplete ascertainment. The proband in each sibship is indicated by an arrow (normal: male □, female ○; affected: male ■, female ●).

Family	Number in sibship		
	Probands	Affected sibs	Total sibs
1	1	0	3
2	1	1	2
3	1	0	2
4	1	1	4
5	2	{ 1 1	4 4
6	1	1	1
7	1	1	4
8	1	0	3
9	1	0	1
10	2	{ 1 1	2 2
	Total	8	32

Davie (1979) has proposed a modification of the 'singles' method for use when there is multiple incomplete ascertainment. Here

$$p = \frac{R - J}{T - J}$$

Where R and T are as before, but J is here the number of *sibships with one proband* and not the number with only one affected individual as in the method of Li & Mantel. The variance of this estimate of ' p '

$$= \frac{(R - J)(T - R)}{(T - J)^3} + \frac{2Q(T - R)^2}{(T - J)^4}$$

where Q is the number of sibships with two probands. Thus using the data in Table 4.7

$$\begin{aligned} p &= \frac{16 - 8}{36 - 8} \\ &= 0.286 \end{aligned}$$

and the variance

$$= \frac{(16 - 8)(36 - 16)}{(36 - 8)^3} + \frac{4(36 - 16)^2}{(36 - 8)^4}$$

$$= 0.010$$

The final result may be stated as

$$p = 0.286 \pm 0.100$$

This is a very simple yet efficient method but is only applicable when probands are ascertained independently of other affected sibs.

Morton has developed what is probably the best method of segregation analysis when ascertainment is incomplete which estimates both π and p using a maximum likelihood approach. The calculations are complex, but a computer program (SEGRAN) is available (Morton et al, 1983).

X-Linked inheritance

In X-linked inheritance (whether dominant or recessive) for rare disorders there is never male to male transmission. Simple pedigree inspection may therefore exclude the possibility of X-linked inheritance.

In X-linked *dominant* inheritance, if fully penetrant, *all* the daughters of affected males will be affected. In the case of affected females, on average, half their daughters and half their sons will be affected. A departure from the expected 1 : 1 ratio of affected to normal offspring of affected females may be tested for in the same way as for autosomal dominant inheritance (p. 37).

In X-linked *recessive* inheritance *all* the daughters of affected males will be carriers. In the case of carrier females, on average, half their daughters will also be carriers and half their sons will be affected. A departure from the expected 1 : 1 ratio of affected to normal sons of *known* carriers may be tested for as in the case of autosomal dominant inheritance (p. 37) provided the selection of carriers was because they were the daughters of affected males. If they have been selected in any other way this introduces biases into the calculations which would have to be taken into account, for example if carriers have been selected because they have had at least one affected son and also another affected maternal male relative. In this situation one might apply the method of multiple incomplete ascertainment to the male progeny of such carriers.

In serious disorders where affected males are infertile or do not survive to have children it may be very difficult to prove X-linkage. However there are statistical methods, though somewhat complex, for getting round this problem (Morton & Chung, 1959). Otherwise one may have to resort to evidence other than segregation analysis; for example, the same woman may have had affected sons by more than one father, though this does not exclude autosomal dominant inheritance with male limitation. The best proof is the

demonstration of linkage with an X-linked marker trait such as Xg blood group, glucose-6-phosphate dehydrogenase deficiency, colour blindness, or an X-linked restriction fragment length polymorphism.

Multifactorial inheritance

In many common disorders (e.g. diabetes mellitus, schizophrenia, peptic ulcer and hypertension) and a number of congenital malformations (e.g. spina bifida and anencephaly, congenital pyloric stenosis and congenital dislocation of the hip) there is a definite familial tendency, the proportion of affected relatives being greater than in the general population but the proportion of affected relatives is often only of the order of 5% or less, and therefore much less than would be expected on a simple unifactorial basis. The most likely explanation is that these disorders are inherited on a *multifactorial* basis (Bishop, 1983). This implies that the cause is partly environmental and partly due to the effects of many genes each of small effect.

If the observed familial aggregation in a particular disorder is suspected of being the result of multifactorial inheritance this may be tested for in a number of ways. It must always be remembered, however, that a confounding feature will be if there is genetic heterogeneity in the disorder being studied. This possibility must be carefully considered and, as far as possible, excluded before combining data from different individuals and their families.

Tests for multifactorial inheritance

A number of models have been proposed for multifactorial inheritance but the one which is most widely used is referred to as the 'threshold model' (Falconer, 1965). According to this model it is assumed that there is some underlying graded attribute which is related to the causation of a particular disorder or congenital malformation. This is referred to as the individual's *liability*, which includes not only his genetic predisposition but also the environmental circumstances which render him more or less likely to develop the disease. According to the model the curve of liability has a normal distribution in both the general population and relatives of probands but the curve for relatives is shifted to the right because they have a higher mean liability (Fig. 5.1). The point on the curve beyond which all individuals are affected is the *threshold*. In the general population the proportion above the threshold is the population frequency and among relatives the proportion above the threshold is the familial frequency.

There are several consequences of such a model and if these are demonstrable in a particular family study it indicates that the disorder in question is probably inherited on a multifactorial basis.

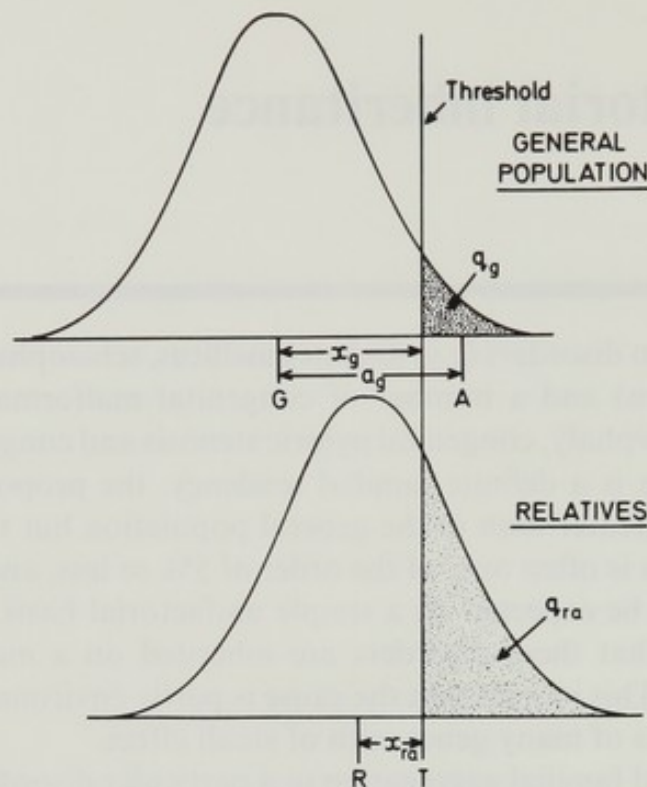


Fig. 5.1 Hypothetical curves of liability in the general population and in relatives of probands

1. The fall off in frequency from first-degree to second-degree to third-degree relatives can be predicted from the threshold model and will be greater than that predicted on the basis of unifactorial inheritance. Examples of this phenomenon have been given by Carter (1976).

2. The frequency will be greatest among the relatives of more severely affected individuals because presumably they are more extreme deviants along the curve of liability.

3. The frequency among sibs born subsequent to index cases will be greater the more affected relatives there are in a family, presumably because this indicates that there are more abnormal genes segregating in the family and/or the family has been more exposed to a precipitating environmental factor(s).

4. When there is a sex difference in the population frequency, the frequency among relatives of affected individuals of the less frequently affected sex will be greater than the frequency among relatives of affected individuals of the more frequently affected sex. This is presumably because affected individuals of the less frequently affected sex will tend to be more extreme deviants from the population mean and so the risk to their relatives will be correspondingly higher.

Other expectations of multifactorial inheritance, though *not* directly consequent on the threshold model, are:

5. The *upper limit of the frequency* among first-degree relatives is approximately equal to \sqrt{q} or $q^{\frac{1}{2}}$, among second-degree relatives is $q^{\frac{1}{4}}$, and among third-degree relatives is $q^{\frac{1}{8}}$, where 'q' is the frequency in the general population (Edwards, 1960, 1976). Note that it is customary in discussing multifactorial inheritance for 'q' to denote the frequency of the disorder and not gene frequency.

6. The *relative frequency* (referred to as 'K' by Penrose, 1953b) is, for each sex, the frequency in relatives divided by the frequency in the general population. The observed relative frequencies can be compared with the expected values for various modes of inheritance.

Thus the expected relative frequencies in *sibs* are:

$$\frac{1}{2q} \text{ for an autosomal dominant trait}$$

$$\frac{1}{4q} \text{ for an autosomal recessive trait}$$

$$\frac{1}{\sqrt{q}} \text{ for a multifactorial trait.}$$

Thus in one study of sacro-iliitis (a manifestation of ankylosing spondylitis) the results in Table 5.1 were obtained. The fairly close agreement between the observed relative frequencies and those expected for multifactorial inheritance suggests that this disorder is inherited on this basis.

Table 5.1 Frequencies and relative frequencies of sacro-iliitis in sibs (data from Emery & Lawrence, 1967)

	Frequency		Observed (s/q)	Relative frequency Expected		
	Gen. pop. (q)	Sibs. (s)		Dominant (1/2q)	Recessive (1/4q)	Multi- factorial (1/√q)
Males	0.049 18	0.1585	3.22	10.17	5.08	4.51
Females	0.015 15	0.1666	10.99	33.00	16.50	8.12

Estimation of heritability from family studies

Having decided that the disorder in question appears to be inherited on a multifactorial basis, it is useful to estimate the heritability. This may be defined as the proportion of the total phenotypic variance (genetic and non-genetic) which is due to additive genetic variance. It is therefore expressed as a percentage and abbreviated to the symbol 'h²'. The greater the value for the

heritability the greater the contribution of genetic factors to aetiology. There are, however, some important precautions to be borne in mind in estimating the heritability.

The estimation of heritability is only meaningful if there is no genetic heterogeneity in the disorder being studied (which in some cases increasingly seems less likely) and if no major gene contributes to the causation of the disorder. If a dominant gene contributes significantly to aetiology then the estimated heritability may exceed 100%. If a recessive gene contributes significantly to aetiology then the estimated heritability from sibs will be much higher than that from parents and children. If, therefore, a 'reasonable' estimate of heritability is obtained, and this is roughly the same for sibs as for parents and children, then it would seem likely that the disorder in question is inherited on a multifactorial basis. Some estimates of heritability are given in Table 5.2.

Table 5.2 Estimates of heritability for various disorders affecting man

Disorder	Frequency (%)	Heritability
Schizophrenia	1	85
Asthma	4	80
Cleft lip \pm cleft palate	0.1	76
Pyloric stenosis (congenital)	0.3	75
Ankylosing spondylitis	0.2	70
Club foot (congenital)	0.1	68
Coronary artery disease	3	65
Hypertension (essential)	5	62
Dislocation of the hip (congenital)	0.1	60
Anencephaly and spina bifida	0.5	60
Peptic ulcer	4	37
Congenital heart disease (all types)	0.5	35

In estimating heritability there are two important possible sources of error. Firstly, since heritability is estimated from the degree of resemblance between relatives, expressed as a correlation coefficient, a sharing of common environment by family members may result in the estimate being too high due to non-genetic causes of resemblance between relatives. This error is likely to affect sibs more than other relatives. For this reason it is therefore important to derive estimates from different kinds of relatives and to measure the frequency in relatives reared or living apart and in unrelated individuals living together, such as spouses. In this way it may be possible to assess the contribution from shared environmental factors. The second source of error only occurs when estimates are based on full sibs. This is due to the fact that non-additive genetic variance contributes to correlations between full sibs. For these two reasons heritability estimates should ideally be based not only on sibs but also on parents and offspring and, where possible, on second- and third-degree relatives, though clearly this is usually a counsel of perfection.

Finally, in estimating heritability it is assumed that the variance of liability

is the same in all groups being compared. It is therefore important that the 'general population' should be representative of the population from which affected individuals and their relatives are selected.

Calculation of heritability

In practice, the most usual situation is that in which the frequency of the disorder has been estimated in the general population ('g') and in relatives of affected individuals ('ra') (Method I in Falconer, 1965).

If A = affected individuals in a sample

N = total number of individuals in the sample

q = frequency = A/N

$p = 1 - q$

x = deviation of the threshold from the mean of the population

a = deviation of the mean of affecteds from the mean of the population

r = correlation between relatives and probands.

V = sampling variance of r

$$W = \frac{p}{a^2 A}$$

then

$$r = \frac{x_g - x_{ra}}{a_g}$$

and $h^2 = r$ for identical (MZ) twins

= $2r$ for first-degree relatives and non-identical (DZ) twins

= $4r$ for second-degree relatives

= $8r$ for third-degree relatives.

That is $h^2 = r/R$ where 'R' is the coefficient of relationship (see p. 21).

From tables of the normal distribution, given a frequency 'q', it is possible to determine the normal deviate 'x' (single-tailed), in standard deviation units, of the threshold from the population mean and also 'a' the deviation of the mean of the affecteds from the population mean (Appendix 5). If the frequency in the general population is assumed to have been estimated without serious error (see Falconer, 1965) then the variance can be calculated thus:

$$V \cong \left(\frac{1}{a}\right)_g^2 W_{ra}$$

and

$$\text{SE } h^2 = 2\sqrt{V} \quad \text{for first-degree relatives}$$

$$= 4\sqrt{V} \quad \text{for second-degree relatives}$$

$$= 8\sqrt{V} \quad \text{for third-degree relatives.}$$

If the frequency of a disorder differs in the two sexes, then the sexes of both

probands and relatives must be treated separately giving four estimates of heritability: male relatives of male probands, female relatives of male probands, male relatives of female probands and female relatives of female probands. Here we have three frequencies: general population comparable to affected individuals (' g '), general population comparable with relatives (' gr ') and relatives of affected individuals (' ra '). In this situation

$$r = \frac{x_{gr} - x_{ra}}{a_g}$$

and

$$V \simeq \left(\frac{1}{a}\right)_g^2 (W_{gr} + W_{ra})$$

If the four separate estimates of the heritability do not differ significantly they can be combined into a single estimate by weighting each by the reciprocal of its sampling variance and taking a weighted mean.

Another situation is when the data consist of the frequencies in relatives of affected individuals (' ra ') and in relatives of unaffected controls matched for age and sex with the affected individuals (' c '). In this situation

$$r = \frac{p_c(x_c - x_{ra})}{a_c}$$

and

$$V \simeq \left(\frac{p}{a}\right)_c^2 W_{ra}$$

Worked examples involving these various methods are given by Falconer (1965) and also the special case in which there is variable age of onset (Falconer, 1967). The reader can be no better advised than to refer to the original publications. However, a simple example may perhaps be helpful in illustrating the method of calculation.

Wynne-Davies (1970) has made a study of the frequency of congenital dislocation of the hip in various relatives of affected individuals. If we use the data on so-called 'late-diagnosis' cases the population frequency is about one per 1000, i.e. $q_g = 0.1\%$. From Appendix 5, for $q_g = 0.1\%$, values for x_g and a_g are 3.090 and 3.367. Among first-degree relatives there were 35 affected individuals out of 1777, i.e. $q_{ra} = 1.97\%$. From Appendix 5, for $q_{ra} = 1.97\%$, values for x_{ra} and a_{ra} are 2.060 and 2.426.

Since

$$r = \frac{x_g - x_{ra}}{a_g}$$

therefore

$$r = \frac{3.090 - 2.060}{3.367}$$

$$= 0.306$$

$$h^2 = 2r \text{ for first-degree relatives}$$

therefore

$$h^2 = 61.2\%$$

Now

$$\begin{aligned} V &= \left(\frac{1}{a}\right)_g^2 W_{ra} \\ &= \left(\frac{1}{a}\right)_g^2 \left(\frac{p}{a^2 A}\right)_{ra} \\ &= \left(\frac{1}{3.367}\right)^2 \left(\frac{0.9803}{(2.426)^2 35}\right) \\ &= 0.000420 \end{aligned}$$

$$\text{SE } h^2 = 2\sqrt{V} \text{ for first-degree relatives}$$

therefore

$$\text{SE } h^2 = 0.041$$

or

$$4.1\%$$

Similarly the heritability and the standard error can be calculated for second- and third-degree relatives (Table 5.3), the estimates obtained being 45.6 ± 9.8 and 46.4 ± 26.3 .

In these calculations it has been assumed that the frequency in the general population (0.10 %) is known without error, i.e. that the number of affecteds upon which q_g was estimated was very large. This of course simplifies the situation but tends to reduce the estimated standard errors slightly.

The three estimates of the heritability can be combined by weighting each by the reciprocal of its sampling variance and taking a weighted mean, i.e. by dividing each of the individual heritability estimates by its variance and summing, and dividing this by the sum of the reciprocals of the variances. Thus

$$\begin{aligned} \text{weighted mean of } h^2 &= \frac{\frac{h_1^2}{(\text{SE}_1)^2} + \frac{h_2^2}{(\text{SE}_2)^2} + \frac{h_3^2}{(\text{SE}_3)^2}}{\frac{1}{(\text{SE}_1)^2} + \frac{1}{(\text{SE}_2)^2} + \frac{1}{(\text{SE}_3)^2}} \\ &= \frac{\frac{61.2}{(4.1)^2} + \frac{45.6}{(9.8)^2} + \frac{46.4}{(26.3)^2}}{\frac{1}{(4.1)^2} + \frac{1}{(9.8)^2} + \frac{1}{(26.3)^2}} \\ &= 58.6\% \end{aligned}$$

The sampling variance of this combined estimate is approximately the reciprocal of the sum of the weights. Thus

$$\begin{aligned}
 \text{SE of the weighted mean} &= \frac{1}{\sqrt{\frac{1}{(\text{SE}_1)^2} + \frac{1}{(\text{SE}_2)^2} + \frac{1}{(\text{SE}_3)^2}}} \\
 &= \frac{1}{\sqrt{\frac{1}{(4.1)^2} + \frac{1}{(9.8)^2} + \frac{1}{(26.3)^2}}} \\
 &= 3.7\%
 \end{aligned}$$

Thus the combined estimate of the heritability with its standard error is $58.6 \pm 3.7\%$.

Table 5.3 Heritability of congenital dislocation of the hip (late diagnosis) from frequencies in various relatives (data from Wynne-Davies, 1970)

	<i>A</i>	<i>N</i>	<i>q</i> %	<i>x</i>	<i>a</i>	<i>r</i>	<i>V</i> × 1000	\sqrt{V}	$h^2 \pm \text{SE}$
Population	—	—	0.10	3.090	3.367	—	—	—	—
Relatives:									
first-degree	35	1777	1.97	2.060	2.426	0.306	0.420	0.0205	61.2 ± 4.1
second-degree	16	4746	0.34	2.706	3.012	0.114	0.606	0.0246	45.6 ± 9.8
third-degree	8	4220	0.19	2.894	3.185	0.058	1.085	0.0329	46.4 ± 26.3

Not all investigators agree with Falconer's model and consider that it is perhaps somewhat artificial to imagine a sharp cut-off beyond which individuals are affected. Instead Edwards (1969) and Curnow (1972) have suggested that a more realistic model is to consider that the genetic component of liability is normally distributed in both the general population and in affected individuals and according to Curnow (1972) the risk of being affected increases in a sigmoid manner from 0 at a low genetic level to 1 at a high genetic level. However, Falconer's model and this latter model are mathematically equivalent and lead to similar results.

Smith (1970) has produced a very useful graph (Fig. 5.2) from which it is possible to derive an approximate estimate of the correlation in liability between relatives, knowing the frequency of a disorder in the general population and in relatives of affected individuals. Knowing the correlation coefficient (*r*) it is possible to calculate the heritability since $h^2 = r/R$ where '*R*' is the coefficient of relationship (see p. 21). Thus in the case of renal calculi (an example chosen by Falconer, 1965) the frequency in relatives of controls is 0.4% and the frequency in first-degree relatives of patients is 2.5%. From Smith's graph $r = 0.25$ and therefore the heritability is 50%. From Figure 5.2 it is also possible to estimate the standard error of the heritability (Smith, 1970), but the method is complex and it is easier to calculate using Falconer's method (Falconer, 1965) as illustrated on page 59. Note that because of sampling error the frequency in relatives might, in a particular study, appear

to be less than the population frequency which would give rise to a negative estimate of heritability (Fig. 5.2). Such negative estimates should be included when pooling estimates from different sources.

Provided the population frequency of a disorder is known it is also possible from Smith's graph (Fig. 5.2) to derive an approximate estimate of the *upper limit* of recurrence risks for relatives by assuming h^2 is 100%. Thus if the frequency in the general population is 1.0%, the maximum frequency (recurrence risk) among first-degree relatives ($r = 0.5$) is 13%, among second-degree relatives ($r = 0.25$) is 4.5% and among third-degree relatives ($r = 0.125$) is about 2.2%. The values obtained by Falconer's method are very similar and from the practical point of view of genetic counselling the differences are small enough to be ignored.

So far we have only been concerned with the estimation of heritability of *discontinuous* characters. Prominence has been given to this subject because most disease states are regarded in this way. However mention should also be made of the estimation of heritability of *continuous* characters such as stature and blood pressure. Here heritability can be estimated from parent-offspring correlations or sib-sib correlations. In the former we take the average value of the offspring in each family and then calculate the correlation (r) between these values and the average value for both parents in each family (so-called *mid-parent* value). In general statistical terms, the correlation coefficient between two variables (x and y) is equal to the square root of the regression (b) of y on x multiplied by the regression of x on y :

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

The expected correlation between mid-parent and child is therefore equal to the square root of the regression of child on mid-parent (which is 1) times the regression of mid-parent on child (which is 0.5);

$$r = \sqrt{0.5} = 0.71$$

if the trait is completely genetically determined. Otherwise the heritability

$$h^2 = \frac{r}{0.71}$$

Alternatively we can calculate the correlation (r) between the average values of the offspring in each family and the value for *one* parent in each family (the mother-child and father-child correlations being treated separately), in which case

$$h^2 = 2r$$

In the case of sibs one calculates the so-called *intraclass correlation coefficient* by an analysis of variance. The reader is referred to one of the standard text books of statistics for details which are outside the scope of this book (for example, see Snedecor & Cochran, 1967, p. 294 et seq). The

derivation of heritability from the intraclass correlation for sibs is not straightforward and the problem is discussed by Falconer (Falconer, 1981). The details of assessing heritability in this way are usefully discussed in regard to stature by Roberts et al (1978).

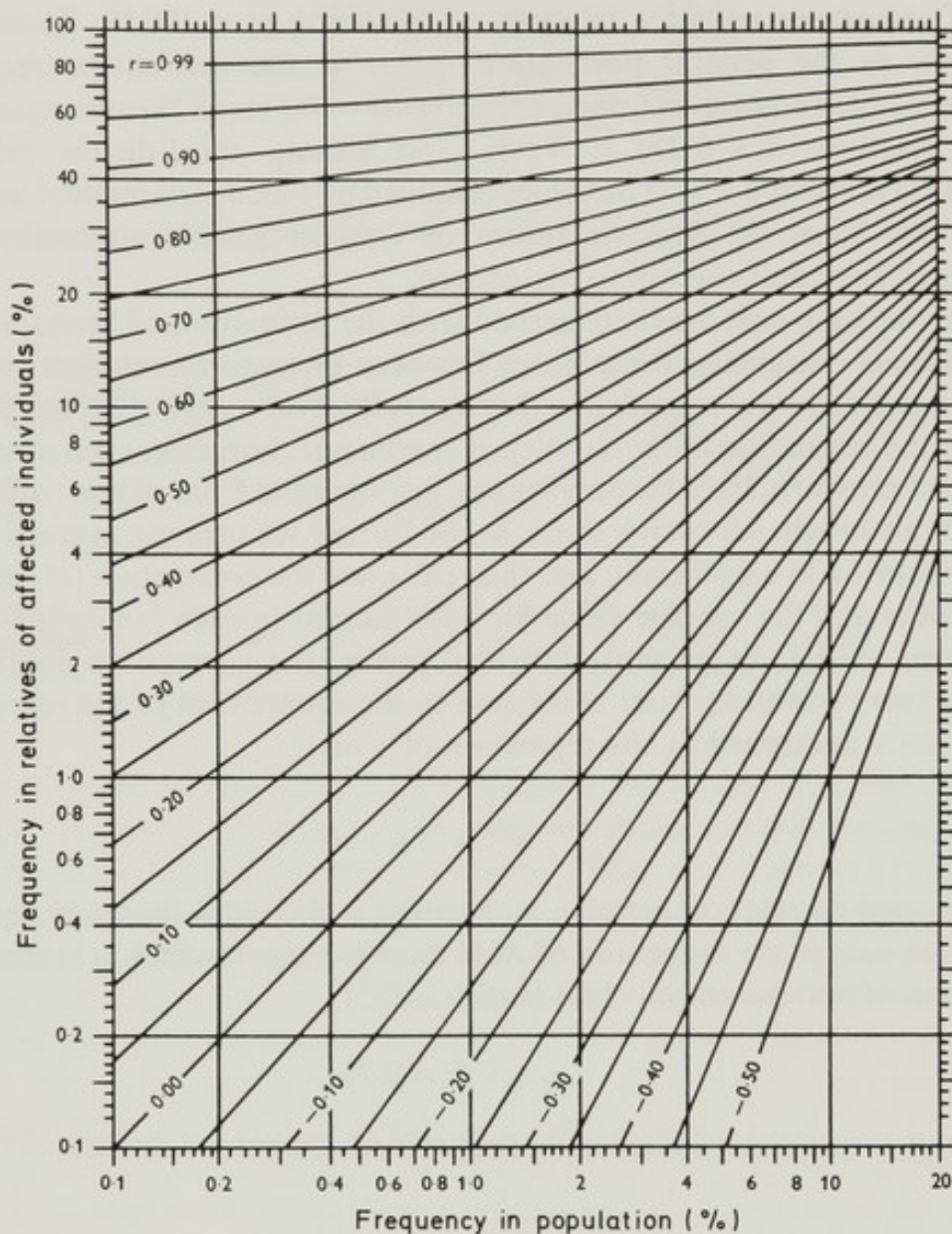


Fig. 5.2 Graph of correlations in liability (r) between relatives and probands. From this the heritability (h^2) can be derived since $h^2 = r$ for MZ twins, $h^2 = 2r$ for first-degree relatives, $h^2 = 4r$ for second-degree relatives and $h^2 = 8r$ for third-degree relatives. (From Smith, 1970.)

Estimation of heritability from twin studies

Twins are said to be *concordant* when both exhibit the same trait. If only one twin has the trait they are said to be discordant (see p. 87).

If in a particular disorder the population frequency and the proband concordance rate (C_p) in twins are known then it is possible to estimate the correlation (r) in liability and from this the heritability (Smith, 1974).

An estimate of 'r' is given by

$$r = \frac{x_g - x_{ra}}{a_g}$$

where values of 'x' and 'a' can be obtained from Appendix 5 in the usual manner. Thus if the frequency of a disorder in the population is 0.5% ($x_g = 2.576$ and $a_g = 2.892$) and if the proband concordance rate is 5.0% ($x_{ra} = 1.645$) then

$$\begin{aligned} r &= \frac{2.576 - 1.645}{2.892} \\ &= 0.32 \end{aligned}$$

or a more precise estimate may be obtained from

$$r = \frac{x_g - [x_{ra}\sqrt{1 - (x_g^2 - x_{ra}^2)(1 - x_g/a_g)}]}{a_g + [x_{ra}^2(a_g - x_g)]}$$

which in the above example is

$$= \frac{2.576 - [1.645\sqrt{1 - (2.576^2 - 1.645^2)(1 - 2.576/2.892)}]}{2.892 + [1.645^2(2.892 - 2.576)]} = 0.36$$

The standard error of the correlation

$$SE = \sqrt{\left(\frac{1}{a_g^2}\right)\left(\frac{1}{a_{ra}^2}\right)\left(\frac{1 - C_p}{A}\right)}$$

Where A in this case is the number of twin pairs in which both members are affected. Thus, if in the above example $A = 9$ then

$$\begin{aligned} SE &= \sqrt{\left(\frac{1}{2.892^2}\right)\left(\frac{1}{2.063^2}\right)\left(\frac{0.95}{9}\right)} \\ &= 0.05 \end{aligned}$$

Thus the estimate of the correlation and its standard error in the above example is

$$0.36 \pm 0.05$$

It is also possible from twin data to estimate the correlation and therefore the heritability simply from Figure 5.2, provided the concordance rate in twins and the frequency in the population are known. Thus in schizophrenia the population frequency is about 1.0% and in one recent study the proband concordance rates were approximately 58% in MZ twins and 12% in DZ twins (Gottesman & Shields, 1972). From Figure 5.2 if the concordance in MZ twins is 58% then r_{MZ} is 0.92, and if concordance in DZ twins is 12% then

r_{DZ} is 0.48. Therefore for MZ twins

$$h^2 = r_{MZ} = 92\%$$

and for DZ twins

$$h^2 = 2r_{DZ} = 96\%$$

Pooling the results of several such twin studies the average estimate of the heritability of schizophrenia works out to be about 85% (Gottesman & Shields, 1973).

It should be noted that in the absence of environmental similarities, concordance rates in MZ twins will not be expected to be high unless the heritability and population frequencies are high (Smith, 1970). Despite a high heritability the concordance may be low if the population frequency is low.

It should also be noted that the index (H) proposed by Holzinger (Holzinger, 1929), which depends on concordance rates in MZ and DZ twins, i.e. $H = (C_{MZ} - C_{DZ}) / (1 - C_{DZ})$, is an arbitrary index and has no specific genetic interpretation (p. 91). It is not a measure of heritability and therefore should not be used for this purpose (Smith, 1974).

The heritability of *continuous* characters may also be estimated from twin studies, in this case it is derived from the intraclass correlation coefficient and this is discussed later (p. 90).

In conclusion, an estimate of heritability of liability for a particular disorder is valuable for a number of reasons. Firstly, if a 'reasonable' estimate is obtained this tends to support the hypothesis that the disorder is inherited on a multifactorial basis. Secondly, it gives an idea of the relative contribution of genetic and environmental factors to aetiology. Thirdly, it can be useful in genetic counselling in helping to predict the possible frequency (and therefore the chances of recurrence) in relatives. However a word of caution is necessary. The application of the multifactorial model, with subsequent estimates of heritability, is only justified when certain specific criteria are met (p. 56). The uncritical application of the concept is therefore to be avoided, a point which has been well argued by Fraser (1976).

Genetic linkage

In recent years a variety of laboratory techniques have provided a great deal of information on gene localization in man (Francke, 1983). Pedigree analysis, however, will continue to be of value particularly for localizing genes for traits not expressed in cultured cells and for measuring distances between loci.

Much has been written on the subject of pedigree analysis for genetic linkage studies and detailed expositions are to be found in Edwards (1971), Renwick (1971) and Smith (1968). An eminently readable introduction to the subject is to be found in Race & Sanger's *Blood Groups in Man* (Race & Sanger, 1975).

The method adopted in determining linkage is the maximum likelihood estimate of the recombination fraction (usually referred to as θ) based upon the relative probability (P_R) of having obtained the family. The latter is determined by calculating the probability of having obtained the various combinations of the particular traits under consideration on the assumption of there being no *measurable* linkage ($\theta = 0.5$) and comparing this with the probabilities based on a range of recombination fractions from 0.00 to 0.50, i.e.

$$P_R = \frac{P(\text{family, given } \theta = 0 \text{ to } 0.5)}{P(\text{family, given } \theta = 0.5)}$$

For convenience P_R is often expressed as its logarithm. The \log_{10} of the relative probability is called the 'log of the odds' or the *lod score* (Morton, 1955). The maximum likelihood estimate of θ may be obtained by plotting the sum of the lod scores (or the relative probabilities) for all the families studied against various values of θ from 0.00 to 0.50, and is the value of θ corresponding to the peak of the curve. Gene loci located on the same chromosome are said to be *syntenic*, and syntenic loci showing less than 50% recombination between them ($\theta < 0.50$) are said to be *linked*.

Autosomal linkage

Three generation families

Linkage phase refers to whether two linked genes are on the same particular

chromosome (= *coupling phase*) or on different homologous chromosomes (= *repulsion phase*). In three generation families the linkage phase of individuals in the second generation may be obvious from inspection of the pedigree. In such a situation the recombination fraction can be determined quite simply. Thus in the family in Figure 6.1, if black represents myotonic dystrophy, a dominant disorder, and the sector alleles are represented as *Se* (secretor) which is dominant to *se* (non-secretor) for the presence of ABH substances in body secretions, we see that II₁ must be heterozygous for both loci, and the secretor and myotonic genes are in coupling in II₁. Therefore in the third generation III₂₋₆ are all non-recombinants but III₁ is a recombinant.

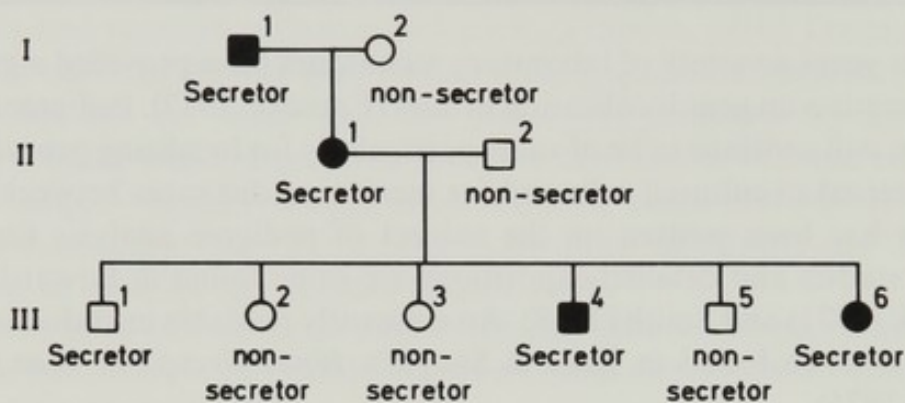


Fig. 6.1 Family in which myotonic dystrophy (dominant) and secretor status are segregating.

In this situation the recombination fraction (θ) is therefore 1 out of 6 or 0.17. In fact, the study of a large number of families in which secretor status and myotonic dystrophy were segregating gives a value of θ close to 0.07.

Two generation families

When information is available only in two generations of a family, the measurement of linkage is more involved and demands a resort to some mathematics. If '*G*' and '*g*' are alleles at the 'main' (disease) locus and '*T*' and '*t*' are alleles at the 'test' (genetic marker) locus, then if an individual has the genotype *GT/gt* (i.e. coupling phase) there are four possible types of gametes: two non-recombinants (*GT* and *gt*) and two recombinants (*Gt* and *gT*). If the frequency of recombination is θ then:

$$\begin{aligned} \text{frequency of non-recombinant gametes} &= 1 - \theta \\ \text{therefore frequency of } GT \text{ or } gt \text{ gametes} &= \frac{1 - \theta}{2} \\ \text{and frequency of recombinant gametes} &= \theta \\ \text{therefore frequency of } Gt \text{ or } gT \text{ gametes} &= \frac{\theta}{2} \end{aligned}$$

If these two loci were not on the same chromosome (not syntenic), or if on the same chromosome but were far apart ($\theta > 0.5$), then there would be equal numbers of all four types of gametes.

Now let us consider a family in which Lutheran blood groups and secretor status are segregating. The Lutheran alleles are Lu^a and Lu^b where Lu^a is dominant to Lu^b (Fig. 6.2).

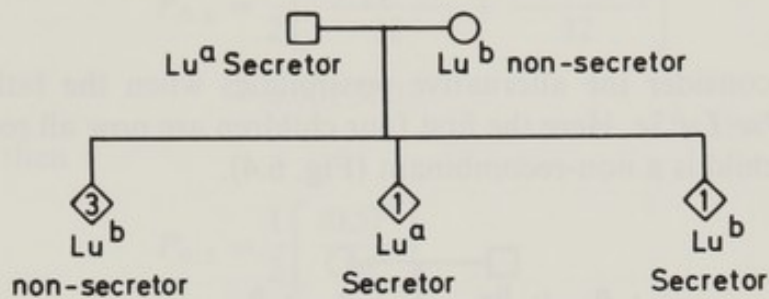


Fig. 6.2 Segregation of Lutheran blood groups and secretor status

From this pedigree the mother must have the genotype $Lu^b se/Lu^b se$ and father must be $Lu^a Se/Lu^b se$ or $Lu^a se/Lu^b Se$. Depending on which genotype the father has affects the assessment of his offspring as to which are recombinants and which are not. Let us first consider that the arrangement is $Lu^a Se/Lu^b se$, in which case the first four children are all non-recombinants and the last child is a recombinant (Fig. 6.3).

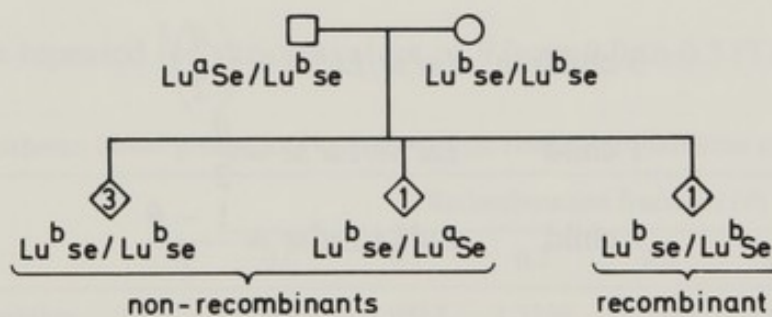


Fig. 6.3 Segregation of Lutheran blood groups and secretor status

In this case the probability of getting

$$3 \text{ children } Lu^b se/Lu^b se = \left(\frac{1-\theta}{2}\right)^3$$

$$1 \text{ child } Lu^b se/Lu^a Se = \frac{1-\theta}{2}$$

$$1 \text{ child } Lu^b se/Lu^b Se = \frac{\theta}{2}$$

Therefore the probability of getting this family if father has the genotype Lu^aSe/Lu^bse is

$$\begin{aligned} & \left(\frac{1-\theta}{2}\right)^3 \left(\frac{1-\theta}{2}\right) \left(\frac{\theta}{2}\right) \\ & = \frac{\theta(1-\theta)^4}{32} \end{aligned}$$

Now let us consider the alternative possibilities when the father has the genotype Lu^ase/Lu^bSe . Here the first four children are now all recombinants and the last child is a non-recombinant (Fig. 6.4).

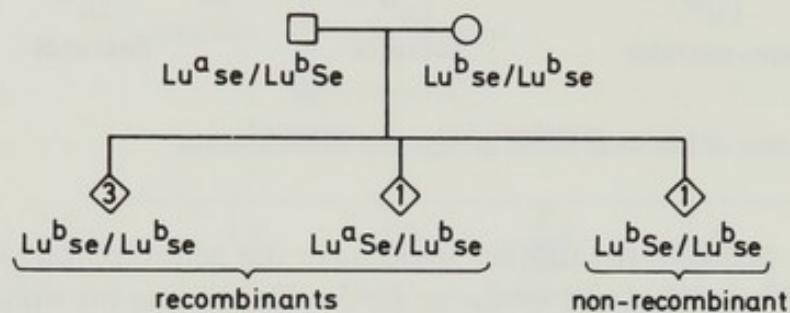


Fig. 6.4 Segregation of Lutheran blood groups and secretor status.

In this case the probability of getting

$$\begin{aligned} 3 \text{ children } & Lu^bse/Lu^bse = \left(\frac{\theta}{2}\right)^3 \\ 1 \text{ child } & Lu^aSe/Lu^bse = \frac{\theta}{2} \\ 1 \text{ child } & Lu^bSe/Lu^bse = \frac{1-\theta}{2} \end{aligned}$$

Therefore the probability of getting this family if father has the genotype Lu^ase/Lu^bSe is

$$\begin{aligned} & \left(\frac{\theta}{2}\right)^3 \left(\frac{\theta}{2}\right) \left(\frac{1-\theta}{2}\right) \\ & = \frac{\theta^4(1-\theta)}{32} \end{aligned}$$

Now what we have to decide is the probability of obtaining the observed phenotypes of the children under two different assumptions namely that father is either in coupling or in repulsion. Since coupling and repulsion are equally likely the probability of getting this family is the *average* of the

probabilities assuming coupling and repulsion, i.e.

$$\frac{1}{2} \left[\frac{\theta(1-\theta)^4}{32} + \frac{\theta^4(1-\theta)}{32} \right]$$

If we assume $\theta = 0.2$ then

$$\begin{aligned} P_{0.2} &= \frac{1}{2} \left[\frac{(0.2)(0.8)^4}{32} + \frac{(0.2)^4(0.8)}{32} \right] \\ &= 0.001\ 300 \end{aligned}$$

If $\theta = 0.5$ then

$$\begin{aligned} P_{0.5} &= \frac{1}{2} \left[\frac{(0.5)^5}{16} \right] \\ &= 0.000\ 976 \end{aligned}$$

Therefore the relative probability (P_R) when $\theta = 0.2$

$$\begin{aligned} &= \frac{P(\text{family}|\theta = 0.2)}{P(\text{family}|\theta = 0.5)}^* \\ &= \frac{0.001\ 300}{0.000\ 976} \\ &= 1.3320 \end{aligned}$$

This is then repeated for various values of θ from 0.0 to 0.5 (Table 6.1).

Table 6.1 Lutheran blood groups and secretor status: relative probabilities and lod scores

	Recombination fractions (θ)					
	0.0	0.1	0.2	0.3	0.4	0.5
Relative probability	0.0	1.0517	1.3320	1.2439	1.0758	1.0000
Lod score	$-\infty$	0.022	0.124	0.095	0.031	0.00

The relative probabilities are then plotted against the various recombination fractions and in this way the maximum likelihood estimate of the recombination fraction is obtained (Fig. 6.5). In this example the maximum likelihood estimate of θ is approximately 0.21. In fact the study of a number of families in which the Lutheran blood groups and secretor status were segregating indicates that θ is about 0.15.

* ' $P(\text{family}|\theta = 0.2)$ ' means the probability of the family *given that* $\theta = 0.2$. The vertical line means 'given that'.

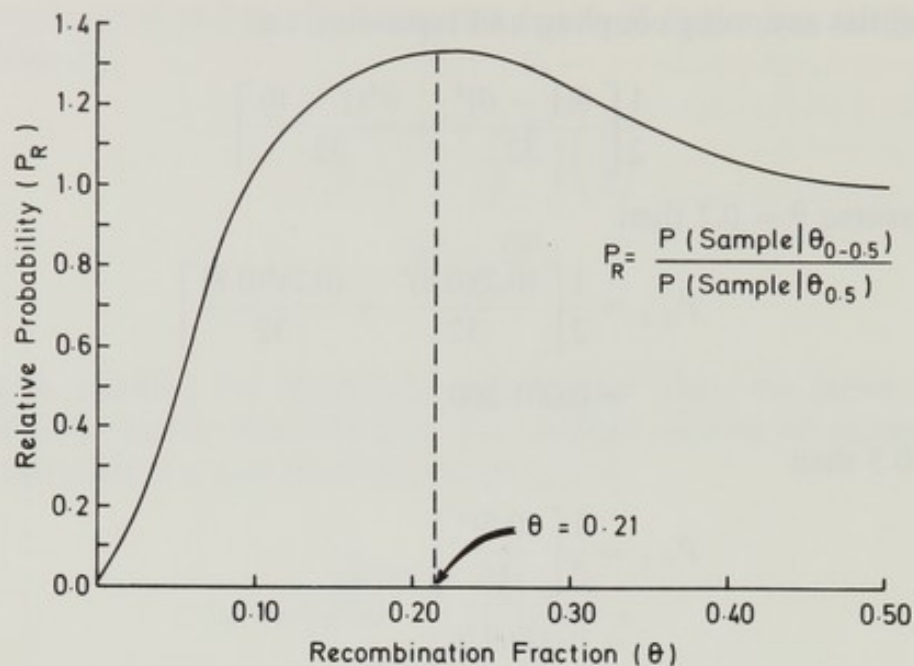


Fig. 6.5 Relative probability of linkage for various recombination fractions

Fortunately it is not necessary to go through such laborious calculations each time because tables of lod scores are now available (see Appendix 6).

When the parent's 'phase' is known (as it may be in three-generation families) then lod scores for various numbers of non-recombinant and recombinant offspring in each family can be read off directly from the table. When the parents' phase is not known (as in two generation families) the z_1 score and its correction (e_1) have to be determined. The z_1 score is determined in the following manner. The offspring are divided up according to whether or not they possess the main character and/or the test character. In the example on page 68 this would have been necessary if there had been no information on individuals in generation I. Thus:

Main character (G) + test character (T)	= 2	}	5
(myotonic dystrophy + secretor)			
Not main character (g) + not test character (t)	= 3		
(healthy + non-secretor)			
Main character (G) + not test character (t)	= 0	}	1
(myotonic dystrophy + non-secretor)			
Not main character (g) + test character (T)	= 1		
(healthy + secretor)			

Thus the z_1 score is indicated as 5:1 (the larger number is conventionally written first) and is looked up in tables under the heading ' z_1 5:1'. The e_1 correction is necessary only when the test character genotype of a parent can only be derived from an offspring involved in the count for z_1 . It is equal to the number of individuals with or without the main character. In the above

example this would be 4 : 2 (the larger number is again written first). The z_1 and e_1 scores, indicated as ' z_1 5:1' and ' e_1 4:2', are then obtained from Appendix 6 and would be added to the lod scores from other families.

The anti log of the sum of the lod scores for individual families for various values of θ gives the relative probability of linkage for each value of θ .

In extensive pedigrees of more than three generations the same logical approach is followed, the z_1 and e_1 scores are calculated and from these the lod scores are obtained.

The work involved however can be extremely tedious and time consuming and there is plenty of room for mistakes in logic particularly in extensive pedigrees. Fortunately computer programs, such as LIPED (Ott, 1974) and LINKAGE (Lathrop et al., 1985), are now available for such computations.

Prior probabilities of linkage

So far, we have considered that all values of the recombination fraction are equally likely. This is an oversimplification and Renwick (1969, 1971) has emphasized the need to take into account the initial or 'prior' probabilities of different values of θ .

Since chromosomes vary in length the probability that a given gene is located on a specific chromosome is proportional to the length of the chromosome. By considering the relative lengths of all 22 autosomes it has been calculated that the prior odds of linkage for any two genes, i.e. that two genes are located on the same autosome, is 1 : 17.5 (Renwick, 1969). Other prior odds to be considered include the differential rate of recombination in males and females (p. 75), known chromosomal location of one of the loci being studied and so on. To obtain the final odds for linkage the prior odds for each value of θ are taken into account. The calculation of prior probabilities is rather complicated and the reader is referred to the original papers by Renwick (1969, 1971). Fortunately, in practice, the inclusion of prior probabilities is not necessary if only a rough estimate of the recombination fraction is required and since human pedigree data are usually meagre, this is often all that is possible anyway.

Probability of linkage

The average height (H) of the relative probability curve indicates the odds on linkage, which for autosomal linkage are approximately equal to $H : 20$. The average height of the relative probability curve is equal to the sum of the antilogs of the lod scores for $\theta = 0.05, 0.10, 0.15 \dots 0.45$ (Appendix 6) divided by 9. Thus if H is 100 then the odds on linkage would be 100 : 20 or 5 : 1. For X-linked loci the probability of linkage is equal to $H : 1$.

Probability limits

The 0.95 *probability* limits (they are not really *confidence* limits in the true

sense of the term) of the maximum likelihood estimate of the recombination fraction may be determined approximately by simply subtracting 2.5 % of the total area under the relative probability curve from each end of the curve. The total area of the curve may be determined by planimetry or simply by counting the number of one centimetre squares covered by the curve.

Recombination fraction and map distance

The relative distance between different loci on any particular chromosome is related to the frequency with which crossing-over occurs between them : 1 % crossing-over (i.e. $\theta = 0.01$) being equal to one map unit or centiMorgan (cM). The relationship between the recombination fraction and actual map distance however, is not linear. The farther apart the loci, the greater the discrepancy since double cross-overs will occur and be scored as non-recombinants. Several formulae have been derived (e.g. Haldane, 1919; Kosambi, 1944; Carter & Falconer, 1951) which permit recombination fractions to be converted into map units. The formula of Kosambi (1944) is convenient.

$$D = 25 \log_e \left(\frac{1 + 2\theta}{1 - 2\theta} \right)$$

where D = map distance in centiMorgans
 θ = recombination fraction.

If natural logarithms (\log_e) are not available then since $\log_e = 2.3026 \log_{10}$, therefore

$$D = 57.57 \log_{10} \left(\frac{1 + 2\theta}{1 - 2\theta} \right)$$

Because of the limitations to the amount of human data that are usually available, it is unlikely that linkage will be detected if θ is much greater than 0.20. Since the relationship between map distance and θ is essentially linear up to $\theta = 0.20$, from a practical point of view, recombination frequencies can be converted directly into map distances with little loss of precision. Thus Race & Sanger (1975, p. 595) calculated map distances for various values of the recombination fraction using the three different formulae and in each case the correction made little difference up to 0.20:

θ	Haldane	Kosambi	Carter & Falconer
0.10	11	10	10
0.20	26	21	20
0.30	46	34	31

A table for converting θ to map distances is given in Rao et al (1977).

For reasons which are not clear recombination is often greater in females than in males. Therefore in studying genetic linkage, if possible, families should be divided into those in which the mother is the doubly heterozygous parent and those in which the father is the doubly heterozygous parent and the recombination frequencies determined for males and females separately.

X-linkage

In assessing the linkage relationships between a main locus and a test locus, the offspring of doubly heterozygous females are scored. It has to be remembered, however, that if no one else in the family is affected a mother can only be considered heterozygous at the main locus (disease locus) if she has had at least two affected sons. If there is only one affected son this could be the result of a new mutation, in which case the mother would not be a carrier. Of course if in a particular disorder there is a reliable test for the heterozygous

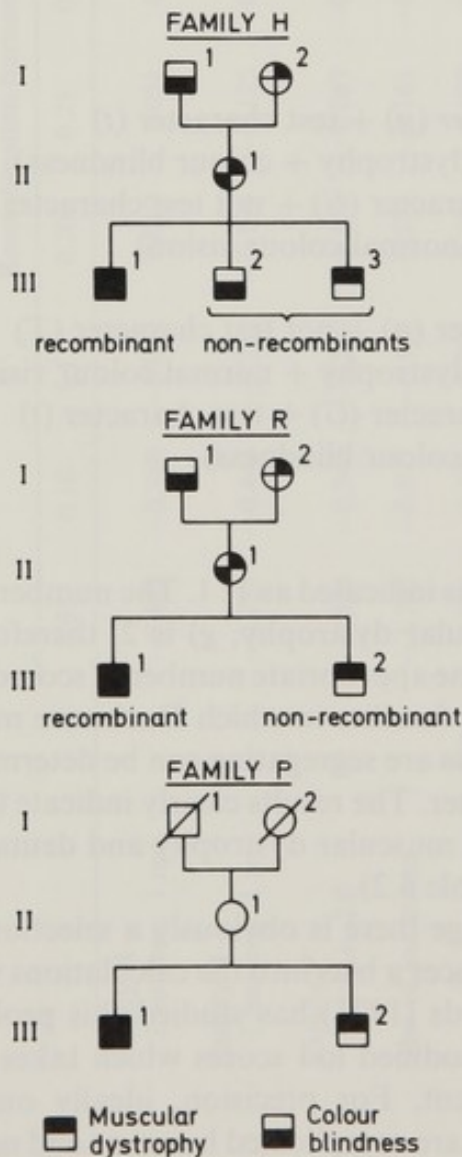


Fig. 6.6 Families in which Duchenne muscular dystrophy and deutan colour blindness are segregating. (From Emery, 1966.)

state then even if the mother has only one affected son her genotype at the main locus can be firmly established as either GG or Gg .

Consider three families in which Duchenne muscular dystrophy (main character) and deutan colour blindness (test character) were segregating (Emery, 1966). In family H (Fig. 6.6), since the maternal grandfather was colour blind and his wife a carrier of Duchenne muscular dystrophy (she also had other male relatives with this disease), these two loci must be in repulsion in their daughter II_1 . Therefore III_1 must be a recombinant and III_2 and III_3 must be non-recombinants. Similarly in family R the mother (II_1) is doubly heterozygous and the colour blind and muscular dystrophy loci are in repulsion and therefore III_1 must be a recombinant and III_2 must be a non-recombinant.

In family P the grandfather (I_1) was dead. It was not known whether he was colour blind and therefore the linkage phase in the mother (II_1) is not known. In this case lod scores can be determined in the usual way. To determine the z_1 score (see p. 72):

$$\left. \begin{array}{l} \text{Main character } (g) + \text{test character } (t) \quad = 1 \\ \quad \text{(muscular dystrophy + colour blindness)} \\ \text{Not main character } (G) + \text{not test character } (T) = 0 \\ \quad \text{(healthy + normal colour vision)} \end{array} \right\} 1$$

$$\left. \begin{array}{l} \text{Main character } (g) + \text{not test character } (T) \quad = 1 \\ \quad \text{(muscular dystrophy + normal colour vision)} \\ \text{Not main character } (G) + \text{test character } (t) \quad = 0 \\ \quad \text{(healthy + colour blindness)} \end{array} \right\} 1$$

Therefore the z_1 score is indicated as 1 : 1. The number of individuals with the main character (muscular dystrophy, g) is 2, therefore the e_1 correction is indicated as 2 : 0. For the appropriate number of scored children the lod scores for each of these three families in which Duchenne muscular dystrophy and deutan colour blindness are segregating can be determined (Appendix 6) and then added to each other. The results clearly indicate (Emery et al, 1969) that the loci for Duchenne muscular dystrophy and deutan colour blindness are not closely linked (Table 6.2).

In studying X-linkage there is obviously a selection of such 'informative' families, which introduces a bias into the calculations which may alter the lod scores slightly. Edwards (1971) has studied this problem in detail and has provided a table of modified lod scores which takes into account different modes of ascertainment. For precision, ideally one should perhaps use Edwards' tables. They are complicated however and most will find that for all practical purposes the lod scores in Appendix 6 are quite adequate and in fact give very similar results.

Table 6.2 Duchenne-type muscular dystrophy and deutan colour blindness: lod scores and antilogs (relative probabilities).

	Recombination fraction (θ)									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	
Family H										
2 non-rec.: 1 rec.	-0.442	-0.188	-0.062	0.010	0.051	0.070	0.073	0.061	0.037	
Family R										
1 non-rec.: 1 rec.	-0.721	-0.444	-0.292	-0.194	-0.125	-0.076	-0.041	-0.018	-0.004	
Family P										
z_1 1:1; e_1 2:0	-0.584	-0.340	-0.215	-0.138	-0.087	-0.052	-0.028	-0.012	-0.003	
Sum of lod scores	-1.747	-0.972	-0.569	-0.322	-0.161	-0.058	0.004	0.031	0.030	
Antilog = relative probability	0.018	0.106	0.269	0.476	0.690	0.875	1.009	1.074	1.072	

Finally it should be appreciated that if the results of pedigree analysis suggest that two loci (whether autosomal or X-linked) are within measurable distance of each other, one will be more confident of the linkage relationship the greater the lod scores and the narrower the 0.95 probability limits. This point is illustrated diagrammatically in Figure 6.7.

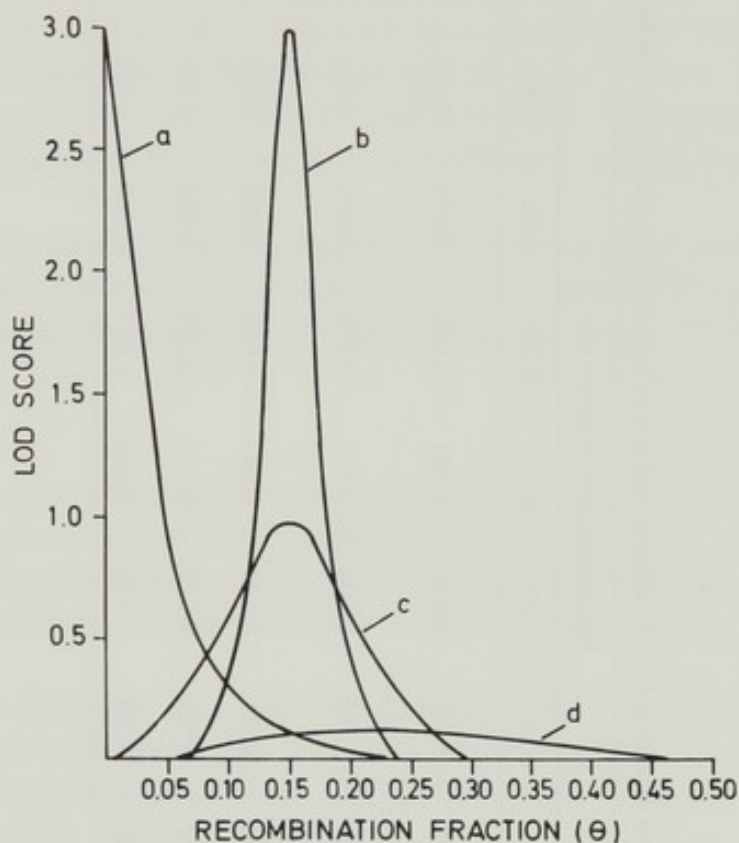


Fig. 6.7 Diagram of lod scores plotted for various values of the recombination fraction (θ) indicating: (a) high probability of very close linkage (θ less than 0.05); (b) high probability that θ equals 0.15; (c) suggestion of linkage; (d) no discernible linkage.

The precise locations of several hundred genes in the human genome is now known. This has been established by a variety of techniques quite apart from classical linkage analysis (McKusick, 1980; Francke, 1983) and summaries are to be found for example in the *American Journal of Human Genetics* (1983) **35**: 134–156, *Human Gene Mapping 7*, *Cytogenetics and Cell Genetics* (1984) **37** Nos 1–4. The use of linkage specifically with DNA markers for genetic counselling purposes will be discussed in Chapter 8.

Twin studies, their use and limitations

Here we shall not be concerned so much with twinning as a biological phenomenon, which has been dealt with in detail by Bulmer (1970) and MacGillivray et al (1975), but rather with the value of twin studies in genetic analysis. There are definitely two, and possibly three, types of twins. Firstly, there are dizygous (DZ), non-identical or fraternal twins derived from the independent release and subsequent fertilization of two separate ova. Such twins are genetically no more alike than sibs. Secondly, there are monozygous (MZ), or identical twins derived from the splitting of a single fertilized ovum at an early stage in development. They therefore have the same genotype and are of the same sex. A third type of twin resulting from the fertilization by two separate sperms of two division products of the *same* oocyte is theoretically possible but evidence of the existence in man is inconclusive. In man this form of fertilization usually results in a mosaic individual.

The frequency of MZ twinning is essentially the same throughout the world at about 3 to 4 per 1000 births. But whereas the frequency of DZ twins in Western Europeans is about 6 to 9 per 1000 it is about twice this in Negroes and less than half this in Orientals. Also MZ twinning is little influenced by

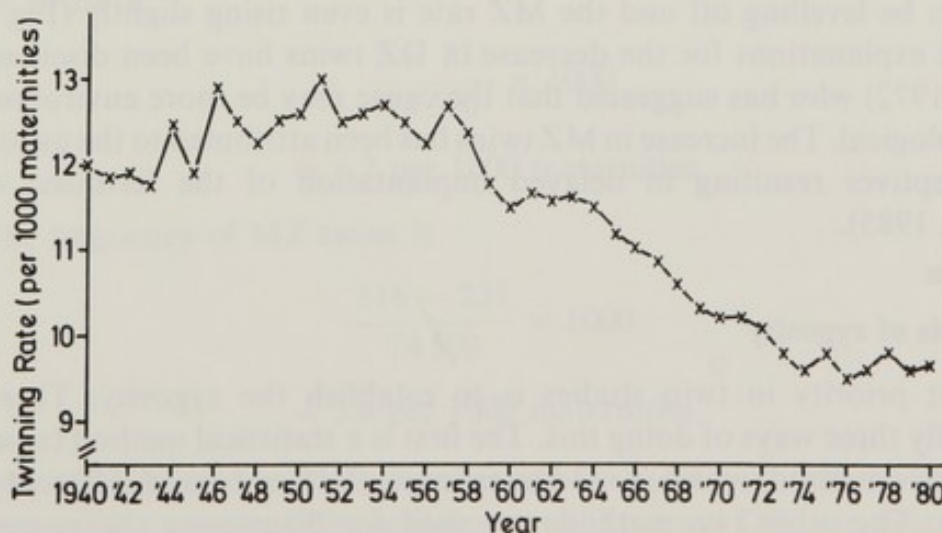


Fig 7.1 Frequency of twin births per 1000 maternities for England and Wales (source: *OPCS Birth Statistics*).

maternal age or parity but the frequency of DZ twins increases significantly with parity and with maternal age, reaching a maximum at 35 to 40 years (Bulmer, 1959). Hereditary factors are important in DZ twinning (mother's genotype being much more important than father's genotype) but much less so in MZ twinning (Parisi et al, 1983; Philippe, 1985). There has been a gradual decline in the frequency of twin births since the early 1950s (Fig. 7.1).

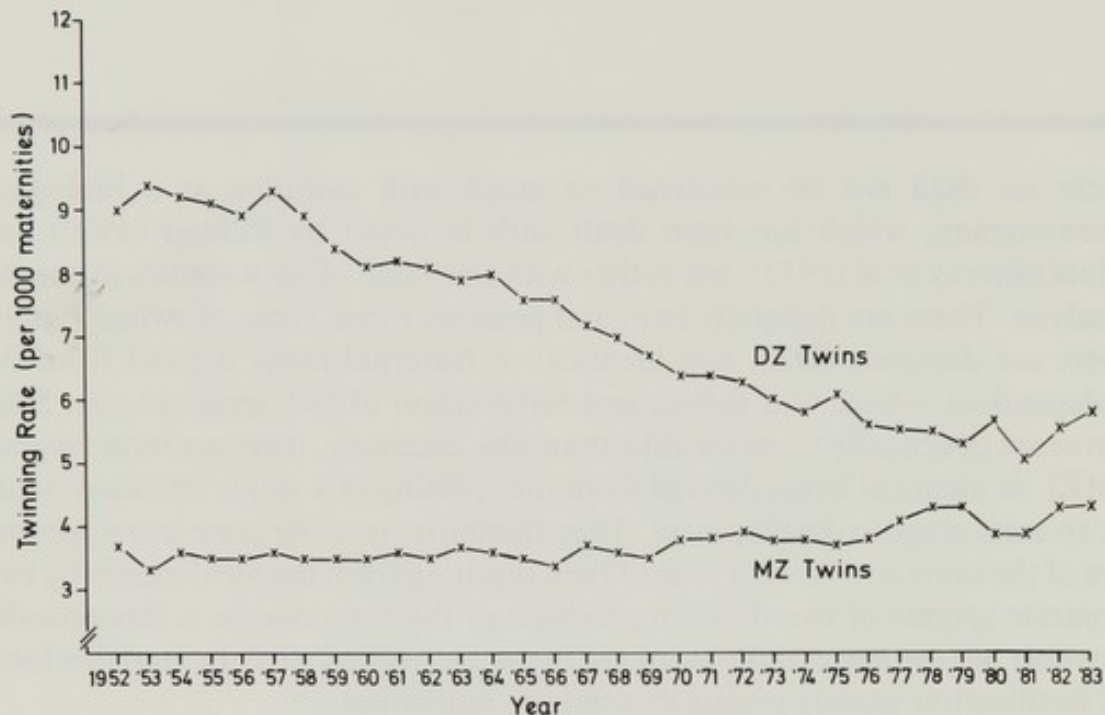


Fig. 7.2 DZ and MZ twinning rates (live and still births per 1000 maternities) for England, Scotland and Wales.

This decline has been due to a fall in the frequency of DZ twins, but this now seems to be levelling off and the MZ rate is even rising slightly (Fig. 7.2). Possible explanations for the decrease in DZ twins have been discussed by James (1972) who has suggested that the cause may be more environmental than biological. The increase in MZ twins has been attributed to the use of oral contraceptives resulting in delayed implantation of the fertilized ovum (Emery, 1985).

Diagnosis of zygosity

The first priority in twin studies is to establish the zygosity. There are essentially three ways of doing this. The first is a statistical method (so-called Weinberg's method) which is used to estimate the numbers of different types of twins. The other two methods are used for diagnosing the zygosity in individual twin pairs: one depends on the fetal membranes and the other on similarities between the twins.

Weinberg's method (Weinberg, 1901)

Since all MZ twins are of like-sex but half of DZ twins will be of unlike sex, then the number of DZ twins can be estimated by doubling the number of unlike-sex twins, and the number of MZ twins is the difference between the numbers of like-sexed and unlike-sexed twins.

Thus the proportion of DZ twins is

$$\frac{2U}{N}$$

where U = number of unlike-sexed twins
 N = total number of maternities

and the proportion of MZ twins is

$$\frac{L - U}{N}$$

where L = number of like-sexed twins.

Therefore per 1000 maternities the DZ twinning rate is

$$\frac{2U}{N} \times 1000$$

and the MZ twinning rate is

$$\frac{L - U}{N} \times 1000$$

Thus in 1973 in Scotland there were 74 500 maternities of which 747 resulted in twin births: 516 of like sex and 231 of unlike sex. Therefore the frequency of DZ twins is

$$\begin{aligned} & \frac{2 \times 231}{74\,500} \times 1000 \\ & = 6.2 \text{ per } 1000 \text{ maternities} \end{aligned}$$

and the frequency of MZ twins is

$$\begin{aligned} & \frac{516 - 231}{74\,500} \times 1000 \\ & = 3.8 \text{ per } 1000 \text{ maternities.} \end{aligned}$$

Weinberg's method is sufficiently accurate for most purposes though in fact there is a slight excess of like-sexed over unlike-sexed DZ twin pairs (James, 1976).

The frequency of multiple births very roughly follows Hellin's law. That is

if the frequency of twins is t , the frequency of triplets is t^2 , the frequency of quadruplets t^3 , etc. However because multiple births may result from treatment with recently introduced 'fertility drugs' this simple relationship may now no longer hold.

Fetal membranes

Examination of the fetal membranes is the time-honoured method of diagnosing zygosity at birth. There are a number of possibilities which are summarized diagrammatically in Figure 7.3. In all cases where there is a single chorion (monochorionic) the twins are unequivocally MZ since this occurs in about 70% of MZ twins but never in DZ twins. In other situations the diagnosis of zygosity is not clear cut and though dichorionic twins are more likely to be DZ, an individual dichorionic twin pair of like-sex cannot be unequivocally diagnosed by fetal membranes alone. For this reason and because information on fetal membranes may not always be available the so-called similarity method of diagnosing zygosity is often resorted to.

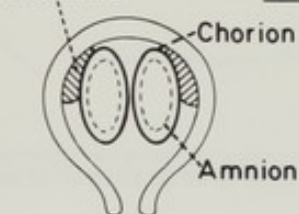



	PLACENTA	CHORION	AMNION	DZ (%)	MZ (%)
	2	2	2	50	15
	1	2	2	50	15
	1	1	2	—	70
	1	1	1	—	rare

Fig. 7.3 Diagrammatic representation of the different types of placentation and their frequencies in DZ and MZ twins

Similarity method

The object of this method is to compare, in the twin pair under study, traits in which MZ twins would be expected to resemble each other more closely than DZ twins. By such studies it is possible to estimate the relative probability of monozygosity to dizygosity. If a pair of twins differs in one simply inherited trait, such as sex, eye colour, or blood group, then the twins must be dizygotic. On the other hand, apart from skin grafting, a pair of twins can never be proved with certainty to be monozygotic, though with a large number of traits the probability that a twin pair is monozygotic may become almost unity. It should be remembered, however, that very rarely MZ twins may have a different phenotype or chromosome constitution as a result of post-zygotic aberrations (Nielsen, 1967; Benirschke & Kim, 1973).

Techniques used in this method include blood group typing, determination of certain polymorphic phenotypes demonstrable in serum, erythrocytes and/or leucocytes (including DNA polymorphic markers (p. 103), PTC tasting, secretor status and dermatoglyphics. The determination of zygosity using this method is much simpler when the parental phenotypes are known. The following example will illustrate the method of calculation. The findings in the father and mother of twin boys J and A were as follows:

	Father	Mother	J	A
<i>Blood groups</i>	<i>O</i>	<i>A₁B</i>	<i>A₁</i>	<i>A₁</i>
	<i>R₁r</i>	<i>R₁R₁</i>	<i>R₁R₁</i>	<i>R₁R₁</i>
	<i>MsMs</i>	<i>NsNs</i>	<i>MsNs</i>	<i>MsNs</i>
	<i>P-pos</i>	<i>P-pos</i>	<i>P-pos</i>	<i>P-pos</i>
	<i>Lu(a-)</i>	<i>Lu(a-)</i>	<i>Lu(a-)</i>	<i>Lu(a-)</i>
	<i>Kell-neg</i>	<i>Kell-neg</i>	<i>Kell-neg</i>	<i>Kell-neg</i>
	<i>Fy(a+)</i>	<i>Fy(a+)</i>	<i>Fy(a+)</i>	<i>Fy(a+)</i>
<i>Secretor status</i>	secretor	secretor	secretor	secretor
<i>Haptoglobin type</i>	1-1	2-1	2-1	2-1
<i>Dermatoglyphics</i>				
Total ridge count	—	—	161	165
Sum of atd angles	—	—	86°	88°

Clearly the only informative data are the ABO and rhesus (Rh) blood groups, haptoglobin types and dermatoglyphic findings. Since approximately 70% of all twins are DZ, the *prior* probability that twins are DZ is 0.70 and the probability of their being MZ is 0.30. In the above example, the *conditional* (see p. 93) probabilities of the second twin having the same sex, ABO and Rh

blood groups and haptoglobin type as the first twin if the twins are DZ is 0.50 in each case. From tables of dermatoglyphic findings in various types of twins (Smith & Penrose, 1955; Smith et al, 1961), a difference in total ridge count of only four occurs in about 4 % of like-sexed DZ twins and 27 % of MZ twins, and a difference in atd angles of only 2° occurs in about 9 % of DZ twins and in 18 % of MZ twins.

The prior, and each of the individual conditional probabilities are multiplied together to give a *joint* probability of either dizygosity (JP_{DZ}) or monozygosity (JP_{MZ}). The probability of dizygosity then

$$= \frac{JP_{DZ}}{JP_{DZ} + JP_{MZ}}$$

and the probability of monozygosity

$$= \frac{JP_{MZ}}{JP_{DZ} + JP_{MZ}}$$

or one minus the probability of dizygosity. In the above example the calculations are as follows:

Character	P_{DZ}	P_{MZ}
Prior probabilities	0.70	0.30
Conditional probabilities		
Sex	0.50	1.00
Blood groups		
ABO	0.50	1.00
Rh	0.50	1.00
Haptoglobin type	0.50	1.00
Dermatoglyphics		
diff. in TRC (4)	0.04	0.27
diff. in atd angles (2°)	0.09	0.18
Joint probability	0.000 157 5	0.01458

The probability of dizygosity is therefore

$$\frac{0.000\ 157\ 5}{0.014\ 737\ 5} \\ = 0.0107$$

and the probability of monozygosity is

$$1 - 0.0107 \\ = 0.9893$$

In this case there is a 99% probability that the twins are monozygous. It should be noted that with continuously variable traits such as dermatoglyphics stature or cephalic index, a diagnosis of the type of zygosity can never be made with certainty.

Now if we had had no information on the parents of these two twins then the probabilities of obtaining the various observed traits in the twins would have to have been based on gene frequencies in the general population. The calculations are very tedious but fortunately tables of relative probabilities of dizygosity are available (Smith & Penrose, 1955; Smith et al, 1961), for blood groups and some other traits based on their population frequencies in the United Kingdom. With such information the method of calculation is as follows. If p_0D , p_1D , p_2D , etc. are the relative probabilities of dizygosity for each trait under consideration, then the overall relative probability of dizygosity (pD) based on this information is

$$= p_0D \times p_1D \times p_2D \times \dots$$

and the total probability of the twins being dizygous is

$$= pD/(1 + pD)$$

and the probability of the twins being monozygous is

$$= 1 - [pD/(1 + pD)]$$

Thus in the above example:

Character	Relative probability of dizygosity
Prior prob.	2.3333 (i.e. 0.7/0.3)
Like-sex	0.5000
Blood groups	
<i>ABO</i> (A_1)	0.6470
<i>Rh</i> (R_1R_1)	0.5021
<i>MNS</i> ($MNss$)	0.4733
<i>P</i> ($P+$)	0.8489
<i>Lu</i> ($A-$)	0.9614
<i>K</i> ($K-$)	0.9485
<i>Fy</i> ($a+$)	0.8036
Secretor status	
secretor	0.8681
Dermatoglyphics	
Total ridge count	0.23
atd angles	0.50
Relative probability of dizygosity (pD)	0.01114

The probability of dizygosity is therefore

$$\frac{0.01114}{1.01114}$$

$$= 0.0110$$

and the probability of monozygosity is

$$1 - 0.0110$$

$$= 0.9890$$

Gaines & Elston (1969) have produced a series of curves (Fig 7.4) from which it is possible to read off directly the relative probability of dizygosity for any gene frequency (q) and which are therefore applicable to any simply inherited trait in any population in which the gene frequency is known. Curve *A* is used when the twins are homozygous as when one of the two alleles is dominant and the twins have the recessive phenotype, or if the two alleles are codominant and the twins have the phenotype of either of the two homozygous genotypes. Here ' q ' is the frequency of the common allele. Curve *B* is used when the twins are heterozygous as when the two alleles are codominant and the twins have the heterozygous phenotype. Here ' q ' is the frequency of *either* allele. Curve *C* is used when the twins could be either homozygous or heterozygous as when one allele is dominant and the twins exhibit the phenotype associated with the dominant allele. Here ' q ' is the frequency of the *dominant* allele. Thus in the above example, the $P +$ phenotype represents the homozygote PP or the heterozygote Pp , the gene frequency of P being about 0.50 in Western Europe. Therefore from curve *C* the relative probability of dizygosity is about 0.86.

Advances in recombinant DNA technology (Emery, 1984) have revealed that DNA polymorphisms, which occur once in every 100 to 200 base pairs, are extremely common in the general population. Such polymorphisms should prove especially valuable in determining twin zygosity (as well as paternity), because if a sufficient number are studied in any individual twin pair, they may well provide enough information in themselves without recourse to any other markers.

The use of twins in genetic analysis

The main value of twin studies in genetic analysis is that they can give an idea of the role of genetic factors in aetiology. From this point of view the most commonly employed techniques are the study of *concordance rates* (for discontinuous characters such as disease states) and *variances* and *correlations* (for continuous characters such as serum lipoproteins).

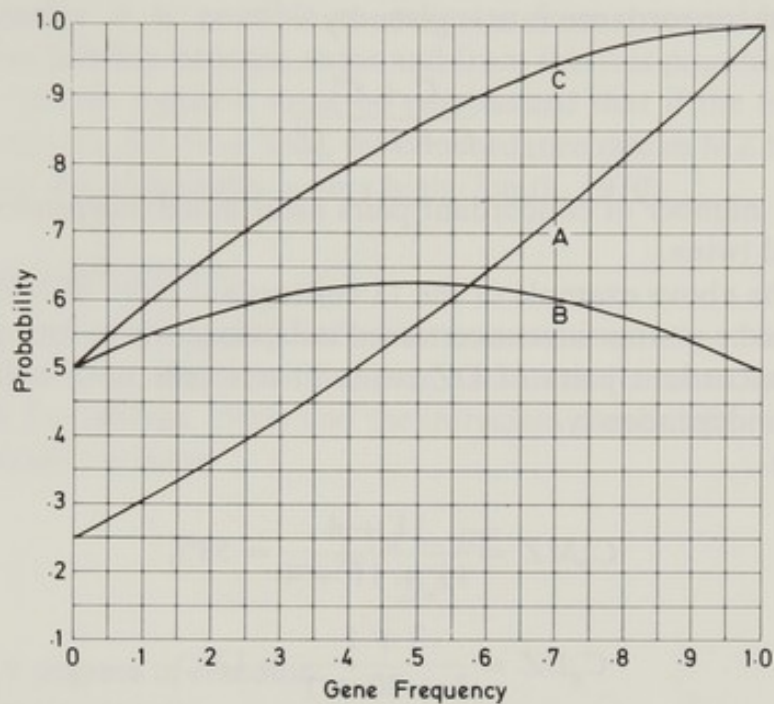


Fig. 7.4 Relative probability of dizygosity as a function of gene frequency (q). For details see text. (From Gaines & Elston, 1969)

Concordance rates

If both twins of a pair are affected they are said to be concordant, while if only one twin is affected they are said to be discordant. Concordance rates can be defined in a number of ways (Allen et al, 1967). Firstly there is the *pairwise concordance rate* (C_w) which may be defined as the proportion of affected twin pairs in which *both* members are affected. The pairwise concordance rate is usually given by

$$\frac{C}{C + D}$$

where

C = total number of concordant pairs

D = number of discordant pairs.

Thus in one recent study of schizophrenia in twins (Gottesman & Shields, 1972) there were 11 concordant and 11 discordant pairs of MZ twins and 3 concordant and 30 discordant pairs of DZ twins.

Therefore

$$C_w MZ = 11/22 = 50\%$$

$$C_w DZ = 3/33 = 9\%$$

Concordance may also be expressed as the *proband concordance rate* (C_p) which may be defined as the proportion of affected individuals among the co-twins of previously ascertained index cases. When both twins are affected and have been independently ascertained, the twin pair is in effect counted twice.

The proband concordance rate is given by

$$\frac{C + C'}{C + D + C'}$$

where C' = number of concordant pairs ascertained *independently* through *both* affected twins.

Thus in the above example of the 11 concordant pairs of MZ twins, both of the affected co-twins were ascertained independently in four pairs, and of the three concordant pairs of DZ twins in one pair both twins had been ascertained independently.

Therefore

$$C_p MZ = \frac{11 + 4}{11 + 11 + 4} = 58\%$$

$$C_p DZ = \frac{3 + 1}{3 + 30 + 1} = 12\%$$

It should be noted that if an attempt is made to ascertain *all* affected twins in the population then $C = C'$ and therefore

$$C_p = \frac{2C'}{2C' + D}$$

Thus by attempting complete ascertainment of all affected twins, this gets around the problem of deciding whether or not a pair of concordant twins have been independently ascertained. It can also be shown (Smith, 1972a) that if *all* twins with the trait in question have been ascertained then

$$C_w = C_p / (2 - C_p)$$

and

$$C_p = 2C_w / (1 + C_w)$$

so the two concordance rates can be derived one from the other. If the condition under consideration is comparatively uncommon and/or ascertainment is low, no pair of twins is likely to be doubly ascertained ($C' = 0$) and therefore the two concordance rates are equivalent.

On balance the proband concordance rate is to be preferred to the pairwise concordance rate for reasons which are discussed in detail by Allen et al (1967). Unfortunately in many twin studies in the past the mode of ascertainment was either not recorded or not taken into account.

The interpretation put on concordance rates is that for a disorder in which genetic factors are important in aetiology, the concordance rate for MZ twins reared apart will be about the same as for MZ twins reared together, and the concordance rate for MZ twins will be greater than for DZ twins. As was

discussed earlier it is possible to estimate from concordance rates the correlation in liability between twins and from this it is possible to derive the heritability (p. 64). Again it must be emphasized that if the frequency of a disorder is low (i.e. 0.1 % or less), the concordance rate in MZ twins will also be low unless the heritability is very high (Smith, 1970).

Variances and correlations

An idea of the degree of genetic influence on a continuously variable trait may be gauged from the intra (within)-pair and inter (between)-pair variances (Osborne & De George, 1959) and the intraclass correlation.

The intrapair variance

$$= \frac{\sum (A - B)^2}{2N}$$

which has N degrees of freedom.

The interpair variance

$$= \frac{1}{N - 1} \left[\frac{\sum (A + B)^2}{2} - \frac{[\sum (A + B)]^2}{2N} \right]$$

which has $N - 1$ degrees of freedom, where

N = number of twin pairs

A and B = values for the members of each twin pair.

Variances may be compared by dividing the larger by the smaller the ratio being referred to as ' F ', the statistical significance of which can be determined from reference to standard tables of ' F ' values (Fisher & Yates, 1963).

The method of calculation is illustrated from some data on serum cholesterol levels in twins (Osborne et al, 1959).

Since a trait such as serum cholesterol level may well be affected by age, sex and environmental factors and for the sake of simplicity in merely wishing to demonstrate the method of calculation, only the authors' data on adult male twins reared together will be considered. Their figures have been rounded-off to one decimal place. For MZ twins the intrapair variance was 279.5 ($N = 14$) and the interpair variance was 2780.5 ($N = 18$). For DZ twins the intrapair variance was 694.3 ($N = 6$) and the interpair variance was 1519.1 ($N = 6$). Thus comparing the *interpair* variances of MZ and DZ twins (MZ/DZ):

$$\begin{aligned} F &= \frac{2780.5}{1519.1} \\ &= 1.83 \end{aligned}$$

Whereas comparing the *intrapair* variances of MZ and DZ twins (DZ/MZ):

$$F = \frac{694.3}{279.5}$$

$$= 2.48$$

Neither of these 'F' values is statistically significant but since the intrapair variance for DZ twins is more than twice that for MZ twins, this suggests that hereditary factors play a role in the control of normal serum cholesterol levels.

However though intrapair and interpair variances can give an idea of the role of genetic factors in aetiology, they are not in themselves a measure of the degree of genetic determination.

Another approach to the problem is to measure the correlation between pairs of twins, but it is not possible to calculate the usual correlation coefficient because there is no way of deciding which measurement on a pair of twins is *X* and which is *Y*. For this reason a different type of correlation coefficient is determined. This is referred to as the *intraclass correlation coefficient* (*r*) which treats the pairs of measurements symmetrically. It is equal to

$$\frac{\text{interpair variance} - \text{intrapair variance}}{\text{interpair variance} + \text{intrapair variance}}$$

From the intraclass correlation coefficient it is then possible to calculate the heritability (p. 57) since

$$h^2 = r/R$$

where

R = coefficient of relationship

Therefore for MZ twins

$$h^2 = r$$

and for DZ twins

$$h^2 = 2r$$

In the above example the intraclass correlation for male MZ twins reared together

$$= \frac{2780.5 - 279.5}{2780.5 + 279.5}$$

$$= 0.82$$

therefore

$$h^2 = 82\%$$

The intraclass correlation for male DZ twins reared together

$$= \frac{1519.1 - 694.3}{1519.1 + 694.3} = 0.37, \text{ therefore } h^2 = 74\%$$

Various useful mathematical models for the analysis of quantitative data *within the families of identical twins* are discussed in detail by Nance & Corey (1976).

Problems and limitations of twin studies

Twin studies have been helpful in fostering a great deal of research. The results of such studies have emphasized the role of genetic factors in aetiology in a variety of conditions. However there are limitations, both statistical and biological, to the twin method. The statistical problems are those of ascertainment and the interpretation to be placed upon such parameters as intrapair and interpair variances. Edwards (1968) has argued that since all diseases have some genetic predisposition the estimation of such parameters may be a costly way of confirming expectations without providing any useful measure of the intensity of this predisposition. However, recent studies have shown that concordance rates and intraclass correlations can be used to estimate the heritability which is meaningful in terms of measuring the degree of genetic determination.

In the literature much use has been made of an index attributed to Holzinger (1929) as estimating the degree of genetic determination from twin data. This index, often referred to as '*H*', has been variously expressed in terms of concordance rates:

$$(C_{MZ} - C_{DZ}) / (1 - C_{DZ})$$

in terms of intraclass correlations:

$$(r_{MZ} - r_{DZ}) / (1 - r_{DZ})$$

and in terms of intrapair variances:

$$(V_{DZ} - V_{MZ}) / V_{DZ}$$

However, this '*H*' index is an arbitrary index and has no specific genetic interpretation (Cavalli-Sforza & Bodmer, 1971). It is not an estimate of heritability and should therefore not be used for this purpose. For these reasons it has been recommended that the use of the '*H*' index should be discontinued (Smith, 1974).

The biological limitations to the twin method are more complex and difficult to cater for. They include prenatal factors such as position in utero, manner of delivery, and the possibility of shared placental circulation, as well as postnatal factors and perhaps here the main problem is the tendency for twins to share the same environment. It is for this latter reason that comparisons between MZ twins reared together and reared apart can be helpful.

It has been suggested by some that the twin method has not vindicated the time spent on the collection of such data. This may have been true to some extent. Certainly considerable care is needed in the collection, analysis and interpretation of twin data.

Estimation of recurrence risks for genetic counselling

Recurrence risks are based upon either Mendelian principles, in the case of unifactorial disorders, or empiric observations on the frequency of a particular disorder among relatives of affected individuals in the case of multifactorial disorders. The estimation of recurrence risks in both these situations has been discussed in detail by Murphy & Chase (1975). Here we shall only be concerned with the principles of such calculations.

Unifactorial disorders

When considering the probability of an individual having a particular genotype (preclinical case or a heterozygous carrier) it is customary to base such calculations on 'anterior' information only. That is the *prior* probability based on knowledge of the individual's antecedents and sibs. But this ignores 'posterior' information based on the individual's phenotype (clinical findings and test results) and that of any of the individual's offspring. From such posterior information it is possible to calculate so-called *conditional* probabilities. The product of the prior and conditional probabilities is the *joint* probability. The final *posterior* probability of an individual having a particular genotype is the joint probability of getting the observed information given the genotype in question, divided by the sum of this probability and the joint probability of getting the observed information if the individual is normal.

The expression of posterior probabilities in terms of prior and conditional probabilities in this way is known as Bayes' theorem or Bayes' law (Bayes, 1763).

In general terms, if the prior probability of an event A occurring is denoted as $P(A)$, and of A not occurring as $P(\text{not } A)$, and if the conditional probability of event O if A occurs (i.e. probability of O given A) is $P(O|A)$, and if the conditional probability of event O if A does not occur is $P(O|\text{not } A)$, then the probability of A given O

$$P(A|O) = \frac{P(A)P(O|A)}{P(A)P(O|A) + P(\text{not } A)P(O|\text{not } A)}$$

This is illustrated in the case of an apparently healthy man aged 50 whose father died of Huntington's chorea and who wishes to know if his own son may one day become affected. This disorder is inherited as an autosomal dominant trait, the first signs of which usually appear sometime between the ages of 25 and 55. The prior probability of having inherited the disorder from his father is $1/2$. Since approximately 80 % of cases of Huntington's chorea develop symptoms before the age of 50, the chance (conditional probability) that he would not have manifested the disease by this age even if he had inherited the gene is about 20 % ($1/5$). Therefore the joint probability of having inherited the disease and being clinically unaffected at age 50 is $1/10$. The prior probability of not having inherited the disease is $1/2$ and of course the conditional probability of being normal if he has *not* inherited the gene is 1, and therefore the joint probability is $1/2$. The posterior probability of having inherited the disease given that he is apparently unaffected at age 50 is therefore $1/6$:

Probability	Inherited the disorder	Not inherited the disorder
● Prior	$1/2$	$1/2$
● Conditional	$1/5$	1
● Joint	$1/10$	$1/2$
Posterior	$\frac{1/10}{1/10 + 1/2} \approx 1/6$	

The prior probability that his son will have inherited the gene is therefore 1 in 12 or 8.5 %. Of course as each year goes by and the father and son remain healthy so their risks of having inherited the gene decrease. The probabilities that an apparently healthy individual and his or her offspring may have inherited Huntington's chorea, polyposis coli or myotonic dystrophy have been calculated in this way from data (based on clinical findings) in the literature and from personal studies and the results expressed graphically in Figures 8.1, 8.2 and 8.3, respectively.

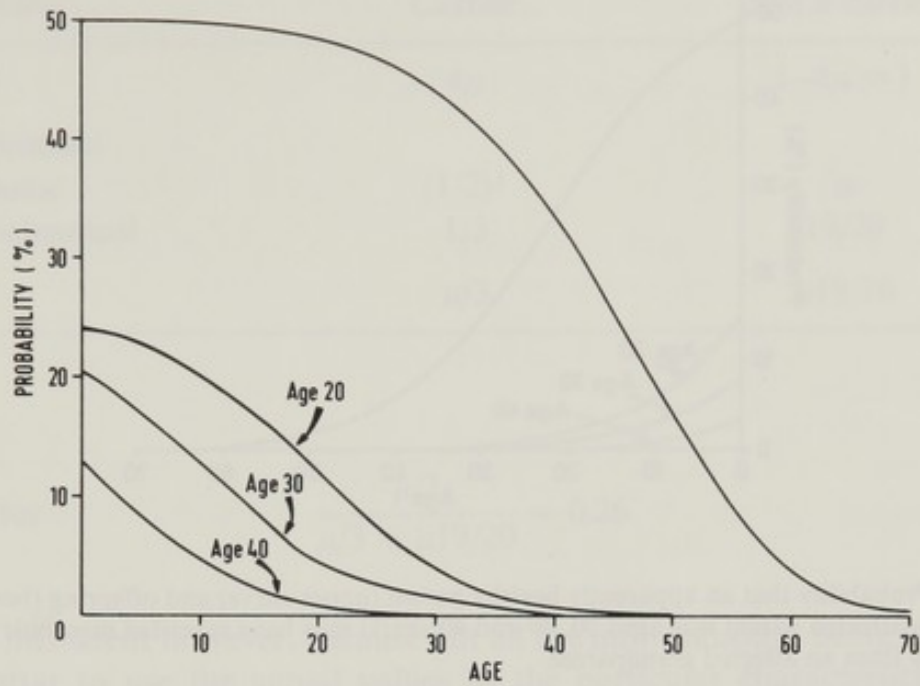


Fig. 8.1 Probability that an apparently healthy parent (upper curve) and offspring (born when the still unaffected parent was aged 20, 30 and 40 years) may have inherited Huntington's chorea from an affected grandparent

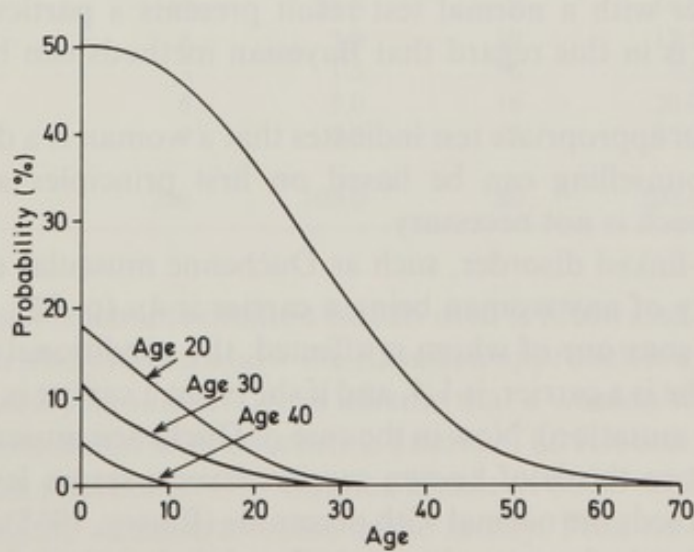


Fig. 8.2 Probability that an apparently healthy parent (upper curve) and offspring (born when the still unaffected parent was aged 20, 30 and 40 years) may have inherited polyposis coli from an affected grandparent

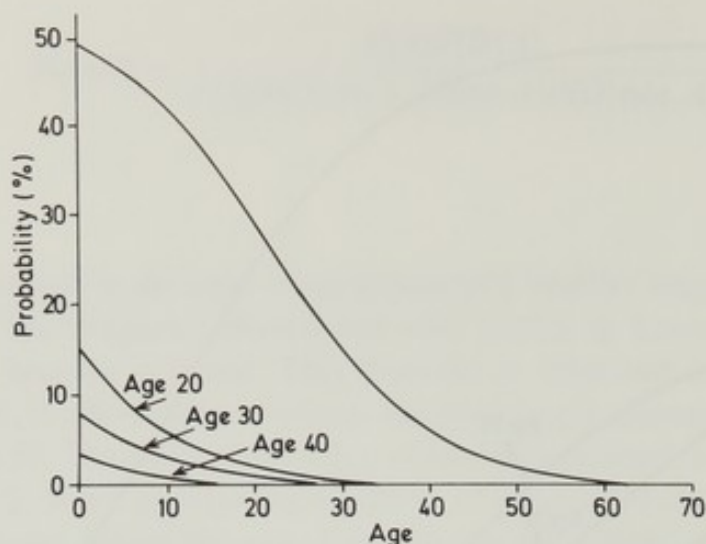


Fig. 8.3 Probability that an apparently healthy parent (upper curve) and offspring (born when the still unaffected parent was aged 20, 30 and 40 years) may have inherited myotonic dystrophy from an affected grandparent

This Bayesian approach to probability calculations and the estimation of genetic risks has been eloquently discussed by Murphy & Mutalik (1969). The method is particularly valuable in the case of X-linked recessive disorders where the problem is to determine the probability of a particular woman being a carrier. The detection of symptomless female carriers is an important problem in genetic counselling. During recent years a number of tests have been devised by means of which it is possible to detect carriers of X-linked recessive disorders. Unfortunately, in some of these tests there is overlap in the results obtained in known carriers and normal women in which case a suspected carrier with a normal test result presents a particularly difficult problem and it is in this regard that Bayesian methods can be particularly helpful.

Of course if an appropriate test indicates that a woman is a definite carrier, then genetic counselling can be based on first principles and this more elaborate approach is not necessary.

In a lethal X-linked disorder, such as Duchenne muscular dystrophy, the prior probability of any woman being a carrier is 4μ (p. 32). If a suspected carrier has two sons one of whom is affected, the conditional probability of this, assuming she is a carrier, is $1/4$, and if she is not a carrier is μ (the affected son being a new mutation). Now in the case of Duchenne muscular dystrophy approximately two-thirds of known carriers have a serum level of creatine kinase which exceeds the normal 95th percentile (Emery, 1965). If a suspected carrier's serum level of creatine kinase is therefore less than the normal 95th percentile then the conditional probability of this if she is a carrier is $1/3$ and if she is not a carrier is $19/20$. Thus:

Probability	Carrier	Not a carrier
● Prior	4μ	$1-4\mu \approx 1$
● Conditional		
genetic	$(1/2)^2$	μ
biochemical	$1/3$	$19/20$
● Joint	$\mu/3$	$\mu 19/20$

$$\text{Posterior} = \frac{\mu/3}{\mu/3 + \mu 19/20} = 0.26$$

This is inefficient however, because not all the information is being used and it is better to use the actual values of the particular characteristic being measured, for example, by taking into account the suspected carriers' actual serum level of creatine kinase and comparing this with values in normal women and known carriers. This is done by first calculating the proportion of normal women and the proportion of known carriers who have a particular serum level of creatine kinase, as shown in Table 8.1.

Table 8.1 Relative probabilities of normal homozygosity to heterozygosity (' h ') for various serum levels of creatine kinase expressed in international units (from Emery, 1980)

Serum creatine kinase (IU)	Controls		Carriers		h (Y_1/Y_2)
	No.	%(Y_1)	No.	%(Y_2)	
11-30	26	13.0	2	2.5	5.20
31-50	112	56.0	10	12.5	4.48
51-70	47	23.5	8	10.0	2.35
71-90	6	3.0	10	12.5	0.24
91-100	3	1.5	6	7.5	0.20
111-170	6	3.0	16	20.0	0.15
> 170	0	0.0	28	35.0	—
Total	200	100.0	80	100.0	—

The way in which such information is then used is illustrated in the following example which also indicates how the Bayesian approach is applied in a more complicated family situation. Let us assume that a woman who seeks genetic counselling has a serum level of creatine kinase of 80 IU, one normal brother, and a sister with a serum level of creatine kinase of 60 IU who has an affected son, there being no one else affected in the family. First we have to go back one generation and consider the mother of these two sisters whose prior probability of being a carrier is of course 4μ . We then consider the conditional probabilities, firstly of her having had a normal son and secondly of having

had a daughter with an affected son and a serum level of creatine kinase of 60 IU. The over-all joint probabilities are then calculated. It is most useful to set out the calculations in clear cut steps as follows:

Consider mother

Probability	Carrier		Not a carrier	
● Prior	4μ		$1 - 4\mu \approx 1$	
● Conditional a normal son	1/2		1	
daughter	Carrier	Not a carrier	Carrier	Not a carrier
● Prior	1/2	1/2	2μ	1
● Conditional affected son	1/2	μ	1/2	μ
SCK 60 IU	0.10	0.24	0.10	0.24
● Joint	0.03	0.12μ	0.10μ	0.24μ
● Joint	0.06μ	$0.24\mu^2$ (negligible)	0.10μ	0.24μ

In the case of the daughter we first determine the prior probabilities of her being a carrier or not a carrier given her mother is or is not a carrier. Secondly, we determine the conditional probabilities of the daughter having an affected son and a serum level of creatine kinase of 60 IU assuming she is or is not a carrier, and finally we determine her joint probabilities. The final over-all joint probabilities are arrived at by multiplying the daughter's joint probabilities by her mother's prior probabilities and her mother's conditional probabilities of having a normal son. The final posterior probability of the *mother* being a carrier, taking into account information on her daughter with an affected son, is the sum of the joint probabilities if she is a carrier (columns 1 and 2) divided by the sum of these probabilities plus the sum of the joint probabilities if she is not a carrier (columns 3 and 4) i.e.:

$$\frac{0.06\mu}{0.06\mu + 0.10\mu + 0.24\mu} = 0.15$$

We now consider the sister who came for counselling who now has a prior probability of being a carrier of 0.075, say 0.08:

Probability	Carrier	Not a carrier
● Prior	0.08	0.92
● Conditional SCK 80 IU	0.13	0.03
● Joint	0.010	0.028

Her (posterior) probability of being a carrier is therefore:

$$\frac{0.010}{0.010 + 0.028} = 0.26$$

Thus despite the fact that both she and her sister have serum creatine kinase levels within the normal range, the sister who requested counselling still has a high chance (i.e. about 1 in 4) of being a carrier.

A general formula for calculating the probability of a woman being a carrier of a *lethal* X-linked disorder, which affects either a brother or a son (*there being no one else affected in the family*) has been derived (Emery & Morton, 1968). If h_c and h_m , based on the results of biochemical and other tests, refer respectively to the relative probabilities of normal homozygosity to heterozygosity (Y_1/Y_2 in Table 8.1) in the suspected carrier and her mother, so that if there is no such information $h = 1$,

and if $q =$ number of normal brothers

and $r =$ number of normal sons

and if $s = 1$ where a son is affected and 0 if a brother is affected

and $t = 0$ where a son is affected and 1 if a brother is affected

then the probability (P) of her being a carrier of a *lethal* X-linked disorder:

$$P = \frac{1 + sa}{1 + sa + ab + tb}$$

where

$$a = h_m 2^q$$

and

$$b = h_c 2^r$$

If the frequency of carrier females is $H\mu$ (i.e. 4μ when the fitness of affected males is 0, as in Duchenne muscular dystrophy, or 18μ when the fitness of affected males is 0.7, as in haemophilia A (see p. 33), then the probability of

a woman being a carrier of *any X-linked disorder*:

$$P = \frac{1 + sa'}{1 + sa' + a'b + tb}$$

where

$$a' = a4/H$$

For example in the case of haemophilia A where $H = 18\mu$

$$\begin{aligned} a' &= a4/18 \\ &= 0.22a \end{aligned}$$

therefore

$$P = \frac{1 + 0.22sa}{1 + 0.22sa + 0.22ab + tb}$$

Returning to the problem of Duchenne muscular dystrophy, formulae have been derived which also take into account information on serum levels of creatine kinase in the first-degree post-pubertal female relatives of a suspected carrier (Emery & Holloway, 1977). The probability of a woman being a carrier (where again there is only one affected individual in the family) then becomes:

$$P = \frac{1 + saf}{1 + saf + afbd + tbd}$$

where

$$d = \prod_{i=1}^n \frac{Y_{1i}}{0.5(Y_{1i} + Y_{2i})}$$

where

Y_1 = proportion of normal women with a particular *daughter's* serum level of creatine kinase

Y_2 = proportion of definite carriers with this serum level of creatine kinase.

Similarly,

$$f = \prod_{i=1}^n \frac{Y_{1i}}{0.5(Y_{1i} + Y_{2i})} \quad \text{for each sister}$$

If there is no such information on daughters then $d = 1$, and if there is no such information on sisters then $f = 1$. Thus if the daughters are currently too young for their serum levels of creatine kinase to be meaningful, and if sisters are not available then the above formula reduces to:

$$\frac{1 + sa}{1 + sa + ab + tb}$$

as before.

To illustrate how the formula is applied, consider the situation in Figure 8.4.

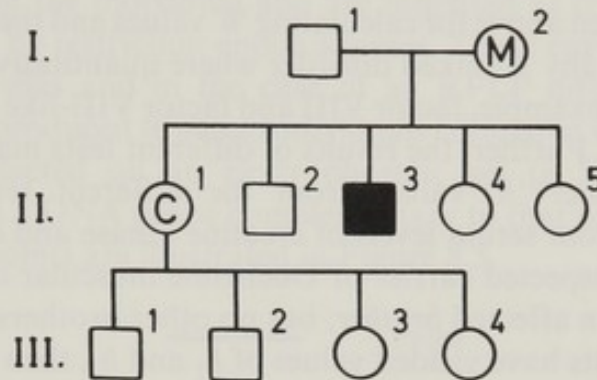


Fig. 8.4 Hypothetical family in which II₁ wishes to know her risk of being a carrier of Duchenne muscular dystrophy

Here the woman who is seeking advice (II₁) has an affected brother ($t = 1$; $s = 0$), one normal brother ($q = 1$), and two normal sons ($r = 2$). If serum levels of creatine kinase are 80 IU in I₂ ($h_m = 0.24$) and 60 IU in II₁ ($h_c = 2.35$), and if serum creatine kinase levels in II₄ and II₅ are respectively 80 IU and 60 IU, then

$$f = \left[\frac{3.0}{0.5(3.0 + 12.5)} \right] \left[\frac{23.5}{0.5(23.5 + 10.0)} \right] = 0.54$$

and if the serum creatine kinase levels in III₃ and III₄ are 20 IU and 40 IU respectively, then

$$d = \left[\frac{13.0}{0.5(13.0 + 2.5)} \right] \left[\frac{56.0}{0.5(56.0 + 12.5)} \right] = 2.74$$

Therefore the probability of II₁ being a carrier, taking into account all this information, is

$$\frac{1}{1 + (0.24)(2)(0.54)(2.35)(4)(2.74) + (2.35)(4)(2.74)}$$

$$= \frac{1}{33.43}$$

or approximately 3%. Though the formulae used in these calculations look a little formidable their attraction is that they are readily amenable to computer programming. A program which can be adapted for this purpose

is PEDIG (Heuch & Li, 1972; Conneally & Heuch, 1974).

One further complication, however, requires consideration. In the Becker type of X-linked muscular dystrophy, serum levels of creatine kinase decrease with age in carriers, which means that in this disorder, in determining the probability of a woman being a carrier, her age as well as her serum level of creatine kinase must be taken into account (Skinner et al, 1975).

The method given above for calculating 'h' values and the general principles involved apply to any X-linked disorder where quantitative data on carriers are available. For example, factor VIII and factor VIII-like antigen in carriers of haemophilia A. Further, the results of different tests may be combined by multiplying together 'h' values from the different tests, for example, combining data from serum levels of creatine kinase and electromyography in the case of a suspected carrier of Duchenne muscular dystrophy. Thus a woman who has an affected *brother*, but no other brothers and no sons, and if two different tests have yielded values of h_1 and h_2 , then the probability of her being a carrier is:

$$\frac{1}{1 + 2h_1h_2}$$

If the problem had been that she had an affected *son*, but no brothers and no other children, then the probability would have been:

$$\begin{aligned} & \frac{1}{1 + (h_1h_2)/2} \\ &= \frac{2}{2 + h_1h_2} \end{aligned}$$

It should be noted however, that it is only legitimate to multiply h_1 and h_2 if the two tests are statistically independent, i.e. they are not positively correlated for controls or carriers.

The Bayesian method of calculating probabilities, based on estimating values of 'h', can provide little information if the data are limited and is unnecessary when the results of a particular test indicate a clear dichotomy between normal women and carriers. The method is most valuable in those X-linked disorders where there is overlap in test results in normal women and carriers. The particular method chosen for estimating 'h' will depend upon the amount of data available. As we have seen, values for 'h' can be estimated from an arbitrary classification into normal and abnormal if the data are limited, or from density functions if the data are extensive.

A major development in recent years has been the introduction of DNA markers for carrier detection and the way in which such information can be combined with, say, serum creatine kinase data and used in genetic counselling deserves special consideration.

Linkage and DNA markers

Of interest in the present context has been the demonstration of linkage between loci for particular genetic diseases and what are referred to as restriction fragment length polymorphisms (RFLPs) (Emery, 1984). The latter are normal variations in base sequences of the DNA which have no apparent effects on the individual and are inherited as co-dominant traits. They are detected by restriction endonucleases. These enzymes cut DNA at sequence specific sites and in the case of an RFLP different sized DNA fragments will be produced in some individuals compared with others. These fragments are detected on an electrophoretic gel (a Southern blot) by hybridization with a DNA probe complementary to that particular region of the DNA. These points are illustrated in Figure 8.5.

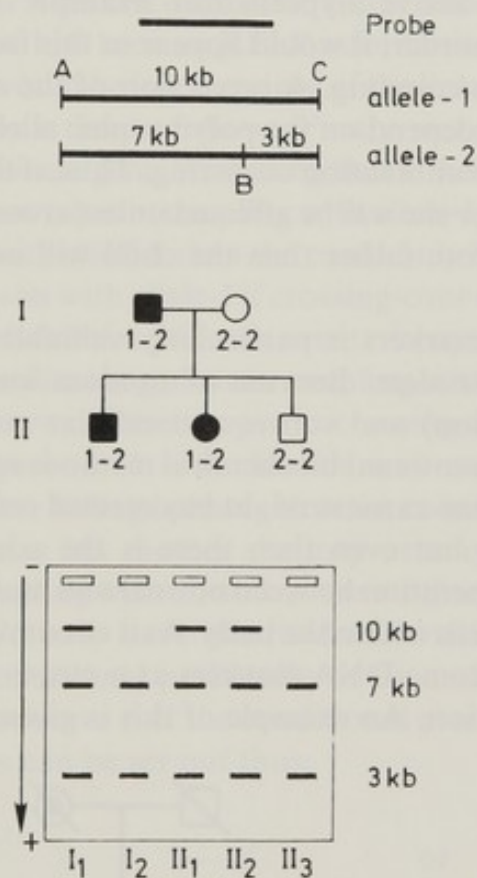


Fig. 8.5 Diagrammatic representation of an autosomal RFLP (above), its inheritance within a family in which three members are affected with an autosomal dominant disorder (middle), and the appearance of the resultant restriction fragments on a Southern blot (below)

Here it is assumed there is a polymorphism at restriction site *B*, the absence of the site is here called allele-1 and the presence of the site is allele-2. When the restriction enzyme cuts the DNA in one chromosome at sites *A* and *C* it generates a single fragment of size 10 kb (1 kb = 1000 base pairs) which corresponds to allele-1. If the enzyme cuts the DNA not only at sites *A* and *C* but also at *B*, two fragments will now be generated of sizes 7 kb and 3 kb

which corresponds to allele-2. Polymorphic genotypes can therefore be deduced from the pattern of bands on an electrophoretic gel. In this example the affected parent is heterozygous (1-2) for the polymorphism as are the two affected children (II_1 and II_2) whereas the youngest unaffected child (II_3) is homozygous for allele-2.

The interest in RFLP's is that if close linkage can be found with a disease locus then this can be useful for preclinical and prenatal diagnosis and, in X-linked disorders, the detection of female carriers. Such information is most useful when linkage is very close, as when the polymorphism occurs within the gene itself or is only a few hundred base pairs distant (1% recombination or 1 centiMorgan is roughly equivalent to 1000 kb). However, if the polymorphism is some distance from the disease locus then the possibility of recombination has to be taken into account. If studies had shown that the polymorphism in the above hypothetical example were linked to an autosomal dominant disorder, it would appear in this family that the disease locus and allele-1 are in coupling. A prediction of the disease status in any subsequent child would depend on the polymorphic allele he or she inherited and the possibility of recombination occurring. Thus, if the next child inherits allele-1 from father, he or she *will* be affected unless crossing-over occurs, but if allele-2 is inherited from father then the child will *not* be affected unless crossing-over occurs.

Linkage with DNA markers is particularly valuable in detecting female carriers of X-linked disorders. Because of random inactivation of the X-chromosome (Lyonisation) and subsequent cellular mosaicism in females, carrier detection by conventional biochemical methods on serum can never be entirely satisfactory. Some carriers might be detected only by studying single cells or clones of cells, but even then there is the added complication of possible metabolic co-operation between normal and mutant cells, or even the suppression of mutant cells within the body. As it circumvents these problems, linkage with X-chromosome DNA markers as a means of detecting carriers has considerable attraction. An example of this is given in Figure 8.6 where

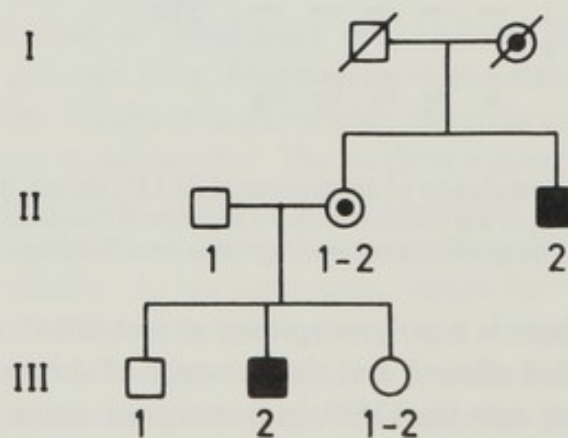


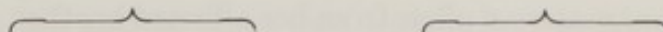
Fig. 8.6 Pedigree of an X-linked recessive disorder linked to an RFLP, the alleles of which are represented below the pedigree symbols

an RFLP (allele-1 and allele-2) has been found to be linked (recombination fraction θ) to an X-linked disease.

Inspection of the pedigree suggests that the disease gene in this family is in coupling with allele-2. The probability of III₃ being a carrier therefore depends on which of her mother's polymorphic alleles she inherited. If she inherited allele-2 then she *will* be a carrier unless crossing-over occurs, and therefore in this instance the probability of her being a carrier is $(1 - \theta)$. But if she inherited allele-1 she will *not* be a carrier unless crossing-over occurs, and therefore the probability of her being a carrier is then equal to the recombination fraction θ . In the example under consideration, III₃ is heterozygous for the RFLP and since she had to inherit allele-1 from her father she must have inherited allele-2 from her mother. She is therefore likely to be a carrier unless crossing-over occurred. Incidentally, contrary to what might be expected in X-linked disorders, useful information can be provided by the father. If there had been *no* RFLP information on the affected male in generation II there would be less certainty of the linkage phase in mother. The method of calculation is then as follows. Firstly, we consider mother who must be a carrier of the disease gene. Assuming that either linkage phase is equally likely (the disease gene (d) is in coupling with allele-1 or allele-2), if the disease gene is in coupling with allele-1 then she could only have an affected son with allele-2 or a normal son with allele-1 if crossing-over occurred in each case, i.e. the conditional probabilities of both events is equal to the recombination fraction θ . But if the disease gene is in coupling with allele-2 then both these events could only occur if there was no crossing-over, i.e. the conditional probabilities are equal to $(1 - \theta)$. In this way we can calculate the joint probabilities for either linkage phase. Now we consider the daughter and for either linkage phase in her mother she may or may not be a carrier. If the disease gene is in coupling with allele-1 in the mother, then given the daughter is a carrier she could only have inherited maternal allele-2 if crossing-over occurred (i.e. the conditional probability is θ), and if she is not a carrier if crossing-over did not occur (i.e. the conditional probability is $1 - \theta$). And so on. The calculations can be set out thus:

Consider mother (II₂)

Probability	$1 - d$	or	$2 - d$
● Prior	$1/2$		$1/2$
● Conditional			
affected son with allele-2	θ		$1 - \theta$
normal son with allele-1	θ		$1 - \theta$
● Joint	$\frac{\theta}{2}$		$\frac{(1 - \theta)^2}{2}$



Consider daughter (III ₃)	Coupling with allele-1		Coupling with allele-2	
	Carrier	Not a carrier	Carrier	Not a carrier
● Prior	1/2	1/2	1/2	1/2
● Conditional Carrier or non-carrier with maternal allele-2	θ	$1 - \theta$	$1 - \theta$	θ
● Joint	$\frac{\theta^3}{4}$	$\frac{\theta^2(1 - \theta)}{4}$	$\frac{(1 - \theta)^3}{4}$	$\frac{\theta(1 - \theta)^2}{4}$

Note that the joint probability in column 1 is the probability that the daughter is a carrier and the disease gene is in coupling with allele-1 in her mother, whereas in column 3 it is the probability that the daughter is a carrier and the disease gene is in coupling with allele-2 in her mother. The overall posterior probability that the daughter is a carrier irrespective of the linkage phase in her mother is therefore the sum of the joint probabilities in columns 1 and 3 divided by the sum of these probabilities plus the sum of the joint probabilities if she is *not* a carrier (columns 2 and 4), i.e.

$$\frac{\theta^3 + (1 - \theta)^3}{\theta^3 + (1 - \theta)^3 + \theta^2(1 - \theta) + \theta(1 - \theta)^2}$$

$$= \frac{1 - 3\theta + 3\theta^2}{1 - 2\theta + 2\theta^2}$$

If linkage is fairly close then in such a situation this refinement does not affect the risks to such an extent as to influence the individual's likely course of action. Thus if $\theta = 0.10$ then disregarding uncertainty of the linkage phase in mother, her daughter's chance of being a carrier is 90%, but if this uncertainty is taken into account then her chance of being a carrier works out to be 89%!

The calculation of risks is more difficult when there is only one affected individual in the family, a situation which is becoming increasingly frequent nowadays, partly as a result of genetic counselling in affected families. The affected individual in such a family may represent a new mutation and there is no certainty as to the linkage phase. Let us assume that mother is heterozygous for the DNA polymorphism (1-2) and that her husband and her only son who is affected both have allele-2 whereas her daughter, whose risk is to be determined, is heterozygous. The latter, having inherited allele-2 from her father, must therefore have inherited allele-1 from her mother, that is a *different* maternal allele from her affected brother. The method of calculation

is as follows. Firstly, we consider mother whose prior probability of being a carrier is of course 4μ and there is an equal chance that the disease gene is in coupling with allele-1 or allele-2. Given that mother is a carrier with the disease gene (d) in coupling with allele-1 then she could only have had an affected son with allele-2 if there had been a cross-over, i.e. the conditional probability of having an affected son with allele-2 is equal to the recombination fraction θ . However if the disease gene in mother is in coupling with allele-2 then she could only have had an affected son with allele-2 if crossing-over did *not* occur (i.e. $1-\theta$). In this way we can calculate the joint probabilities for the mother being a carrier or not being a carrier. Now we consider the daughter who may or may not be a carrier. If her mother is a carrier with the disease gene in coupling with allele-1, then given her daughter is also a carrier, she could only have inherited maternal allele-1 if crossing-over did *not* occur ($1-\theta$), and if she is not a carrier, only if crossing-over did occur (θ); and so on. The calculation can be set out thus:

Consider mother:

Probability	Carrier		Not a carrier
● Prior	4μ		1
	2μ (1 - d)	2μ (2 - d)	
● Conditional affected son with allele-2	θ	$1 - \theta$	μ
● Joint	$2\mu\theta$	$2\mu(1 - \theta)$	μ

Consider daughter:

	Carrier		Not		Carrier		Not	
	1/2	1/2	1/2	1/2	2μ	1		
● Prior								
● Conditional carrier or non-carrier with maternal allele-1	$1 - \theta$	θ	θ	$1 - \theta$	$1/2$	$1/2$		
● Joint	$\frac{2\mu\theta(1 - \theta)}{2}$	$\frac{2\mu\theta^2}{2}$	$\frac{2\mu\theta(1 - \theta)}{2}$	$\frac{2\mu(1 - \theta)^2}{2}$	$\frac{2\mu^2}{2}$	$\frac{\mu}{2}$		
					(negligible)			

● Posterior
(of being a carrier)

$$\frac{4\theta(1 - \theta)}{4\theta(1 - \theta) + 2\theta^2 + 2(1 - \theta)^2 + 1}$$

$$= \frac{4\theta(1 - \theta)}{3}$$

Serum creatine kinase data can also be taken into account by incorporating

it as another conditional probability (see p. 98). If the daughter's serum creatine kinase level is such that Y_1 of normal women and Y_2 of known carriers have this level, then the final probability becomes:

$$\begin{aligned} &= \frac{[4\theta(1-\theta)]Y_2}{[4\theta(1-\theta)]Y_2 + [2\theta^2 + 2(1-\theta)^2 + 1]Y_1} \\ &= \frac{4\theta(1-\theta)}{4\theta(1-\theta) + [2\theta^2 + 2(1-\theta)^2 + 1]h} \end{aligned}$$

where

$$h = \frac{Y_1}{Y_2}$$

Thus if there is no information on serum creatine kinase ($h = 1$), and if $\theta = 0.5$ (i.e. essentially no data from a DNA probe), then the probability of the daughter being a carrier becomes:

$$\frac{2(\frac{1}{2})}{2(\frac{1}{2}) + \frac{1}{2} + 2(\frac{1}{2})^2 + 1}$$

or one-third, which is what would be expected.

The probability of a sister of an isolated case of Duchenne muscular dystrophy being a carrier has been calculated assuming various SCK levels and recombination fractions (Fig. 8.7).

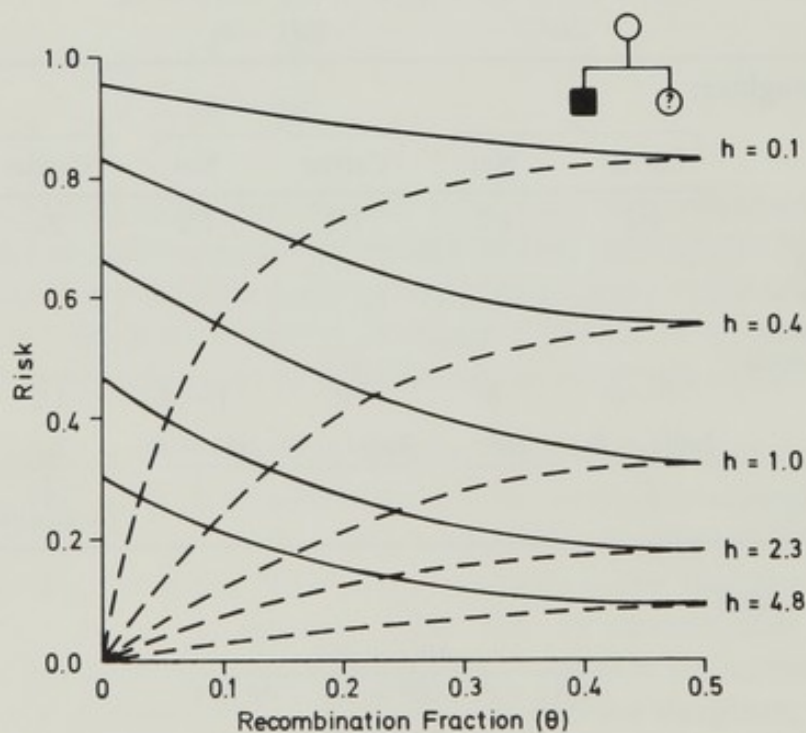


Fig. 8.7 The risks of a sister of an isolated case of Duchenne muscular dystrophy being a carrier assuming she has inherited the same (—) or different (---) maternal RFLP allele as her affected brother, and for various serum creatine kinase levels corresponding roughly to the normal percentiles of 99 ($h = 0.1$), 98 ($h = 0.4$), 95 ($h = 1.0$), 90 ($h = 2.3$), and 50 ($h = 4.8$)

The risks in such situations may be further refined by including information on the number of normal brothers she may have, her mother's serum creatine kinase level, and, particularly important, *her maternal grandfather's haplotype*. To reduce the rate of misdiagnosis resulting from recombination between the disease locus and a DNA marker, information from probes on either side of the disease locus (flanking probes) can also be very helpful. Taking into account all such information in order to give an over-all risk figure is very valuable (Clayton & Emery, 1984).

Computer programs for assessing genetic risks which take into account information on linked DNA probes are now available. The program LIPED (Ott, 1974), which is widely used to compute lod scores in linkage analysis, has been adapted to include DNA probe data for counselling, and details are given, for example, in Conneally et al (1984), Winter (1985), Clayton (1985).

Dominant disorders with reduced penetrance

In autosomal dominant disorders, genetic counselling is relatively straightforward when there is more than one affected individual in the family, even if penetrance is reduced, or if there is only one affected individual but the disorder is always fully penetrant. In the latter situation then, barring illegitimacy, the affected individual must be a new mutation and therefore the chance of recurrence in any subsequent children is negligible. But in disorders with reduced penetrance, an isolated case in a family may not be a new mutation, because one of the parents may be a clinically normal heterozygote. In the case of apparently normal parents who have had a child with an autosomal dominant disorder with reduced penetrance, the chances of recurrence in a subsequent child may be calculated in the following way. The (prior) probability that *either* parent is heterozygous but unaffected because the gene is non-penetrant is $4pq(1 - P)$ where P is the penetrance. Given one of the parents is heterozygous, then the conditional probability of having a child who has inherited the mutant gene and is also affected is $P/2$. On the other hand, given that neither parent is heterozygous, then the conditional probability of their having a child with a new dominant mutation is 2μ which is equal to $2pq(1 - f)$ provided there is balance between mutation and selection. The probability of an affected child is then $2pq(1 - f)P$. The joint and posterior probabilities are then calculated in the usual manner:

Probability	One parent heterozygous	Both parents normal
● Prior	$4pq(1 - P)$	≈ 1
● Conditional	$P/2$	$2pq(1 - f)P$
● Joint	$2pq(1 - P)P$	$2pq(1 - f)P$

The posterior probability that one of the parents is heterozygous is therefore

$$\frac{2pq(1 - P)P}{2pq(1 - P)P + 2pq(1 - f)P}$$

$$= \frac{1 - P}{2 - P - f}$$

The risk of the next child inheriting the mutant gene and also being affected is therefore

$$\frac{P}{2} \cdot \frac{1 - P}{2 - P - f}$$

$$= \frac{P(1 - P)}{2(2 - P - f)}$$

If there is no reduction in fitness ($f = 1$) then the risk simply becomes $P/2$. Alternatively if the disorder confers sterility ($f = 0$) then the risk becomes

$$= \frac{P(1 - P)}{2(2 - P)}$$

However, when dealing with disorders with reduced penetrance, fitness is always assessed only for *affected* heterozygotes (say f') and the true fitness of all heterozygotes (f) will be somewhat greater. Using fitness values derived from affected individuals only, will therefore tend to underestimate risks calculated in this way. However selection against all heterozygotes (s) is equal to selection against those that are affected (s') multiplied by the penetrance:

$$s = s'P$$

therefore

$$(1 - f) = (1 - f')P$$

$$f = 1 - (1 - f')P$$

Substituting this value of f in the above risk equation, the risks to the next child become:

$$\frac{P(1 - P)}{2\{2 - P - [1 - (1 - f')P]\}}$$

which reduces to:

$$\frac{P(1 - P)}{2(1 - Pf')}$$

Thus in tuberous sclerosis where penetrance is around 90 % and fitness of affecteds is about 0.25, then the risks to the sib of an isolated case:

$$\begin{aligned} &= \frac{(0.90)(1 - 0.90)}{2\{1 - (0.90)(0.25)\}} \\ &= 0.058 \end{aligned}$$

or about 6 %. Or in the case of myotonic dystrophy where P is at least 95 % and f' is about 0.75, then the risk is 0.083 (8.3 %).

However, too much faith should not be placed on such figures because estimates of fitness vary considerably in different studies. Further, as techniques for detecting heterozygotes become increasingly more sophisticated and sensitive, so the proportion of heterozygotes who remain completely undetected may become so small that penetrance is virtually 100 %. In the situation where both parents can be shown *not* to be heterozygotes, there would be virtually no risk to subsequent children. However, it has to be admitted that the whole concept of penetrance is a difficult one and has been the subject of much debate in recent years. This is well summarized by Opitz (1981).

Multifactorial disorders

By determining the frequency of a particular disorder among relatives it is possible to predict recurrence risks, for example, to children born subsequent to an affected child in a family. Such information is also important in segregation analysis when unifactorial inheritance is suspected (see Ch. 4) or for calculating the heritability when multifactorial inheritance is suspected (p. 57).

Empiric risk figures for sibs may be determined by considering the proportion of affected individuals among *all* sibs as is usually done in segregation analysis. However this assumes that the risks to children born before the proband are no different from the risks to children born after the proband. This is true for unifactorial disorders but may not be true in other situations, for example, if there is the possibility that the recurrence of the disorder may be related to maternal age or birth order. This problem is illustrated in the case of endocardial fibroelastosis a disorder characterized by progressive cardiac failure beginning in early childhood and associated with gross cardiomegaly and characteristic cardiac histology, possibly at biopsy but usually at autopsy. The cause is not known, but various suggestions have been proposed including autoimmunity, viral infection, a recessive metabolic disorder or a multifactorial aetiology. In an extensive study of 119 families with this disorder, Chen and her colleagues (Chen et al, 1971) found that whereas the frequency of the disorder in the general population is about 0.017 %, the

Table 8.2 Empiric risks (%) for some common disorders. (From Emery, 1983.)

Disorder	Incidence	Sex ratio M:F	Normal parents having a second affected child	Affected parent having an affected child	Affected parent having a second affected child
Anencephaly	0.20	1:2	3-5*	—	—
Asthma	3-4	1:1	10	26	—
Cerebral palsy	0.20	3:2	1†	—	—
Cleft palate only	0.04	2:3	2	7	15
Cleft lip ± cleft palate	0.10	3:2	4	4	10
Club foot	0.10	2:1	3	3	10
Congenital heart disease (all types)	0.50	1:1	1-4	1-4	10
Diabetes mellitus (juvenile, insulin-dependent)	0.20	1:1	6	1-2	—
Dislocation of hip	0.07	1:6	6	12	36
Exomphalos (omphalocele)	0.02	1:1	<1	—	—
Epilepsy (‘idiopathic’)	0.50	1:1	5	5	10
Hirschsprung’s disease short segment	0.02	4:1	3	2	—
long segment			12	—	—
Hydrocephalus (isolated, not XR)	0.05	1:1	3**	—	—
Hypospadias (in males)	0.20	—	10	10	—
Manic-depressive psychosis	0.40	2:3	10-15	10-15	—
Mental retardation (‘idiopathic’)	0.30-0.50	1:1	3-5	10	20
Profound childhood deafness	0.10	1:1	10	8	—
Pyloric stenosis male index	0.30	5:1	2	4	13
female index			10	17	38
Renal agenesis (bilat.) male index	0.01	3:1	3	—	—
female index			7	—	—
Schizophrenia	1-2	1:1	10	16	—
Scoliosis (‘idiopathic, adolescent’)	0.22	1:6	7	5	—
Spina bifida	0.30	2:3	3-5*	4*	—
Tracheo-oesophageal fistula	0.03	1:1	1	1	—

* Risk for anencephaly or spina bifida

† If associated with ataxia, or symmetrical spastic paraplegia or athetosis risk approximately 10%

** Additional 1-2% risk of other neural defects.

frequency among *all* sibs was 3.8% compared with a frequency of 17.7% among sibs born *subsequent* to the index cases. The latter figure is clearly the appropriate one for genetic counselling when parents have already had an affected child. Therefore when determining empiric risks for genetic counselling purposes it is clearly important first of all to exclude the possibilities of a parental age or birth order effect on the recurrence in subsequent sibs. It would be ideal to base recurrence risks always on the

frequency in subsequent sibs. However, in practice, this is often difficult because family limitation, subsequent to the birth of an affected child, may result in insufficient data being available.

Risk tables for genetic counselling in various family situations for cleft lip +/- cleft palate, pyloric stenosis and CNS malformations are available (Bonaiti-Pellié & Smith, 1974). For some relatively common disorders empiric risks to sibs and to the children of affected individuals are given in Table 8.2. These are average figures but are usually adequate for genetic counselling purposes. There are computer programs for risk calculations either disregarding consanguinity, e.g. RISKMF (Smith, 1972b), or taking consanguinity into account (Bonaiti, 1978).

Disease associations

One approach to demonstrating the role of genetic factors in the aetiology of a disorder is to determine if there is any association with an inherited marker trait such as a particular blood group or HLA type. If a disorder is found to be associated with a particular marker more frequently than would be expected by chance, this may suggest a causal relationship, that is the association may be due to multiple effects of the same gene. It should be remembered, however, that association can be due to other causes which include epistatic interaction (e.g. between the Lewis and secretor loci), selective interaction (e.g. between G6PD deficiency, thalassaemia and resistance to malaria in certain areas of the Mediterranean), population stratification (p. 122) and very close linkage resulting in 'linkage disequilibrium' that is certain alleles at adjacent loci are preferentially maintained in coupling (Bodmer et al, 1969).

The first large-scale study of association was made by the late Professor Aird and his colleagues in 1953 (Aird et al, 1953). He had proposed that since cancer of the stomach and blood group O were both commoner in the North of England the two might be associated. In fact the association proved to be not with group O but with group A and the association was highly significant in all parts of the country. Since then, there have been many studies of disease associations either with blood groups (Roberts, 1957; Clarke, 1961; Vogel & Helmbold, 1972) or more recently with HLA types (McDevitt & Bodmer, 1974; Svejgaard et al, 1975; Ryder & Svejgaard, 1981).

Penrose sib method

Penrose (1935) sib method for detecting association (or linkage) depends on the fact that if pairs of sibs are selected at random from a series of families certain types of sib pairs will be more frequent if there is association or linkage than if there is free assortment of the characters studied. The method is also applicable when there are more than two sibs in a family. If there are three sibs then the family will provide three pairs for comparison; if there are four sibs then six pairs can be compared. That is, a family of size s can be partitioned into $s(s-1)/2$ possible pairs. If association or linkage exists the number of

classes where sibs are both alike or are both unlike will be relatively increased over other classes. To determine if such a deviation is statistically significant Fisher's exact probability test may be used (Fisher, 1970). Thus if we consider two traits, say X and Y , then the findings in pairs of sibs can be expressed as:

		Trait X		Total
		Like	Unlike	
Trait Y	Like	a	b	$a + b$
	Unlike	c	d	$c + d$
Total		$a + c$	$b + d$	N

where the total number of sib pair comparisons studied (N) is equal to $(a + b + c + d)$, and the exact probability of observing the proportions in the various classes:

$$= \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{N! a! b! c! d!}$$

where '!' denotes 'factorial' and means successive multiplications in a descending series. Thus $4!$ means $4 \times 3 \times 2 \times 1$ or 24, and by convention $0! = 1$. Tables of factorials are available as well as logarithms of factorials, the latter being necessary when large numbers are involved (Fisher & Yates, 1963). Many pocket calculators are also now available which give factorials.

The application of the method is provided by Penrose who studied the relationship between blood group A and red hair in 60 sib-pair comparisons and obtained the following results:

		Blood group A		Total
		Like	Unlike	
Red hair	Like	40	17	57
	Unlike	0	3	3
Total		40	20	60

The probability of obtaining this distribution of results by chance:

$$= \frac{57! 3! 40! 20!}{60! 40! 17! 0! 3!}$$

$$\approx 1/30$$

Such a result, though indicating a relationship between two traits, does not

give any indication whether this is the result of association or linkage. The Penrose sib method (1935), subsequently elaborated by him (Penrose, 1953a), is useful however inasmuch as the result may indicate that a particular possibility is worthy of further study. Also, since it is not necessary to know parental genotypes, this method can be particularly valuable when there are difficulties in obtaining such data, as in the case of traits that may only become manifest in middle or old age. The method can therefore be of particular value in studying factors associated with longevity. However, though Penrose's method is simple, it is not very efficient, and when there is more than one sib pair per family it becomes less reliable.

Woolf's method

To determine the statistical significance of an association the method most widely used is that of Woolf (1955). This method has the advantage that it allows us to combine data from various centres, in which the marker trait may have different incidences, and it also allows us to test for heterogeneity between centres. The method involves essentially four steps.

1. The relative incidence

The patients and controls are divided into two groups depending on whether they have a particular marker (say α) or not (either β or not α). For example, those with blood group O as compared to those with blood group A, or those without group O (A, B and AB). The relative incidence of the disease in persons with marker α compared to persons with marker β is obtained by cross-multiplication. Thus if

h = number of patients with marker α
 H = number of controls with marker α
 k = number of patients with marker β
 K = number of controls with marker β

then we can draw up a table thus:

marker	patients	controls
α	h	H
β	k	K

and the relative incidence ('x') of the marker α in patients

$$= \frac{hK}{Hk}$$

For example in a large study in Liverpool there were 505 Os and 263 As among patients with duodenal ulcer, and 7536 Os and 6013 As in controls (Clarke, 1961). The relative incidence of duodenal ulcer in persons with group O compared to 1 in persons of group A is therefore

$$\frac{505 \times 6013}{7536 \times 263} \\ = 1.53$$

To test the significance of this finding and in order to combine results from different centres it is necessary to calculate the *total* χ^2 , *pooled* χ^2 and *heterogeneity* χ^2 (for the significance of which see Appendix 2).

2. Total χ^2

If the relative incidence (x)

$$= \frac{hK}{Hk}$$

and if

$$y = \log_e x$$

and

$$w = \frac{1}{\frac{1}{h} + \frac{1}{k} + \frac{1}{H} + \frac{1}{K}}$$

then the significance of 'y' in individual studies is determined by calculating χ^2 for each study which is equal to wy^2 with one degree of freedom. χ^2 values for individual studies are then summed to give the *total* χ^2 ($= \Sigma wy^2$), the number of degrees of freedom of which is equal to the number of studies being combined.

3. Pooled χ^2

This tests the significance of the overall mean value of 'x' from unity and is equal to

$$\frac{(\Sigma wy)^2}{\Sigma w}$$

and has one degree of freedom.

4. Heterogeneity χ^2

This tests the departure of individual values of 'x' from the overall mean. It is obtained by subtracting the pooled χ^2 from the total χ^2 :

$$\Sigma wy^2 - \frac{(\Sigma wy)^2}{\Sigma w}$$

The number of degrees of freedom is one less than the number of studies being combined.

In combining data from several studies the weighted estimated mean value of 'x' is the natural antilogarithm of $\Sigma wy/\Sigma w$ and its SE is the natural antilogarithm of $\sqrt{1/\Sigma w}$.

The method of calculation is illustrated with data from various centres in the UK on the association of peptic ulcer and blood group O (Woolf, 1955). The data and related calculations are summarized in Table 9.1. The results indicate that in all three centres there is a significant association between peptic ulcer and blood group O.

The pooled χ^2

$$\begin{aligned} &= \frac{(\Sigma wy)^2}{\Sigma w} \\ &= \frac{(189.94)^2}{576.0} \\ &= 62.63 \end{aligned}$$

and heterogeneity χ^2

$$\begin{aligned} &= \Sigma wy^2 - \frac{(\Sigma wy)^2}{\Sigma w} \\ &= 65.62 - 62.63 \\ &= 2.99 \end{aligned}$$

With two degrees of freedom $0.2 < P < 0.3$. Therefore there is no apparent heterogeneity in the results of the three studies, which may therefore be combined.

The weighted mean value of 'x' is the natural antilogarithm of

$$\begin{aligned} & \frac{\sum wy}{\sum w} \\ &= \frac{189.94}{576.0} \\ &= 0.33, \text{ the natural antilogarithm of} \\ & \quad \text{which is 1.39.} \end{aligned}$$

Its SE is the natural antilogarithm of

$$\begin{aligned} & \sqrt{1/\sum w} \\ &= \sqrt{1/576.0} \\ &= 0.0417 \end{aligned}$$

The 95% confidence limits are therefore

$$\begin{aligned} & 0.33 \pm (1.96)(0.0417) \\ &= 0.25 \text{ to } 0.41 \end{aligned}$$

Taking natural antilogarithms the 95 % confidence limits are 1.28 to 1.51.

Another instructive example is afforded by data from Los Angeles (Schlosstein et al, 1973) and London (Brewerton et al, 1973) on the association between ankylosing spondylitis and the HLA antigen B27 (Table 9.2). Clearly both studies reveal a highly significant association. Here the total χ^2 is 137.63, and the pooled χ^2 is $(27.20)^2/5.54$ or 133.55. The heterogeneity χ^2 is therefore

$$\begin{aligned} &= 137.63 - 133.55 \\ &= 4.08 \end{aligned}$$

With one degree of freedom this value of χ^2 is just significant ($0.02 < P < 0.05$). There is therefore a suggestion of heterogeneity in the data and it may not be entirely justified to combine the data from these two studies. However if one does, the weighted estimated mean value of the relative incidence is 135 (the natural antilogarithm of $27.20/5.54$ or 4.91). Therefore this association is very much greater than has been observed in the case of any of the blood group associations.

It should be noted however that from a clinical point of view relative incidence is not of great practical value since it only tells us the relative risk of a disease in individuals with a particular HLA type compared with individuals without this HLA type. What a clinician is more likely to want to know is the likelihood of a particular disease if the patient has a particular HLA type. To answer this question we have to use Bayes' theorem (p. 93). Thus in the case of ankylosing spondylitis, the incidence in males is approximately

Table 9.1 The association of peptic ulcer and blood group O relative to blood group A (Woolf, 1955)

Centre	Peptic ulcer				$x = \frac{hK}{Hk}$	$y = \log_e x$	w^*	wy	$\chi^2 = wy^2$	P
	group O (h)	group A (k)	group O (H)	group A (K)						
London	911	579	4578	4219	1.4500	0.3716	304.9	113.30	42.11	<0.001
Manchester	361	246	4532	3775	1.2224	0.2008	136.6	27.43	5.50	<0.02
Newcastle	396	219	6598	5261	1.4418	0.3659	134.5	49.21	18.01	<0.001
					Sum		576.0	189.94	65.62	—

$$*w = \frac{1}{\frac{1}{h} + \frac{1}{k} + \frac{1}{H} + \frac{1}{K}}$$

Table 9.2 The association of ankylosing spondylitis with HLA antigen B27. Data from London (Brewerton et al, 1973) and Los Angeles (Schlossstein et al, 1973.)

Centre	Ankylosing spondylitis				$x = \frac{hK}{Hk}$	$y = \log_e x$	w	wy	$\chi^2 = wy^2$	P
	B27 (h)	not B27 (k)	B27 (H)	not B27 (K)						
London	72	3	3	72	576.00	6.36	1.44	9.16	58.25	<0.001
Los Angeles	35	5	72	834	81.08	4.40	4.10	18.04	79.38	<0.001
					Sum		5.54	27.20	137.63	—

0.2 %, and HLA B27 is present in approximately 90 % of patients and 5 % of controls. Thus if a male patient is suspected of having ankylosing spondylitis and he has HLA B27 then:

Probability	Affected	Normal
● Prior	0.002	0.998
● Conditional HLA B27 positive	0.900	0.050
● Joint	0.0018	0.0499

The posterior probability of such an individual being affected is therefore:

$$\frac{0.0018}{0.0018 + 0.0499} \approx 3.5\%$$

which is much greater than if information on HLA were not available. However it assumes that the physician's suspicion (prior probability) of the disease is merely based on the disease incidence (0.2 %) but in fact it may be considerably greater because of the patient's symptoms and signs. For example, the probability of ankylosing spondylitis in an adult male with persistent low back pain. Ideally such information should be included in the calculations.

Smith's method

Another approach to the problem of disease associations is Professor C. A. B. Smith's method of analysing sibships (see Clarke, 1959a, 1961). The principle of this method is to assess in each sibship in which the particular marker trait under consideration is segregating (in which some sibs have marker trait α and some β) the probability of the proband (the individual with the particular disorder under consideration) having marker trait α and then compare the total 'observed' result with the total 'expected'. Thus in a sibship of four in which two are of group O and two of group A, the 'expected' probability of the proband being group O is 0.5. If in fact he is group O the 'observed' score is 1, whereas if he is in group A the 'observed' score is 0. The observed and expected scores are then summed and the difference compared statistically. The disadvantage of the method is that a great number of families are required for such analysis since many will be uninformative. For this reason the method has not been widely adopted.

Problems of disease association studies

Wiener (1970) has been particularly critical of studies of blood group associations but some of his criticisms are equally valid in *any* study of disease association. Excluding technical problems of erroneous typing and ambiguities in diagnosis and classification of disease, which should really not be problems in present day studies, the other main criticisms are largely concerned with the statistical treatment of the data.

Firstly, if a large enough number of different studies are made between a particular blood group and a particular disease then the results of 1 in 20 of these studies might appear 'significant' by chance alone. Or when studying many different HLA antigens for possible association with one particular disease one would expect that even if none of the antigens is really associated 1 out of 20 would appear associated by chance alone. This statistical problem is referred to as the 'Bonferroni inequality' and Bodmer has suggested one answer to the problem is to multiply each P value obtained by the χ^2 test by the number of antigens tested, i.e. the number of comparisons. Thus with 20 comparisons an individual P value would have to be less than $0.05/20$ or 0.0025 to be significant. Better still the results of a pilot study should be confirmed by a more extensive prospective study.

Secondly, *prior* probabilities of their being an association are not taken into account. If diseases are selected at random and without clear rationale then the likelihood of an association may be remote. In such studies a P value of 0.05 would hardly be enough to overcome the presumption that no association exists. Unless there is a valid biological explanation for an observed association then perhaps a P value of 0.01 might be considered a more appropriate level of statistical significance.

Thirdly, in combining data from different centres there are a number of statistical problems, perhaps the most important of which is pooling heterogeneous data.

Finally, there are the problems of 'stratification' and the choice of adequate controls. For example, there may be a stratum of the population in which both a particular disorder and blood group are especially frequent but with no causal connection between them. Controls must therefore be chosen from the same population as the patients. Also there is the possibility that healthy controls may be biased in favour of blood group O (Vogel, 1970).

With careful selection of controls and appropriate statistical analysis these problems can be avoided. However there remains the problem of the biological relevance of a blood group association. The strongest association is between duodenal ulcer and blood group O and non-secretor, yet even here the contribution of the ABO and secretor loci to the total variance is only about 2.5% (Edwards, 1965). Thus the blood group loci would appear to contribute little to the genetic component of liability. However this seems unlikely to be the case with the HLA loci where the associations with certain diseases are very much stronger (McDevitt & Bodmer, 1974; Svejgaard et al,

1975; Ryder & Svejgaard, 1981). The rather unrewarding results of studies of blood group associations should therefore not deter the investigator from considering disease associations with other marker traits, but always bearing in mind the importance of carefully selecting matched controls and the underlying problems of statistical analysis.

Value of disease association studies

There are a number of important practical reasons for studying disease associations. Firstly, an association with a genetic marker indicates an identifiable genetic component in the aetiology of the disorder. Possible explanations for blood group associations with various non-infectious and infectious diseases have been discussed respectively by Clarke (1961) and Vogel (1970). These associations are comparatively weak. However several associations which have recently been demonstrated with HLA antigens are much stronger (McDevitt & Bodmer, 1974; Svejgaard et al, 1975; Ryder & Svejgaard, 1981). Some significant associations between various blood groups and HLA types are given in Tables 9.3 and 9.4. The figures for relative incidences are only approximate since they continually change as more studies are reported.

One of the most likely interpretations for the strong associations with HLA antigens is that immunological mechanisms, mediated by the HLA loci are involved in pathogenesis perhaps even by immunological cross-reaction between the HLA antigen and a possible aetiological agent(s). It could be that the homozygote for the particular HLA type may be at a higher risk of becoming affected or of manifesting a more severe form of the disease. The identification of groups at risk through their HLA type (or other marker) may be useful in recognizing preclinical cases in families where the marker trait is segregating and where a strong association has been demonstrated between the marker trait and the disorder in question. Clearly such

Table 9.3 Significant associations between blood groups and disease

Disease	Blood group	Relative incidence (ave.)
<i>Non-infectious</i>		
Cancer of various sites	A	1.1-1.6
Pernicious anaemia	A	1.2
Ischaemic heart disease	A	1.2
Duodenal ulcer	O	1.3
Gastric ulcer	O	1.2
<i>Infectious</i>		
Leprosy	A	1.1
Hepatitis	A	1.3
Smallpox	A and AB	6.1
Influenza A ₂	O	1.5

Table 9.4 Significant associations between HLA types and disease

Disease	HLA Antigen	Relative incidence (ave.)
<i>Non-malignant</i>		
Ankylosing spondylitis	B27	135
Reiter's disease	B27	40
Anterior uveitis	B27	10
Coeliac disease	B8	10
Myasthenia gravis	B8	5
	A2	
Multiple sclerosis	B7	5
	A3	1.8
Diabetes mellitus	B8	3
	B8, B15	10
<i>Malignant</i>		
Hodgkins disease	B18	1.9
	B5	1.6
	A1	1.4

information could be valuable in genetic counselling. For example, consider the risk to the children of an individual with ankylosing spondylitis. Family studies have shown that the empiric risk of the disease in first-degree relatives is about 4% (higher in males than females). Further, the chance of having HLA antigen B27 if one has ankylosing spondylitis is approximately 90% whereas the proportion of healthy individuals with this HLA antigen is only about 5%. If an affected parent has B27 antigen and if a particular offspring is also found to have the B27 antigen then the chances of its developing ankylosing spondylitis can be calculated to be in the order of 9%. However if in this case the offspring is found *not* to have B27 then the chances of its developing the disease is less than 1%. Thus in this condition information on HLA typing may significantly affect the genetic advice one gives to relatives.

Finally when there is a strong association with a genetic marker this may be helpful in resolving genetic heterogeneity. For example myasthenia gravis of adult onset may be a heterogeneous disorder because one form has been shown to be associated with HLA B8, has an earlier onset, and thymomas are uncommon. However another form is associated with HLA A2 and has a later onset and thymomas are common (Feltkamp et al, 1974). These findings may have important practical implications as there is a suggestion that the two forms may respond differently to treatment by thymectomy (Fritze et al, 1974). Therefore the resolution of heterogeneity by HLA typing may prove to have considerable practical importance in this disorder.

In conclusion, the study of disease associations has evolved over the last few years. Early studies on blood group associations, though not particularly rewarding, highlighted the importance of carefully choosing controls and applying the right statistical methods. The HLA system is without doubt the most polymorphic locus so far identified in man and therefore there is plenty of scope for studying possible disease associations (Bodmer, 1978; Harris,

1983). The recent demonstration of strong associations with certain HLA types is an exciting new development with important implications and should serve as a stimulus to search for other associations.

Resolution of genetic heterogeneity

It is now well recognized that clinically similar disorders may be genetically different and this is referred to as *genetic heterogeneity*. The recognition of such heterogeneity is important for several reasons: firstly, in order to have accurate risks for genetic counselling; secondly, to know the prognosis in the individual case, since this may be different in disorders which though clinically similar are genetically different; thirdly, a precise genetic diagnosis is essential in interpreting the results of studies designed to investigate aetiology and pathogenesis; finally, genetically different disease entities may well respond differently to any proposed therapy — what is effective in one form of a disease may prove to be ineffective or even deleterious in another.

Genetic heterogeneity may involve mutant genes at different loci or different mutations at the same locus (i.e. different alleles). Heterogeneity may be demonstrated in various ways (Table 10.1). here we shall only be concerned with some relatively simple statistical methods which can be used for demonstrating and resolving genetic heterogeneity.

Table 10.1 Various ways in which genetic heterogeneity may be demonstrated

Clinical

clinical features
response to therapy, etc.

Biochemical

enzyme assays and kinetics
in vitro complementation

Genetic

modes of inheritance
tests for allelism
variations within and between families
consanguinity studies
disease associations
linkage

Molecular

restriction mapping
DNA hybridization, etc.

Pedigree studies

The fact that clinically similar disorders are inherited differently implies that they are due to different genes and there are numerous examples of this phenomenon. In general, recessive forms of a disorder are more severe (earlier onset, more severe manifestations, poorer prognosis) than dominant ones, while X-linked forms tend to be intermediate in severity. This is sometimes referred to as 'Allen's law'. If in a particular disorder clinical differences are sufficiently clearcut then even in an isolated case it may be possible to make a specific genetic diagnosis. For example, of the two commonest forms of mucopolysaccharidosis, in the autosomal recessive form (Hurler's syndrome) there is clouding of the cornea, whereas in the X-linked recessive form (Hunter's syndrome) the cornea is clear and vision is unimpaired. Otherwise these two disorders are clinically almost identical though they differ biochemically. Unfortunately in many genetically heterogeneous disorders there is considerable overlap of clinical features in the different disease entities and a clear diagnosis of a dominant, recessive, or X-linked form may only be possible when there is an extensive pedigree of the disorder. Even if a disorder is clinically similar and inherited in the same manner in two different families this does not necessarily mean that the disorder in the two families is due to exactly the same mutation. Recent studies of the molecular pathology of the thalassaemias, for example, have illustrated this point very well indeed.

From pedigree studies it is possible to test for allelism in two ways. In the case of autosomal recessive traits, when both parents are homozygous for mutant genes but at *different* loci, than all their offspring will be normal. But if they are homozygous for a mutant gene at the *same* locus, then all their offspring will be affected. In this way it has been shown for example that autosomal recessive albinism, deaf mutism and amaurosis can each be due to mutations at different loci. Secondly, in the case of co-dominant traits, if they are due to allelic genes at the same locus, then the individual offspring of a parent who carries two different alleles will inherit either trait, but never both or neither. However, if the mutant genes are not allelic then individual offspring can inherit both traits, one trait or neither trait. That is, genes at the same locus (alleles) segregate, whereas genes at different loci (non-alleles) assort. Such tests for allelism are only possible when the traits are relatively common so that the study of doubly heterozygous individuals is feasible. For this reason they have found most use in studying the genetics of various haemoglobinopathies.

Pedigree studies can provide a simple and valuable approach to resolving genetic heterogeneity but there are serious limitations. Firstly, since many genetic disorders are uncommon, matings which could be informative are often very rare. Secondly, heterogeneity frequently exists between disorders which appear to be inherited in the same manner. Thirdly, many cases may be isolated with no family history. A frequent problem is that the investigator has collected together several cases of a rare disorder in which the clinical

features suggest that more than one disease entity may be involved, but in only a few instances is there another affected relative. It is in this sort of situation that various statistical approaches may be usefully applied.

Analysis of variance

When investigating whether or not significantly heterogeneity exists between families, it is best to consider some characteristic which can be quantified. In practice age at onset is often used but since this is somewhat subjective, more objective measures are preferred, such as age at death, nerve conduction velocity, etc. Differences in the characteristic between families may be merely due to chance, as would be expected for random samples taken from the same population, or may signify real differences between families and thus be indicative of genetic heterogeneity. The usual parametric method for testing whether several independent samples have come from the same population is the so called 'one-way analysis of variance' or F test. But this assumes that the measured characteristic is normally distributed, and family sizes being small, there is no assurance of this. The alternative is to use a non-parametric test, such as the Kruskal-Wallis test (Kruskal & Wallis, 1952; Siegel, 1956). In this test the statistic

$$S = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

where (in studying familial variation)

k = number of families

n_i = number of individuals in each family

N = total number of individuals

R_i = sum of *ranks* in each family

$\sum_{i=1}^k$ = sum over all k families

S is distributed approximately as χ^2 with $(k - 1)$ degrees of freedom.

Consider for example an apparently autosomal recessive disorder in which the age of onset (in months) has been recorded in each sib in 6 families (A to F):

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
	4	4	6	10	2	9
	10	3	6	9	4	7
	6		4		3	8
			12		2	
			8			
mean	6.67	3.50	7.20	9.50	2.75	8.00

Inspection suggests there may be two different diseases with a somewhat earlier onset in families *B* and *E* compared with the other families. Applying the Kruskal-Wallis test, the ages of onset are *ranked*, and when ties occur between two or more values, each score is given the mean of the ranks which are tied. Thus

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
	6.5	6.5	10.0	17.5	1.5	15.5
	17.5	3.5	10.0	15.5	6.5	12.0
	10.0		6.5		3.5	13.5
			19.0		1.5	
			13.5			
n_i	3	2	5	2	4	3
R_i	34.0	10.0	59.0	33.0	13.0	41.0

Therefore

$$S = \frac{12}{19(19 + 1)} \left[\frac{(34)^2}{3} + \frac{(10)^2}{2} + \frac{(59)^2}{5} + \frac{(33)^2}{2} + \frac{(13)^2}{4} + \frac{(41)^2}{3} \right] - 3(19 + 1)$$

$$= 11.96$$

With $(6 - 1)$ degrees of freedom the probability of the observed values occurring by chance is $0.01 < P < 0.05$ (Appendix 2). Thus it seems that in this example age at onset varies significantly between families and suggests that there may well be genetic heterogeneity.

Evidence of bimodality

When studying a particular measurable characteristic in different families, provided the data are sufficient, the frequency distribution curve may appear bimodal, suggesting the existence of two genetically different groups. But it

may be difficult to decide if this apparent bimodality is made up of two distinct but overlapping curves, or merely a dip in an otherwise normal distribution. If a large enough sample is studied mere inspection of the frequency distribution curve may be sufficiently convincing. Otherwise it is necessary to resort to a statistical approach to determine if any apparent bimodality is really significant.

A very simple method was devised by Haldane (1951). If x represents the various measured values, and n_x the number of times each value occurs, if there is bimodality then some value of n_x (after further grouping if necessary) should be significantly different from $\frac{1}{2}(n_{x-1} + n_{x+1})$.

If

$$d_x = n_{x-1} - 2n_x + n_{x+1}$$

and

$$N_x = n_{x-1} + n_x + n_{x+1}$$

the standard deviation $s = \sqrt{2N_x}$

then any value of n_x differs significantly from the mean of its neighbours if

$$\frac{|d_x| - \frac{3}{2}}{s} > 1.96$$

and the exact probability can be read off from tables of the 'normal distribution' (Table I, in Fisher & Yates, 1963).

Consider the following imaginary data of measurements on 100 individuals:

x	0	5	10	15	20	25	30	35	40
n_x	2	4	30	6	4	6	40	5	3

Mere inspection suggests two modes, one at $x = 10$, and the other at $x = 30$.

At $x = 10$

$$d_x = -50$$

$$N_x = 40$$

$$s = 8.944$$

$$\frac{|d_x| - \frac{3}{2}}{s} = 5.423 \quad P < 0.01$$

At $x = 30$

$$d_x = -69$$

$$N_x = 51$$

$$s = 10.100$$

$$\frac{|d_x| - \frac{3}{2}}{s} = 6.683 \quad P < 0.01$$

But at $x = 20$

$$d_x = 4$$

$$N_x = 16$$

$$s = 5.657$$

$$\frac{|d_x| - \frac{3}{2}}{s} = 0.442 \quad P = 0.66$$

Other more complicated examples are given by Haldane (1951) who further refines the test which is a useful preliminary for confirming a suspicion of bimodality.

When biomodality results from the overlapping of two normally distributed curves then the proportion of each group misclassified will be the same (Fig. 10.1), and will be equal to the *one-tail* area under the normal curve. The point of overlap (x), measured in standard deviation units from either mean, can be determined from the means (m_1 and m_2) and standard deviations (s_1 and s_2) of the two curves (Penrose, 1951):

$$x = \frac{m_1 - m_2}{s_1 + s_2}$$

Thus if the point where two curves overlapped corresponded to 1.96 standard deviations from the means of either, then, since 95% of observations lie within 1.96 standard deviations on either side of the mean of a normal curve, 2.5% (one-tail) will lie outside and therefore be misclassified. The percentage misclassification for various values of x can be obtained from tables of the 'normal probability integral' (Table II, in Fisher & Yates, 1963). For convenience the percentage misclassification for various values of x has been plotted (Fig. 10.2).

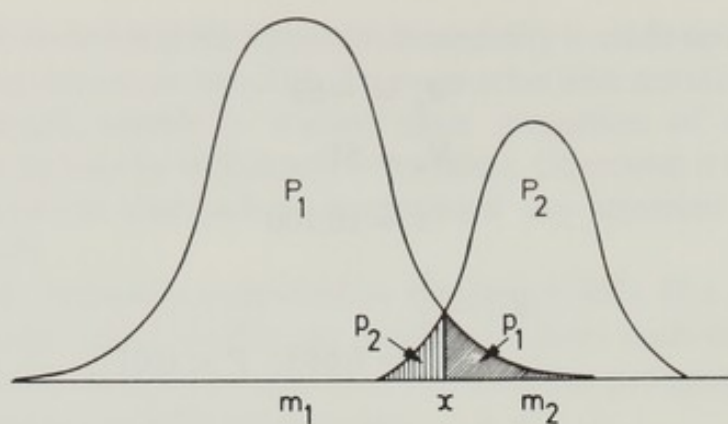


Fig. 10.1 Two overlapping normal distributions. The point of overlap (x) is such that $p_1/P_1 = p_2/P_2$

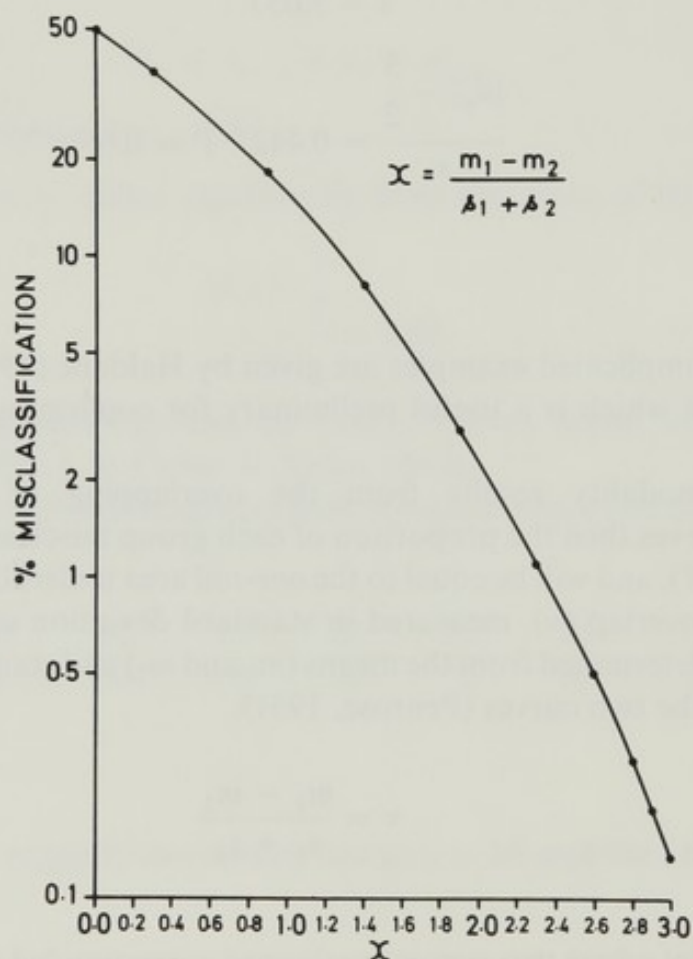


Fig. 10.2 Percentage misclassification for various values of x

This approach has been applied for example to data on Becker and Duchenne types of X-linked muscular dystrophy (Table 10.2). These two disorders are clinically similar in their manifestations though the former is more benign than the latter.

Table 10.2 Clinical course in Becker and Duchenne types of X-linked muscular dystrophy. (From Emery & Skinner, 1976.)

	Onset			Chair-bound			Death		
	No.	Mean age (years)	SD	No.	Mean age (years)	SD	No.	Mean age (years)	SD
Becker	27	11.1	4.9	9	27.1	8.4	10	42.2	13.8
Duchenne	88	2.8	1.5	67	8.6	1.4	26	16.0	2.7
<i>P</i>	<i>P</i> < 0.001			<i>P</i> < 0.001			<i>P</i> < 0.001		
' <i>x</i> '	1.297			1.888			1.588		
% misclassification	9.9			3.1			5.8		

The value of *x* and the percentage misclassification (in parentheses) for age at onset, age of becoming chair-bound and age at death are 1.297 (9.9%), 1.888 (3.1%) and 1.588 (5.8%) respectively. Thus in (100 - 9.9) or 90.1% of boys with Duchenne muscular dystrophy the onset is before the age of 2.8 + (1.297)(1.5) years or 4.7 years, whereas 90.1% of males with Becker muscular dystrophy develop symptoms after the age of 4.7 years. Similarly (100 - 3.1) or 96.9% of boys with Duchenne muscular dystrophy become chair-bound before the age of 8.6 + (1.888)(1.4) years or 11.2 years, whereas 96.9% of males with Becker muscular dystrophy become chair-bound after this age. Finally (100 - 5.8) or 94.2% of boys with Duchenne muscular dystrophy die before the age of 16.0 + (1.588)(2.7) years, or 20.3 years, whereas 94.2% of males with Becker muscular dystrophy die after this age.

This approach is useful because it provides a simple means for determining which of several criteria might be the best for distinguishing between two somewhat similar disorders in, say, an isolated case. In the above example for instance it would appear that the age of becoming chair-bound is the best criterion for distinguishing between Becker and Duchenne types of X-linked muscular dystrophy.

The methods devised by Haldane and Penrose are simple and easy to apply in practice. The problems of resolving bimodality are explored in more detail, for example, by Murphy & Bolling (1967).

Correlations between relatives

An idea of the nature of genetic factors in aetiology and the possibility of genetic heterogeneity may be gained by considering correlations between relatives with regard to some measurable characteristic (such as age at onset, or age at death) associated with a disease (Haldane, 1941; Harris & Smith, 1947). Provided the characteristic being considered is normally distributed, the expected correlations between first-degree relatives will be:

- a. around 0.5 for a major gene with many modifying genes
- b. approach 1.0 if there are two (or more) major genes
- c. approach zero for a single major locus with only random environmental effects.

A number of investigators have used this approach in searching for evidence of heterogeneity. Thus in a study of proximal spinal muscular atrophy of childhood, in 69 sibships, the correlation coefficients (with 95 % confidence limits) for age at onset and age at death (transformed to logarithms because of their skewed distributions) were 0.77 (0.67 to 0.86) and 0.72 (0.47 to 0.87) respectively. These results suggest that heterogeneity exists in this disorder with the operation of at least two major genes (Emery et al, 1975).

Here we have been referring to the usual (product-moment) correlation coefficient. However, if the number of sibs in each sibship varies a great deal then correlations are more reliably estimated from an analysis of variance (see Winter et al, 1981).

If a disease inherited on a multifactorial basis can be split into two (or more) groups on any criterion, Smith (1976) has shown how to test for the groups being genetically distinct. The frequency of the two groups in the general population and among first-degree relatives are determined and the data presented as in Table 10.3.

Table 10.3 Data on probands and their relatives for two disease groups

Proband group	Population frequency	Affected relatives				All relatives
		Group 1		Group 2		
		Number	Proportion	Number	Proportion	
1	P_1	A_{11}	P_{11}	A_{12}	P_{12}	N_1
2	P_2	A_{21}	P_{21}	A_{22}	P_{22}	N_2

Two simple tests can then be applied to the data to determine if the two postulated groups are genetically different. Firstly, a test of *genetic identity* between the two disease groups can be made by a χ^2 test using a 2×2 contingency table whose elements are represented by a , b , c , and d , where $n = a + b + c + d$:

$$\chi^2 = \frac{n \left[|ad - bc| - \frac{n}{2} \right]^2}{(a + b)(c + d)(a + c)(b + d)}$$

Thus in a population and family study of neural tube defects quoted by Smith, probands with anencephaly had 16 first-degree relatives with anencephaly and 13 with spina bifida, while probands with spina bifida had 20 first-degree relatives with anencephaly and 32 with spina bifida. These data can be arranged thus:

Probands with	Relatives with	
	Anencephaly	Spina bifida
Anencephaly	16(a)	13(b)
Spina bifida	20(c)	32(d)

Therefore

$$\chi^2 = \frac{81[|512 - 260| - 40.5]^2}{(29)(52)(36)(45)}$$

$$= 1.48$$

With 1 degree of freedom this is not significant and therefore the genetic liabilities in the two proposed groups do *not* differ significantly. That is, anencephaly and spina bifida are not genetically distinct, but merely different expressions of the same genetic liability.

Secondly, if there is *no* genetic correlation in liability between two proposed groups, that is they are *genetically distinct*, then the proportions P_{12} and P_{21} in Table 10.3 should be equal to the population frequencies P_2 and P_1 respectively. The observed and expected numbers (N_1P_2 and N_2P_1) can be compared using an appropriate statistical test. These two simple tests for suspected heterogeneity in multifactorial disorders are further elaborated by Smith (1976).

Cousins and parental consanguinity

It is possible to estimate the number of gene loci involved in an autosomal recessive disease by:

1. Studying the incidence of the disease in first cousins of affected individuals
2. Comparing the observed incidence of the disease in the general population with that calculated from knowing the frequency of consanguinity in the general population and among the parents of affected individuals
3. Comparing the observed consanguinity rate among parents of affected individuals with that expected if the disorder is due to different numbers of gene loci.

If the incidence of an autosomal recessive disease in the general population is known, it is possible to calculate the expected number of affected first cousins of affected individuals. If the incidence in the general population is I_p , and only one gene locus is involved, then the gene frequency (q) is equal to $\sqrt{I_p}$, and the incidence among first cousins is:

$$\begin{aligned} & \frac{pq}{4} \\ &= \frac{q(1-q)}{4} \\ &\cong \frac{q}{4} \end{aligned}$$

However, if the disease under consideration is genetically heterogeneous with more than one gene locus being involved, then the incidence in first cousins will be *less* than expected. If the disorder is due to mutant genes at n different loci, then the gene frequency for each (assuming they have equal frequencies) is:

$$\sqrt{\frac{I_P}{n}}$$

and therefore the incidence of the disease in first cousins (I_C):

$$\begin{aligned} &= \frac{\sqrt{\frac{I_P}{n}}}{4} \\ &= \sqrt{\frac{I_P}{16n}} \end{aligned}$$

and, following an original suggestion by Steinberg (see Crow, 1965) the number of loci is therefore:

$$= \frac{I_P}{16I_C^2}$$

For simplicity equal gene frequencies have been assumed but it is also possible to make similar calculations for different gene frequencies.

The value of this approach depends very much on the completeness of ascertainment of cases in both the general population as well as among first cousins of affected individuals. Also if the disorder is rare, the differences between the expected incidences in first cousins for different numbers of loci will be slight (Fig. 10.3), and therefore to demonstrate that any differences are statistically significant may require the study of a very large number of affected families, which may not be feasible.

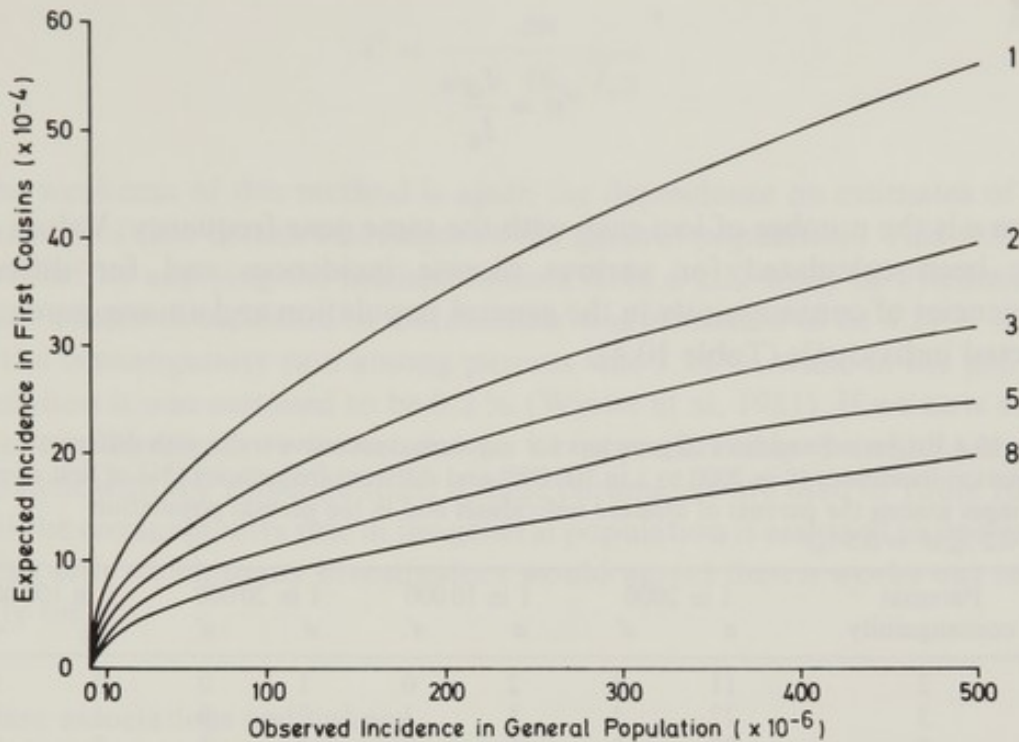


Fig. 10.3 The observed incidence of an autosomal recessive disorder in the general population (I_p), and the expected incidence in first cousins ($(I_p/16n)^{\frac{1}{2}}$) for different numbers of gene loci (1 to 8)

The second approach involves comparing the incidence in the general population with that calculated from the known consanguinity rates in the general population and among parents of affected individuals. Dahlberg (1947) has shown that for an autosomal recessive trait, the gene frequency (q) can be estimated from:

$$\frac{a(1 - C)}{16C - Ca - 15a}$$

where a and C are the frequencies of first cousin marriages in the general population and among the parents of affected individuals respectively (p. 22). The expected incidence of a recessive disorder ($I_E = q^2$) derived in this way can then be compared with the observed incidence (I_O) in the general population. Note that here the expected incidence refers to the expected incidence in the *general population* and not in first cousins as previously. If only one locus is involved then the observed and expected values will be roughly equal. However, if the observed incidence of affected individuals is greater than the expected incidence this could indicate that there is more than one disease locus. In fact the ratio of the observed incidence of a disorder to the expected incidence provides an estimate of the number of loci, homozygosity at any one of which can produce the disease.

Thus

$$n = \frac{I_O}{I_E}$$

where n is the number of loci each with the same gene frequency. Values of n have been calculated for various disease incidences and for different frequencies of consanguinity in the general population and among parents of affected individuals (Table 10.4).

Table 10.4 Estimated numbers of gene loci for autosomal recessive traits with different population incidences (1 in 2000 to 1 in 100 000) and different frequencies (%) of first cousin marriages among the parents of affected individuals and in the general population ($a = 0.2\%$; $a' = 0.5\%$)

Parental consanguinity	1 in 2000		1 in 10 000		1 in 20 000		1 in 100 000	
	a	a'	a	a'	a	a'	a	a'
2	11	1	2	0	1	0	0	0
3	27	3	5	1	3	0	1	0
5	82	12	16	2	8	1	2	0
7	172	25	34	5	17	3	3	1
10	380	57	76	11	38	6	8	1

It will be seen that the values obtained are very much affected by relatively small changes in the frequency of first cousin marriages in the general population and therefore this figure should be determined as accurately as possible in the population under consideration. From the table a rough estimate of the possible number of loci involved in any particular disorder can be obtained. A more rigorous approach to the problem is to be found, for example, in Dewey et al (1965).

Finally, the observed consanguinity rate among parents of affected individuals can be compared with that expected as calculated from Dahlberg's formula (p. 22). If more than one gene locus is involved then the observed value will be *greater* than the expected value. When there is only one gene locus the frequency of first cousin marriages (C) among parents of individuals affected with an autosomal recessive disorder which is relatively rare (and therefore q is very small) is approximately (p. 22):

$$= \frac{a}{a + 16q}$$

$$= \frac{a}{a + 16\sqrt{I_p}}$$

where ' a ' is the frequency of such marriages in the general population. But it can be shown that if there are n different loci then:

$$C = \frac{an}{an + 16\sqrt{I_p n}}$$

The weakness of this method is again the dependence on estimates of the frequency of first cousin marriages in the general population. This point is illustrated by applying the method to data from a UK study of Friedreich's ataxia. The birth incidence of the disorder was estimated to be 4.29×10^{-5} , and the consanguinity rate among parents was 5.38 %, while in the general population it was assumed to be 0.2 % (Winter et al, 1981). If we now solve the above equation, n works out to be approximately 8 (which is also consistent with the value obtained if these parameters are used in Table 10.4). But if the consanguinity rate in the general population is assumed to be nearer 0.5 % (with which many investigators would agree) then n works out to be nearly unity!

Disease associations and linkage

Disease associations and genetic linkage studies may also be used to resolve genetic heterogeneity. For example, two clinically similar forms of myasthenia gravis have been shown to be associated with different HLA types and this has been discussed already (p. 124).

Linkage studies also have great value in the detection and analysis of genetic heterogeneity. One form of a disease may be linked to a genetic marker trait ('test character') and another not, as in the case of elliptocytosis and the rhesus blood group (Morton, 1956). Alternatively two forms of a disease may both be linked to a genetic marker but at different distances so that the distribution of recombination fractions is bimodal. Further, two forms of a disease may both be located on the same side and at similar distances from a genetic marker, which then suggests that heterogeneity is between alleles. It should be noted however that though linkage studies may disprove allelism, because of their relative crudity they can never *prove* allelism. Testing for heterogeneity using linkage is considered in detail by C. A. B. Smith (1963). He describes a method for detecting heterogeneity by comparing likelihood values assuming that in a proportion of families the disease locus is linked to a specific genetic marker with a particular recombination fraction, whereas in the remainder the locus is unlinked. This latter approach has recently been applied, along with other methods, to the possibility of resolving heterogeneity in insulin dependent diabetes mellitus (Harris et al, 1985).

In this brief discussion of methods which may be used in attempting to resolve genetic heterogeneity, emphasis has been on those methods which are relatively simple and easy to apply. However, no one method is likely to produce an entirely convincing answer. Evidence should always be drawn from as many sources (clinical, biochemical and genetic) and analysed in as many ways as possible.

Parental age and birth order

Probably the earliest report of a significant effect of parental age on the incidence of a genetic disorder was Sewall Wright's demonstration in 1926 of a maternal age effect in polydactyly and colour pattern in guinea pig (Wright, 1926). The rationale of studying parental age and birth order effects in human disorders and congenital malformations is that the results of such studies may throw some light on pathogenesis. Thus the demonstration of a parental age effect in sporadic cases of a chromosomal, autosomal dominant or X-linked disorder would indicate that mutation was related to parental age. Conversely in sporadic disorders of unknown aetiology where affected individuals do not reproduce, and so dominant inheritance cannot be proved, the demonstration of a parental age effect would suggest that such cases are perhaps due to fresh dominant mutations. In disorders not inherited in any simple manner (such as many congenital malformations) the demonstration of a parental age or birth order effect provides strong presumptive evidence of an environmental influence. Further, if the incidence of a disorder is shown to be related to parental age or birth order this information could be valuable for genetic counselling provided the effect is large enough. Such information is important in the derivation of empiric risks (p. 111).

So far the only abnormalities shown to be unequivocally related to *maternal* age are certain chromosomal disorders: trisomy-13 (Patau's syndrome), trisomy-18 (Edwards' syndrome) and trisomy-21 (Down's syndrome), and the XXX and XXY (Klinefelter's syndrome). On the other hand a number of unifactorial disorders have been shown to be related to paternal age. These include autosomal dominant disorders such as acrocephalosyndactyly (Apert's syndrome), achondroplasia, Marfan's syndrome, myositis ossificans, bilateral retinoblastoma and to a lesser extent some other dominant disorders (Jones et al, 1975). There is also evidence that paternal age may be a factor in new mutations in X-linked haemophilia A and perhaps Duchenne muscular dystrophy, in these cases the maternal grandfather's age being the important factor. Finally, certain sporadic disorders of unknown aetiology have also been shown to be related to paternal age, such as progeria and acrodysostosis (Jones et al, 1975).

Birth order effects have also been studied extensively (Carter, 1965). First

born children are more often affected in congenital dislocation of the hip and to a lesser extent in congenital pyloric stenosis. On the other hand, haemolytic disease of the newborn is commoner in later born children, and whereas CNS malformations (anencephaly and spina bifida) are commonest in first born children, the incidence rises again in high birth orders. In fact it seems that in CNS malformations among primiparae it is *younger* mothers who are at high risk, whereas among parities of three or more it is *older* mothers who are at greater risk (Fedrick, 1970).

In all such studies the main problem is disentangling the separate effects of maternal age, paternal age and birth order which are all correlated with each other. A number of statistical techniques have been developed for tackling this problem.

Method of Haldane and Smith

Haldane & Smith's (1947) method is perhaps the one most widely used for determining if there is a parental age or birth order effect. In this method the sum of the birth orders of all affected sibs (A) is compared with the theoretical value calculated on the assumption that there is no birth order effect. If A exceeds the theoretical value by more than about twice its standard error we may conclude that later born sibs are more likely to be affected, whereas if A is less than the theoretical value by more than twice its standard error then earlier born sibs are more often affected. The arithmetic is much simplified by testing $6A$ rather than A . Unclassified members of a sibship which occur only at the beginning or at the end of a sibship may be omitted. Thus if 'N' denotes a normal sib, 'a' an affected sib and '—' an unclassified sib, then a sibship —Na would be recorded as a sibship of size 2 with A equal to 2.

From knowing the total number of *classified* sibs (k) and affected sibs (h) in a sibship it is possible to determine the mean and variance of $6A$ from Table 11.2. A special case is when unclassified sibs do not occur at the beginning or end of a sibship. In such a case we have to calculate the mean and variance of $6A$.

$$\text{The mean} = \frac{6hS_1}{k}$$

and the variance

$$= \frac{36h(k-h)(kS_2 - S_1^2)}{k^2(k-1)}$$

where h = number of affected sibs

k = number of classified sibs

S_1 = sum of the birth orders of all ' k ' classified sibs

S_2 = sum of the *squares* of the birth orders for all ' k ' classified sibs

Thus in a sibship $N - a$, the mean of $6A$

$$= \frac{(6)(1)(1 + 3)}{2}$$

$$= 12$$

and the variance of $6A$

$$= \frac{36(2 - 1)(2 \times 10 - 4^2)}{2^2(2 - 1)}$$

$$= 36$$

Having determined the birth order (A) and the mean and variance of $6A$ for each sibship, the data may then be tabulated as in the form given in Table 11.1, which in this case is based on data from 70 non-familial cases of adult onset myasthenia gravis, kindly made available by Dr Anne Jacob (Jacob et al, 1968). In this example the theoretical mean value is 1240 and its standard error is $\sqrt{6399.50}$ or 80.0. The difference between the sum of $6A$ and the theoretical mean value is only 62 which is even less than the standard error. We may therefore conclude that in this disorder there is no significant parental age or birth order effect.

Table 11.1 Analysis of birth order in non-familial adult onset myasthenia gravis. N = normal; a = affected; — = unclassified; k = number of classified sibs; h = number of affected sibs; A = birth order

Family no.	Sibship	k	h	A	$6A$	Mean	Variance
1	aN	2	1	1	6	9	9
2	N——NaN	4	1	6	36	28.5	186.75
3	a	1	1	1	6	6	0
4	NaN	3	1	2	12	12	24
5	aN	2	1	1	6	9	9
6	aNNNNNNNN	9	1	1	6	30	240
7	NNNaNNNN	8	1	4	24	27	189
*8	——NNNNaNN	7	1	5	30	24	144
.							
.							
.							
.							
.							
70	NNNNa	5	1	5	30	18	72
	Total	330	70	217	1302	1240	6399.50

* Unclassified sibs omitted.

Though this test is easy to apply it does not answer the question whether it is maternal age, paternal age, birth order or a combination of these factors which is important. Other methods have to be used to determine the separate effects of these various factors. Further, this test will not be informative if the

Table 11.2—continued

h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
11	36	72	108	144	180	216	252	288	324	360	396	—	—	—	—	—	—	—	—
	360	648	864	1008	1080	1080	1008	864	648	360	0	—	—	—	—	—	—	—	—
12	39	78	117	156	195	234	273	312	351	390	429	468	—	—	—	—	—	—	—
	429	780	1053	1248	1365	1404	1365	1248	1053	780	429	0	—	—	—	—	—	—	—
13	42	84	126	168	210	252	294	336	378	420	462	504	546	—	—	—	—	—	—
	504	924	1260	1512	1680	1764	1764	1680	1512	1260	924	504	0	—	—	—	—	—	—
14	45	90	135	180	225	270	315	360	405	450	495	540	585	630	—	—	—	—	—
	585	1080	1485	1800	2025	2160	2205	2160	2025	1800	1485	1080	585	0	—	—	—	—	—
15	48	96	144	192	240	288	336	384	432	480	528	576	624	672	720	—	—	—	—
	672	1248	1728	2112	2400	2592	2688	2688	2592	2400	2112	1728	1248	672	0	—	—	—	—
16	51	102	153	204	255	306	357	408	459	510	561	612	663	714	765	816	—	—	—
	765	1428	1989	2448	2805	3060	3213	3264	3213	3060	2805	2448	1989	1428	765	0	—	—	—
17	54	108	162	216	270	324	378	432	486	540	594	648	702	756	810	864	918	—	—
	864	1620	2268	2808	3240	3564	3780	3888	3888	3780	3564	3240	2808	2268	1620	864	0	—	—
18	57	114	171	228	285	342	399	456	513	570	627	684	741	798	855	912	969	1026	—
	969	1824	2565	3192	3705	4104	4389	4560	4617	4560	4389	4104	3705	3192	2565	1824	969	0	—
19	60	120	180	240	300	360	420	480	540	600	660	720	780	840	900	960	1020	1080	1140
	1080	2040	2880	3600	4200	4680	5040	5280	5400	5400	5280	5040	4680	4200	3600	2880	2040	1080	0
20	63	126	189	252	315	378	441	504	567	630	693	756	819	882	945	1008	1071	1134	1197
	1197	2268	3213	4032	4725	5292	5733	6048	6237	6300	6237	6048	5733	5292	4725	4032	3213	2268	1197

disorder in question is associated with *both* early and late pregnancies, these two effects tending to cancel each other out. If such a situation is a possibility then the method of Barton & David (1958) may be used to resolve the problem.

Choice of controls

The simplest way of demonstrating a parental age or birth order effect is to make comparisons with the birth of unaffected sibs. However this assumes that the disorder in question is not likely to affect the parents' decision to have further children. A serious disorder present at birth or with onset in childhood may well deter some parents from having further children. In such a situation the use of normal sibs as controls would result in an apparent greater parental age and birth order. This method is therefore only justified in the case of disorders with onset in adulthood.

An alternative approach is to make comparisons with a sample of control families in which the disorder in question does not occur. But here the difficulty is to choose as controls parents who are truly comparable to the parents of affected individuals. This is notoriously difficult.

A simple method which has often been used and which avoids some of these difficulties is to compare maternal age, paternal age and birth order in a series of families with comparable data from the general population. This has been done for example in Apert's syndrome (Blank, 1960), myositis ossificans (Tünte et al, 1967), achondroplasia (Murdoch et al, 1970) and Marfan's syndrome (Murdoch et al, 1972). Some population data on parental age and birth order in three different countries are given in Table 11.3. Unfortunately in Britain population statistics for paternal age were not available before 1961. However on the basis of studies over a number of years of a large series of births in England and Wales, Fraser & Friedmann (1967) have calculated that between 1900 and 1962 the mean difference between paternal and maternal ages at the birth of a child was about 3.1 years. Therefore for this period an approximate estimate of paternal age can be derived from maternal age by adding 3.1 years. Bunday and her colleagues (1975) have derived more accurate estimates of paternal age by determining the appropriate age difference between spouses *according to the mother's age*, since this difference is not the same for all maternal ages, and then adding this difference to maternal age. Armed with such information one may then compare not only the mean parental ages in patients' families with those expected, but also the mean *differences* between parental ages in patients' families with those expected.

It should be noted that in making such comparisons with the general population secular changes in parental ages over the years covered by the births of affected individuals being studied must be taken into account. Comparisons should therefore be made with controls of roughly the *same period of time* and from a *similar environment*. In Table 11.4 are given data on

Table 11.3 Population data on parental age and birth order

Source	Maternal age		Paternal age		Birth order	
	Mean	SD	Mean	SD	Mean	SD
England and Wales, 1950 (Blank, 1960)	28.04	5.97	—	—	2.24	1.57
London, 1960 (Blank, 1960)	—	—	31.69	6.47	1.86	1.21
*England and Wales, 1983	26.45	5.39	29.80	6.12	1.96	1.10
Australia, 1953 (Blank, 1960)	27.65	5.84	31.04	6.79	—	—
United States, 1955 (Murdoch et al, 1970)	26.54	6.07	29.85	6.95	2.64	1.73

* Calculated from data in Registrar General's *Statistical Review of England and Wales, Part 2*. HMSO, London, 1983

parental ages and birth order calculated from information in the Registrar General's Reports for England and Wales.

A simple and effective graphical way of demonstrating a parental age effect is to determine for each age group the number of parents of affected individuals relative to the number in the general population (Fig. 11.1). A five-year interval size is useful for this purpose.

To determine if there is any significant difference between the mean ages of parents and controls one may use the 'student's *t* test'. If a *large* general population is being used for comparison then

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

where *m* = mean parental age in the sample

s = standard deviation of parental age in the sample

n = number of fathers or mothers in the sample

μ = mean parental age in the general population

Thus in a study of Marfan's syndrome (Murdoch et al, 1972) the mean maternal age of 23 sporadic cases was 29.30 (SD 5.36) compared with the mean maternal age in the general population of 26.54 (SD 6.07), i.e. a difference of 2.76 years.

Therefore

$$\begin{aligned} t &= \frac{29.30 - 26.54}{5.36/\sqrt{23}} \\ &= 2.5 \end{aligned}$$

With (*n* - 1) degrees of freedom, i.e. 22, from Tables of student's *t* distribution, *P* = 0.02, and therefore the mean age of mothers at the birth of offspring with Marfan's syndrome is significantly greater than in the general population. In this study however, the mean paternal age was 36.61 (SD 9.06) compared with a mean paternal age in the general population of 29.85 (SD 6.95). Here the difference is 6.76 years which is highly significant (*P* < 0.01).

Table 11.4 Parental age and birth order in England and Wales calculated from information in the Registrar General's Reports

Year	Maternal age		Paternal age		Birth order	
	Mean	SD	Mean	SD	Mean	SD
1940	28.53	5.98	—	—	2.37	1.90
1941	28.55	6.06	—	—	2.36	1.91
1942	28.66	5.99	—	—	2.25	1.81
1943	28.84	6.06	—	—	2.21	1.75
1944	29.06	6.09	—	—	2.26	1.68
1945	29.12	6.19	—	—	2.27	1.69
1946	29.01	5.91	—	—	2.16	1.59
1947	28.54	5.90	—	—	2.09	1.53
1948	28.26	5.99	—	—	2.15	1.56
1949	28.03	5.94	—	—	2.16	1.54
1950	28.04	5.97	—	—	2.24	1.57
1951	28.02	5.89	—	—	2.22	1.54
1952	27.79	5.81	—	—	2.24	1.56
1953	27.70	5.74	—	—	2.23	1.54
1954	27.62	5.75	—	—	2.23	1.54
1955	27.57	5.75	—	—	2.23	1.54
1956	27.46	5.75	—	—	2.22	1.53
1957	27.40	5.75	—	—	2.22	1.53
1958	27.30	5.73	—	—	2.22	1.52
1959	27.22	5.73	—	—	2.24	1.53
1960	27.20	5.77	—	—	2.27	1.53
1961	27.09	5.80	30.17	6.68	2.28	1.54
1962	26.96	5.80	30.04	6.66	2.29	1.54
1963	26.87	5.77	29.93	6.62	2.31	1.54
1964	26.80	5.77	29.85	6.62	2.32	1.53
1965	26.63	5.77	29.67	6.64	2.27	1.49
1966	26.38	5.74	29.42	6.64	2.24	1.47
1967	26.26	5.69	29.27	6.63	2.20	1.43
1968	26.13	5.59	29.11	6.56	2.18	1.40
1969	26.02	5.51	28.94	6.50	2.14	1.36
1970	25.87	5.41	28.75	6.42	2.11	1.31
1971	25.78	5.32	28.60	6.35	2.06	1.27
1972	25.78	5.22	28.59	6.25	2.01	1.22
1973	25.79	5.10	28.58	6.13	1.95	1.16
1974	25.72	5.23	28.59	6.04	1.94	1.13
1975	25.81	5.21	28.71	6.02	1.93	1.11
1976	25.89	5.15	28.80	5.95	1.92	1.09
1977	26.05	5.18	29.01	5.93	1.90	1.07
1978	26.16	5.23	29.15	5.95	1.90	1.07
1979	26.24	5.26	29.28	5.96	1.91	1.07
1980	26.25	5.30	29.32	6.00	1.93	1.08
1981	26.33	5.32	29.49	6.05	1.94	1.08
1982	26.37	5.37	29.63	6.09	1.96	1.09
1983	26.45	5.39	29.80	6.12	1.96	1.10

Thus paternal age is elevated more than maternal age but in fact both are significantly greater than the general population.

Birth order is not normally distributed and therefore it is not statistically legitimate to make comparisons in this way. Further, this approach is limited in distinguishing the separate effects of parental age and birth order. The so-called *Greenwood-Yule method* (Greenwood & Yule, 1914), subsequently modified by McKeown & Record (1956), was developed in order to separate

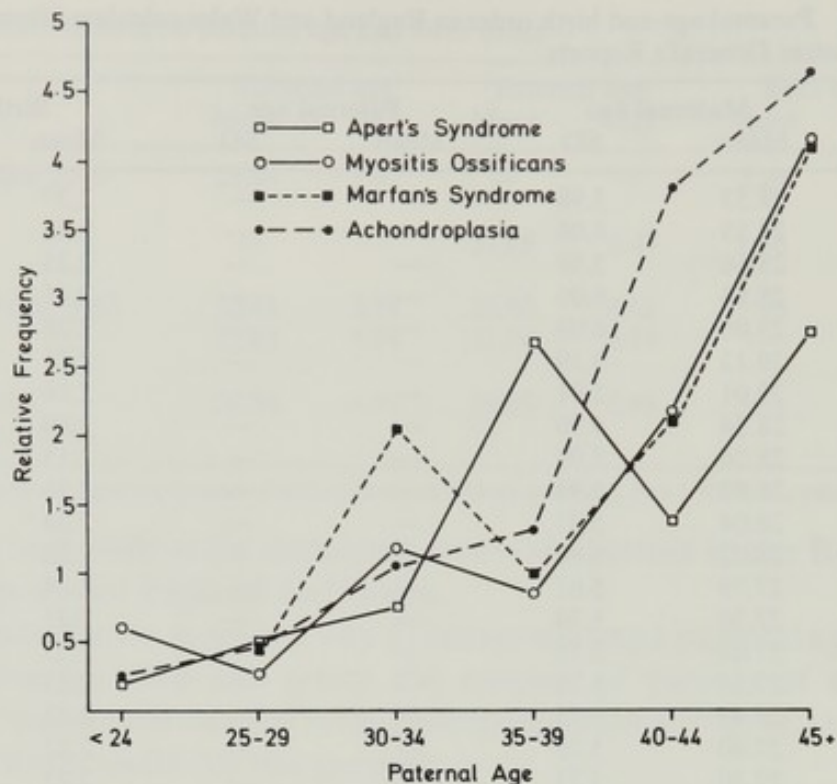


Fig. 11.1 Number of fathers of affected offspring relative to the number in the general population in various age groups (calculated from the original data)

the effects of maternal age and birth order but the method does not take into account paternal age. To evaluate separately these various factors the method of partial correlations is usually used.

Table 11.5 Correlations from various population studies

	Source	Correlation	Reference
Paternal and maternal age	Australia (1953)	0.73	Blank (1960)
	USA (1955)	0.76	Murdoch et al (1970)
	England and Wales (1973)	0.72	Unpublished*
	Scotland (1973)	0.77	Unpublished*
Paternal age and birth order	London (1960)	0.30	Blank (1960)
Maternal age and birth order	England and Wales (1950)	0.49	Blank (1960)
	USA (1955)	0.52	Murdoch et al (1970)
	England and Wales (1973)	0.45	Unpublished*
	Scotland (1973)	0.49	Unpublished*

* Calculated from data in: Registrar General (1975) *Statistical Review of England and Wales for the year 1973*. Part 2. HMSO, London. Registrar General, Scotland (1974) *Annual Report for 1973*. Parts 1 + 2. HMSO, Edinburgh.

Method of partial correlations

This method has been widely used in determining the separate effects of paternal age, maternal age and birth order (Penrose, 1957), though it has to

be recognized that it may not be statistically entirely satisfactory (C.A.B. Smith, 1972).

Essentially the method allows one to compare the effect of a single variable on incidence while other variables are held constant, for example, to estimate the effect of paternal age on incidence while maternal age and birth order are held constant. In statistics wherever there are more than two variables and a correlation between any pair is to be determined, the effect of one or more of the remaining variables being eliminated (held constant), this is referred to as a *partial correlation coefficient*. Thus if there are four variables represented as 1, 2, 3 and 4, then we first determine the usual (product-moment) correlation between each pair of variables: r_{12} , r_{13} , r_{23} , etc. The partial correlation between any pair of variables (e.g. incidence and maternal age, paternal age or birth order) eliminating the other two variables can then be calculated. Thus between 1 and 2 eliminating 3 and 4 (written as $r_{12.34}$) by

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}}$$

where coefficients like $r_{12.4}$ can be calculated by

$$r_{12.4} = \frac{r_{12} - r_{14}r_{24}}{\sqrt{(1 - r_{14}^2)(1 - r_{24}^2)}}$$

In practice, to determine the independent effects of parental age and birth order we calculate the following correlations:

A. From population data (see Table 11.5)

- | | |
|----------------------------------|-----------|
| 1. Paternal age and maternal age | $:r_{PM}$ |
| 2. Paternal age and birth order | $:r_{PA}$ |
| 3. Maternal age and birth order | $:r_{MA}$ |

B. From families being studied

- | | |
|---------------------------------------|-----------|
| 4. Paternal age and disease incidence | $:r_{PI}$ |
| 5. Maternal age and disease incidence | $:r_{MI}$ |
| 6. Birth order and disease incidence | $:r_{AI}$ |

and from these we derive the partial correlations between

- | | |
|--------------------------------------------------------------------------------------|--------------|
| 7. Paternal age and disease incidence, maternal age and birth order being eliminated | $:r_{PI.MA}$ |
| 8. Maternal age and disease incidence, paternal age and birth order being eliminated | $:r_{MI.PA}$ |
| 9. Birth order and disease incidence, paternal age and maternal age being eliminated | $:r_{AI.PM}$ |

The partial correlation between paternal age and disease incidence, birth order being eliminated, is

$$r_{PI \cdot A} = \frac{r_{PI} - r_{PA}r_{IA}}{\sqrt{(1 - r_{PA}^2)(1 - r_{IA}^2)}}$$

and between paternal age and maternal age, birth order being eliminated, is

$$r_{PM \cdot A} = \frac{r_{PM} - r_{PA}r_{MA}}{\sqrt{(1 - r_{PA}^2)(1 - r_{MA}^2)}}$$

and between maternal age and disease incidence, birth order being eliminated, is

$$r_{MI \cdot A} = \frac{r_{MI} - r_{MA}r_{IA}}{\sqrt{(1 - r_{MA}^2)(1 - r_{IA}^2)}}$$

and so on. From these partial correlation coefficients we can then calculate the partial correlation between paternal age and disease incidence, maternal age and birth order being eliminated:

$$r_{PI \cdot MA} = \frac{r_{PI \cdot A} - r_{PM \cdot A}r_{MI \cdot A}}{\sqrt{(1 - r_{PM \cdot A}^2)(1 - r_{MI \cdot A}^2)}}$$

Thus in Blank's study of Apert's syndrome (Blank, 1960):

from population data

$$r_{PM} = 0.73$$

$$r_{PA} = 0.30$$

$$r_{MA} = 0.49$$

from the families being studied

$$r_{PI} = 0.34$$

$$r_{MI} = 0.31$$

$$r_{AI} = 0.14$$

Therefore

$$\begin{aligned} r_{PI \cdot M} &= \frac{0.34 - (0.73)(0.31)}{\sqrt{(1 - 0.73^2)(1 - 0.31^2)}} \\ &= 0.18 \end{aligned}$$

Similarly

$$\begin{aligned} r_{PI \cdot A} &= 0.32, \quad r_{MI \cdot P} = 0.10, \quad r_{MI \cdot A} = 0.28, \\ r_{AI \cdot P} &= 0.04, \quad r_{AI \cdot M} = -0.01, \quad \text{and} \quad r_{PM \cdot A} = 0.70 \end{aligned}$$

Finally the partial correlation between paternal age and disease incidence, maternal age and birth order being eliminated, is calculated

$$\begin{aligned} r_{PI \cdot MA} &= \frac{0.32 - (0.70)(0.28)}{\sqrt{(1 - 0.70^2)(1 - 0.28^2)}} \\ &= 0.18 \end{aligned}$$

Similarly

$$r_{MI \cdot PA} = 0.09 \quad \text{and} \quad r_{AI \cdot PM} = 0.00$$

Therefore paternal age is the main factor because when maternal age and birth order are eliminated there remains a positive partial correlation of 0.18 between paternal age and disease incidence. On the other hand when paternal age and birth order are eliminated then the correlation between maternal age and disease incidence is only 0.09, and there is no correlation between birth order and disease incidence when paternal and maternal age are eliminated.

In all these calculations it is necessary to re-emphasize a word of caution in using population data on parental ages or correlations between parental ages and birth order. The general population with which comparisons are being made must be similar to the parents being studied both in time and place since these parameters are known to be affected by a variety of socio-economic factors.

The significance of an ordinary correlation coefficient, for a relatively small sample size (as is the case in most family studies), may be determined from tables (Appendix 3) or by calculating student's t , which for an ordinary correlation coefficient is

$$= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with $(n - 2)$ degrees of freedom. In the case of a partial correlation coefficient where two variables have been eliminated, student's t is

$$= \frac{r\sqrt{n-4}}{\sqrt{1-r^2}}$$

with $(n - 4)$ degrees of freedom. The significance of t values can be determined from tables (Appendix 1).

In order to determine if two correlation coefficients differ significantly they are first transformed to so-called 'z' values where:

$$z_1 = \frac{1}{2} \log_e \frac{1+r_1}{1-r_1}$$

and

$$z_2 = \frac{1}{2} \log_e \frac{1+r_2}{1-r_2}$$

Fortunately there are tables (see Appendix 4) for transforming 'r' values into 'z' values. We can then calculate the normal deviate:

$$x = \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

If the second sample represents the general population, and therefore n_2 is very large, then the normal deviate becomes:

$$\frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3}}}$$

We can then determine if the difference is significant from tables (p. 155).

Partial correlation coefficients may be compared in the same way as ordinary (product-moment) correlation coefficients except that in the above formulae 'n' is replaced by n minus as many variables as have been eliminated from the comparison in question. Thus if we were comparing the partial correlation coefficients of paternal age and disease incidence with maternal age and disease incidence, in each case eliminating birth order and the other parental age, then the normal deviate would be:

$$\frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 5} + \frac{1}{n_2 - 5}}}$$

In conclusion, proving that a parental age or birth order effect on disease incidence exists, if not obvious on casual inspection of family data, may be difficult. The method of partial correlations is relatively simple to apply but it may not be statistically entirely satisfactory, though it can be used to give at least a first approximation. Probably the best method of estimating the separate effects of maternal age, paternal age and birth order is by *multiple regression analysis*. Multiple regression analysis is a method for analysing the relationships between several variables and for estimating their relative importance. The multiple regression equation is given by

$$y = a + b_1x_1 + b_2x_2 \dots b_nx_n$$

where b_1, b_2, \dots, b_n are the partial regression coefficients of the dependent variable (y , say the incidence of Down's syndrome) on each independent variable ($x_1, x_2 \dots x_n$, say birth order, maternal age and paternal age), and a is the point of intersection on the y axis. Details of the computations, which are complex, are given in textbooks of statistics and nowadays computer programs for multiple regression analysis are available. The use of the technique for dissecting out the independent effects of parental ages are detailed in C.A.B. Smith (1972).

All of these approaches assume that parental ages are normally distributed and birth incidence is linearly related to parental age, assumptions which are not really justified. The resultant weakness of methods of assessing parental age and birth order effects are critically reviewed by Stene & Stene (1977) who propose an alternative 'conditional probability test'. It has to be admitted, however, that in some cases, even with the most sophisticated statistical analysis it may be difficult to separate the effects of birth order from maternal or paternal age which may be more convincingly demonstrated by a simple 3-dimensional grid of birth order and parental age (see for example Record et al, 1969).

Should a parental age or birth order effect be demonstrated this should not be regarded as the end of the investigation. It is rather the beginning of an enquiry, since it may suggest possible lines for further research. In the case of a congenital malformation of unknown aetiology it suggests the importance of environmental factors in causation, which should then be sought for in relation to parental age and/or birth order.

Thus persistent patent ductus arteriosus is commoner in first borns. Now the normal closure of the ductus shortly after birth depends upon adequate oxygenation of the blood. Since difficulties in labour, with possible resultant fetal anoxia, are commoner in first pregnancies than in later pregnancies this may be the explanation for the birth order effect in this abnormality. In fact the incidence of fetal distress is higher among affecteds than would be expected.

Finally if the effect is sufficiently marked, information on parental age and birth order may also be used to construct risk tables for use in genetic counselling as in the case of Down's syndrome and maternal age.

Recognition and estimation of changes in disease frequency

In studying the possible relevance of environmental factors in the aetiology of say a particular congenital malformation, one approach to the problem is to study changes in frequency over time. If there is a dramatic change at a particular point in time one would then attempt to identify the environmental factor which caused this change. Such a change might be recognized by an astute observer without recourse to statistical techniques. An outstanding example of this was the recognition by Lenz in Germany and McBride in Australia of the teratogenic effects of thalidomide, a drug which first appeared on the market as a sedative in the late 1950s. During 1961, Lenz (1961) and McBride (1961) reported that they were seeing many more cases of a rare form of limb deformity (a type of phocomelia) than had been their previous experience. On taking a careful history they discovered that the mothers of these children had all taken the drug thalidomide in early pregnancy. This approach, however, is not possible with relatively common disorders because a very large number of cases would be needed to detect any significant change in frequency. In such situations we have to rely on statistical methods.

Incidence and prevalence

So far, for the sake of simplicity, we have usually referred to the number of cases of a disorder in a population as its *frequency*. There are, however, two estimates of frequency which are not necessarily identical. *Incidence* refers to the number of *new* cases per unit of population. For example, the incidence of Down's syndrome is 1.4 per 1000 live births or about 1 in 700 live births. *Prevalence* on the other hand refers to *all* cases present in a population, either within a given period (so-called *period* prevalence rate) or a particular point in time (so-called *point* prevalence rate), per unit of population at risk at that time. In the case of Down's syndrome prevalence is much less than incidence because of early mortality in this condition. In the case of congenital malformations incidence is more precisely known than prevalence, the latter being notoriously unreliable.

Comparison of proportions

If we wish to determine if there has been a significant change in the frequency (incidence or prevalence) of a particular disorder we could merely compare the proportion of cases in one year with the proportion in another using standard statistical techniques.

Thus if the proportion of cases (P_1) in the first period was

$$n_1/N_1$$

and in the second period (P_2) was

$$n_2/N_2$$

and if

$$P_0 = \frac{n_1 + n_2}{N_1 + N_2}$$

then in the usual manner (see Snedecor & Cochran, 1967) we can calculate 'x', the so-called normal deviate, where

$$x = \frac{|P_1 - P_2|}{\sqrt{P_0(1 - P_0)(1/N_1 + 1/N_2)}}$$

(The vertical lines $|P_1 - P_2|$ mean that we subtract whichever is the smaller from the larger of the proportions.) If the sample size (N) is relatively small (say less than 200) then in such calculations a *correction for continuity* may be included. Details are to be found in most standard statistical texts. If the value of 'x' exceeds 1.96 then the proportions differ significantly ($P = 0.05$). The exact level of significance of 'x' can be determined from tables of the normal distribution (Fisher & Yates, 1963). The points most commonly required for significance are given in Table 12.1.

Table 12.1 Probability (P) of deviations (x) in units of standard deviation from the mean assuming normal distribution

P	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001
x	1.28	1.65	1.96	2.33	2.58	3.09	3.29	3.89

It should be noted that $x^2 = \chi^2$ and the same results may be obtained by presenting the data in a 2×2 table and determining the significance of the χ^2 value with 1 degree of freedom. The calculation, however, is more laborious than using proportions.

The method of calculation is illustrated with data on the incidence of CNS malformations (anencephaly and spina bifida) in the Edinburgh region. For several years the incidence of these malformations had remained fairly steady at about 1 in 200 total (still and live) births. However in 1971 there appeared

to be an increase in the incidence to 1 in 120 births and the question arises as to whether this figure is significantly different from previous years. The actual figures were 50 CNS malformations out of 9706 total births in 1970, and 79 out of 9771 births in 1971. Thus:

for 1970

$$P_1 = \frac{50}{9706}$$

$$= 0.0052$$

for 1971

$$P_2 = \frac{79}{9771}$$

$$= 0.0081$$

and

$$P_0 = \frac{50 + 79}{9706 + 9771}$$

$$= 0.0066$$

and

$$x = \frac{|P_1 - P_2|}{\sqrt{P_0(1 - P_0)(1/N_1 + 1/N_2)}}$$

$$= \frac{(0.0081 - 0.0052)}{\sqrt{(0.0066)(0.9934)(1/9706 + 1/9771)}}$$

$$= 2.5$$

Thus the difference in proportions is statistically significant ($P < 0.02$). In subsequent years the incidence returned again to about 1 in 200 births and no satisfactory explanation could be found for the increase in 1971, though when, as here, many comparisons are made, 1 in 20 could differ by chance alone.

This method is only applicable if n_1 and n_2 are reasonably large (say more than 20).

Another approach is to consider an overall significance test using a $2 \times k$ contingency table and determine whether there is a significant trend in the proportions from group 1 to group k (Armitage, 1955). However the method of calculation is somewhat tedious and does not have the immediate visual appeal of the so-called cumulative sum or *cusum* techniques.

Cumulative sum techniques

These techniques (Woodward & Goldsmith, 1964) were originally developed for use in industry to demonstrate phenomena such as trends in productivity, but they can also be used to pinpoint the onset of an epidemic or an increase in the incidence of a particular congenital malformation.

The basic procedure merely consists of subtracting a previously defined 'reference value' (k) from each number in the series and accumulating the sum of the differences as each additional figure is introduced. The successive accumulated differences are referred to as the 'cumulative sums' (*cusums*) and the graph of these sums is known as the 'cumulative sum chart'. Thus if the individual numbers of cases in successive years are

$$n_1, n_2, n_3, \dots n_r$$

then

$$S_1 = (n_1 - k)$$

$$S_2 = (n_1 - k) + (n_2 - k) = S_1 + (n_2 - k)$$

$$S_3 = S_2 + (n_3 - k)$$

and

$$S_r = S_{r-1} + (n_r - k) = n_1 + n_2 + \dots n_r - rk$$

The reference value is chosen as the number around which the results are expected to vary, usually the mean value of the results at the beginning of a period of study. To simplify the calculation of cusums, ' k ' is suitably rounded off. If the average of the results is close to the reference value, some of the differences will be positive and some negative so that the cusum chart will be essentially horizontal. However, if the average begins to rise more of the differences will become positive and the cusum chart will slope upwards.

The value of the technique is illustrated in the following example. Suppose the annual incidence (say number per 10 000 births suitably rounded to the nearest whole number for convenience) of a particular congenital malformation is as given in Table 12.2. If the annual incidence is plotted there is no clear trend or change over the period of study (Fig. 12.1). However, if the cusums are calculated with $k = 20$ (Table 12.2), and plotted (Fig. 12.1) it becomes clear that the annual incidence began to rise in 1967. The average incidence during the period 1950 to 1966 was about 21. The importance of choosing a reference value close to this is illustrated in Figure 12.2. If too low a value is chosen the cusum plot increases steadily throughout, whereas if too high a value is chosen all the cusums are negative.

If the method is applied to absolute *numbers* of cases, then in an expanding population the number of cases would automatically increase, and be reflected by a change in cusums, even if the relative incidence remained the same. The method is therefore best applied to *incidence rates* (e.g. number per 10 000 births) as in the above example.

It should be noted that this method is essentially merely a graphical means of demonstrating a change. If the reader wishes to define such a change in precise terms then it is best to apply one of the more conventional statistical methods which are available for this purpose (e.g. see Armitage, 1971).

Table 12.2 Annual incidence and cusums ($k = 20$)

Year	Incidence	Difference from k	Cusums
1950	22	+2	2
1951	29	+9	11
1952	28	+8	19
1953	23	+3	22
1954	9	-11	11
1955	28	+8	19
1956	12	-8	11
1957	30	+10	21
1958	14	-6	15
1959	28	+8	23
1960	15	-5	18
1961	18	-2	16
1962	27	+7	23
1963	9	-11	12
1964	30	+10	22
1965	14	-6	16
1966	24	+4	20
1967	27	+7	27
1968	30	+10	37
1969	15	-5	32
1970	29	+9	41
1971	26	+6	47
1972	30	+10	57
1973	26	+6	63
1974	19	-1	62
1975	32	+12	74

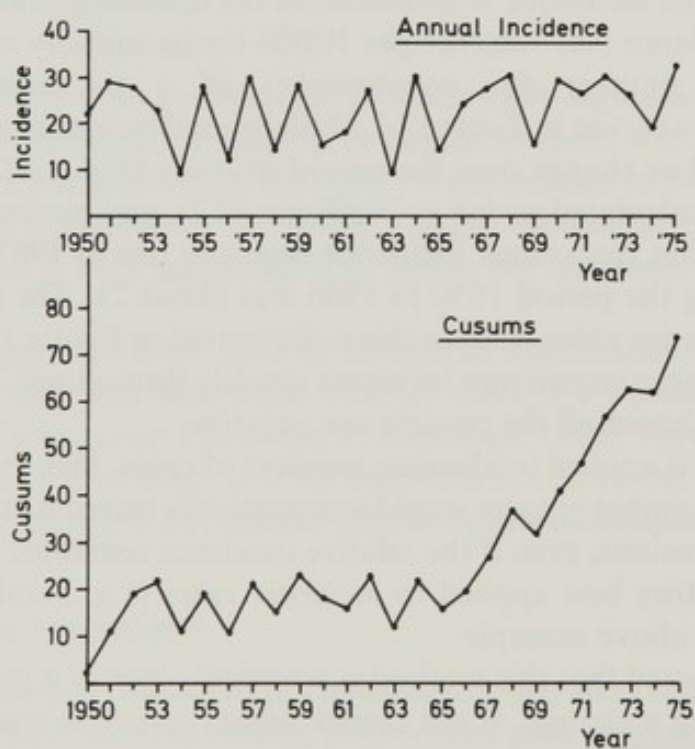


Fig. 12.1 Annual incidence. Above: standard chart. Below: cumulative sum chart. (Data from Table 12.2)

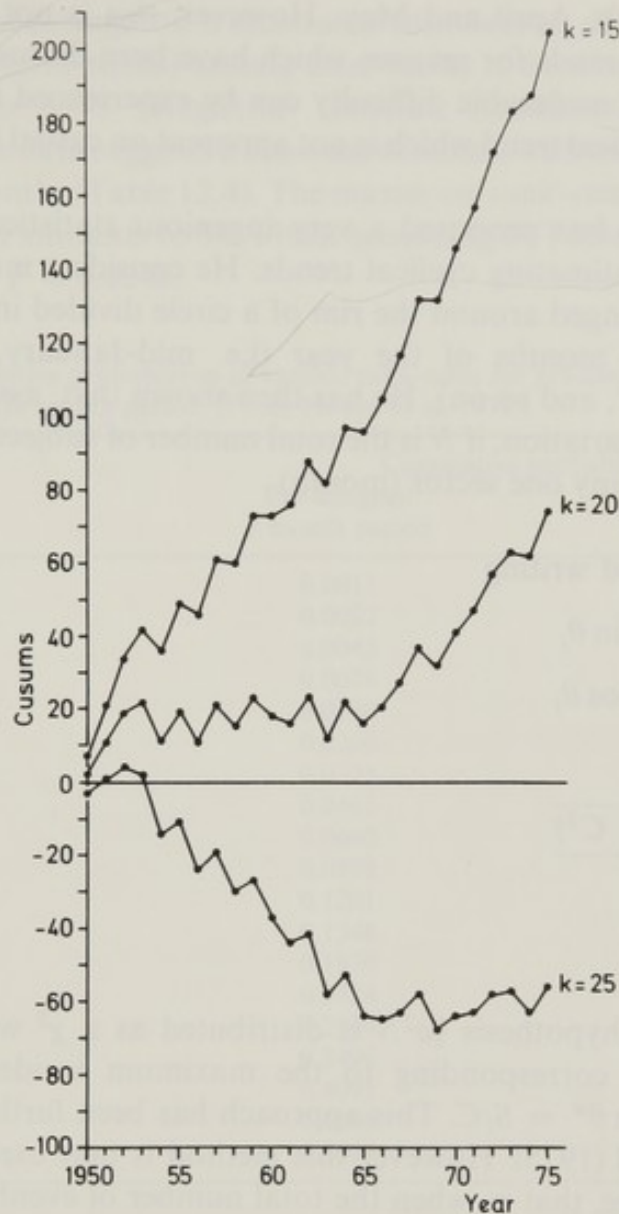


Fig. 12.2 Cumulative sum charts for various values of k . (Data from Table 12.2)

Cyclical changes

There are a number of disorders which show seasonal variations or cyclical trends in incidence; for example, hay fever is without doubt seasonal, with the highest incidence in summer months. The demonstration of a seasonal variation in incidence for a particular disorder or congenital malformation would indicate that environmental factors are involved and thus give a clue to aetiology. For this reason a number of studies in recent years have been directed to this problem. A commonly used method of demonstrating seasonal variation is to compare the observed incidence with the expected incidence using a simple χ^2 test. Using this method, for example, Nielsen et al (1975) showed that there was a significant seasonal variation in the birth of children with sex chromosomal abnormalities, with the highest incidences

occurring in March, April and May. However, this is not a good test for detecting cyclical trends for reasons which have been discussed by Edwards (1961). Further, considerable difficulty can be experienced in attempting to demonstrate a cyclical trend which is not apparent on casual inspection of the data.

Edwards (1961) has proposed a very ingenious statistical technique for recognizing and estimating cyclical trends. He considers monthly incidence data as being arranged around the rim of a circle divided into equal sectors corresponding to months of the year (i.e. mid-January, $\theta = 15^\circ$; mid-February, $\theta = 45^\circ$; and so on). He has then shown that, assuming there is a simple harmonic variation, if N is the total number of subjects in a study with N_i the number in any one sector (month)

i.e. $N = \sum N_i$, and writing

$$S = \sum \sqrt{N_i} \sin \theta_i$$

$$C = \sum \sqrt{N_i} \cos \theta_i$$

$$W = \sum \sqrt{N_i}$$

$$d = \frac{\sqrt{(S^2 + C^2)}}{W}$$

$$a = 4d$$

then on the null hypothesis $\frac{1}{2}a^2N$ is distributed as a χ^2 with 2 degrees of freedom, and θ^* corresponding to the maximum incidence is equal to $\tan^{-1} S/C$, i.e. $\tan \theta^* = S/C$. This approach has been further developed by Walter & Elwood (1975). However this method is best used only when the sample size is large, that is, when the total number of events (e.g. congenital malformations) exceeds 50.

A much simpler ranking (non-parametric) method has been introduced by Hewitt et al (1971) which can also be used for sample sizes less than 50 provided that at least 6 of the 12 months have non-zero frequencies. The method first consists of ranking the incidence rates for each month, the highest incidence as 12 and the lowest as unity. The next step is to decide if there is a prior hypothesis for specifying a six month period of higher expected incidences, or if the likely nature of any seasonal variation has to be inferred from the data. In the latter case when no prior hypothesis exists it is necessary to determine the six month period which yields the highest value of the rank-sum. With a pre-assigned six month period a rank-sum equal to or greater than 50 would be significant whereas for *any* six month period a rank-sum equal to or greater than 55 is required for significance (Table 12.3).

A six month period is chosen for these calculations because the chance probability of obtaining the largest possible rank-sum is smallest for this period (Hewitt et al 1971).

The method of calculation is illustrated with data on the monthly incidence of anencephaly (still births) among total births in Scotland for the five year period 1969 to 1973 (Registrar General, Scotland. *Annual Reports*). Inspection of the data suggests a seasonal variation with most births occurring in the winter months (Table 12.4). The maximum rank-sum is from September to February and amounts to 56, which according to Table 12.3 is statistically significant with $P = 0.0248$.

Table 12.3 Cumulative probabilities of various rank-sums for pre-assigned six month periods or for any six month period. (From Hewitt et al, 1971.)

Rank-sum	Cumulative probabilities	
	Pre-assigned 6 month period	Any 6 month period
57	0.0011	0.0134
56	0.0022	0.0248
55	0.0043	0.0464
54	0.0076	0.0766
53	0.0130	0.1260
52	0.0206	0.1914
51	0.0325	0.2908
50	0.0465	0.3826
49	0.0660	0.4958
48	0.0898	0.6086
47	0.1201	0.7258
46	0.1548	0.8310
45	0.1970	0.9138
44	0.2424	0.9614
43	0.2944	0.9904
42	0.3496	0.9986
41	0.4091	1.0000
40	0.4686	—
39 or less	1.0000	—

Table 12.4 Seasonal variation in the incidence of anencephaly in Scotland from 1969 to 1973 inclusive

Month of birth	Total births	Anencephalics		Rank	Maximum rank
		No.	Incidence/1000		
January	34 110	94	2.76	10	10
February	30 840	93	3.02	12	12
March	35 400	69	1.95	3	—
April	32 548	79	2.43	7	—
May	33 635	64	1.90	2	—
June	32 523	71	2.18	5	—
July	33 037	67	2.03	4	—
August	32 115	56	1.74	1	—
September	30 791	79	2.57	9	9
October	33 365	83	2.49	8	8
November	29 565	82	2.77	11	11
December	31 304	74	2.36	6	6
				Rank-sum	56

When data can be grouped (as in the case of monthly birth incidences) and when there is no particular reason to expect a specific parametric alternative (such as a sinusoidal curve of period 12 months) then Freedman (1979) has proposed the use of a Kolmogorov-Smirnov type statistic to test for seasonal variation. If

N = total number of subjects in a study

t = time of the occurrence of the event in question measured from the beginning of the year. If the study covers leap years then an 'average' year consists of $365\frac{1}{4}$ days with February having $28\frac{1}{4}$ days. Thus for an event in March, $t = 31 + 28\frac{1}{4} + 31 = 90\frac{1}{4}$

j = cumulative number of events over the year(s)

$$F_N = j/N$$

$$F_t = t/365.25$$

then for each month the difference ($F_N - F_t$) is obtained, and the sum of the maximum difference and the *absolute* minimum difference is denoted as V_N . This is a Kolmogorov-Smirnov type statistic (see Siegel, 1956) and estimated percentiles for the distribution of $V_N\sqrt{N}$ are given in Table 12.5, though a slightly more sensitive statistic is also available (Freedman, 1981).

Applying this approach to the data on the monthly birth incidence of anencephaly given in Table 12.4, the calculations required to calculate V_N are set out in Table 12.6.

The maximum value of ($F_N - F_t$) is 0.043 in February and -0.015 in August. Therefore

$$\begin{aligned} V_N &= 0.043 + 0.015 \\ &= 0.058 \end{aligned}$$

$$\begin{aligned} \text{and } V_N\sqrt{N} &= 0.058\sqrt{911} \\ &= 1.751 \end{aligned}$$

and from table 12.5, $P < 0.01$.

This non-parametric test is a little more complicated than Hewitt's test but has the advantage of being more powerful.

The relative merits of the various tests devised by Edwards (1961), Walter & Elwood (1975) and Hewitt et al (1971) for studying cyclical changes are discussed by Walter & Elwood (1975) and Walter (1977). In general, whenever the sample size is small ($N < 50$) then a non-parametric method is preferable. But his method will only detect a fairly marked and consistent seasonal variation. A parametric method is preferable when there is a substantial amount of data.

It should always be borne in mind, of course, that in studying a particular disorder or congenital malformation the demonstration of a significant change in incidence, which may or may not be cyclical, is not an end in itself but merely the first step in attempting to identify possible aetiological factors.

Table 12.5 Estimated percentiles of the distribution of $V_N \sqrt{N}$. (From Freedman, 1979.)

Percentile	Estimate	90% confidence limits
10%	0.58	0.576-0.589
20%	0.67	0.670-0.680
30%	0.75	0.744-0.754
40%	0.82	0.811-0.822
50%	0.89	0.882-0.893
60%	0.96	0.950-0.961
70%	1.03	1.028-1.040
80%	1.14	1.128-1.145
85%	1.21	1.199-1.217
90%	1.29	1.280-1.297
95%	1.41	1.400-1.422
99%	1.66	1.641-1.683

Table 12.6 Calculations required to calculate V_N for grouped data

Month	t	Cumulative No. (j)	F_N (j/N)	F_t ($t/365.25$)	$F_N - F_t$
Jan	31	94	0.103	0.085	0.018
Feb	59.25	187	0.205	0.162	0.043
Mar	90.25	256	0.281	0.247	0.034
Apr	120.25	335	0.368	0.329	0.039
May	151.25	399	0.438	0.414	0.024
June	181.25	470	0.516	0.496	0.020
Jul	212.25	537	0.589	0.581	0.008
Aug	243.25	593	0.651	0.666	-0.015
Sep	273.25	672	0.738	0.748	-0.010
Oct	304.25	755	0.829	0.833	-0.004
Nov	334.25	837	0.919	0.915	0.004
Dec	365.25	911 (N)	1.000	1.000	0

maximum

minimum

Appendices

1. Student's t distribution
2. χ^2 distribution
3. Correlation coefficient
4. Transformation of r to z
5. Normal distribution for estimation of h^2
6. Lod scores

Appendices 1, 2 and 3 are from N. T. J. Bailey (1969) *Statistical Methods in Biology*, English Universities Press, London; Appendix 4 from R. R. Sokal and F. James Rohlf (1973) *Introduction to Biostatistics*, W. H. Freeman & Company © 1973; Appendix 5 from D. S. Falconer (1965) *Annals of Human Genetics (London)* 29:51–76; and Appendix 6 calculated from C. A. B. Smith (1968) *Annals of Human Genetics (London)* 32:127–150, with some modifications, and to which values for family sizes greater than 7 have been added.

Appendix 1: Student's *t* distribution

The table gives the percentage points most frequently required for significance tests and confidence limits based on student's *t* distribution. Thus the probability of observing a value of *t*, with 10 degrees of freedom, greater in *absolute value* than 3.169 (i.e. < -3.169 or $> +3.169$) is exactly 0.01 or 1%.

Degrees of freedom	Value of <i>P</i>					
	0.10	0.05	0.02	0.01	0.002	0.001
1	6.314	12.71	31.82	63.66	318.3	636.6
2	2.920	4.303	6.965	9.925	22.33	31.60
3	2.353	3.182	4.541	5.841	10.21	12.92
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.767
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646

Appendix 2: χ^2 distribution

The table gives the percentage points most frequently required for significance tests based on χ^2 . Thus the probability of observing a χ^2 with 5 degrees of freedom *greater* in value than 11.07 is 0.05 or 5%. Again, the probability of observing a χ^2 with 5 degrees of freedom *smaller* in value than 0.554 is $1 - 0.99 = 0.01$ or 1%.

Degrees of freedom	Value of <i>P</i>				
	0.99	0.95	0.05	0.01	0.001
1	0.000 157	0.003 93	3.841	6.635	10.83
2	0.0201	0.103	5.991	9.210	13.82
3	0.115	0.352	7.815	11.34	16.27
4	0.297	0.711	9.488	13.28	18.47
5	0.554	1.145	11.07	15.09	20.51
6	0.872	1.635	12.59	16.81	22.46
7	1.239	2.167	14.07	18.48	24.32
8	1.646	2.733	15.51	20.09	26.13
9	2.088	3.325	16.92	21.67	27.88
10	2.558	3.940	18.31	23.21	29.59
11	3.053	4.575	19.68	24.72	31.26
12	3.571	5.226	21.03	26.22	32.91
13	4.107	5.892	22.36	27.69	34.53
14	4.660	6.571	23.68	29.14	36.12
15	5.229	7.261	25.00	30.58	37.70
16	5.812	7.962	26.30	32.00	39.25
17	6.408	8.672	27.59	33.41	40.79
18	7.015	9.390	28.87	34.81	42.31
19	7.633	10.12	30.14	36.19	43.82
20	8.260	10.85	31.41	37.57	45.31
21	8.897	11.59	32.67	38.93	46.80
22	9.542	12.34	33.92	40.29	48.27
23	10.20	13.09	35.17	41.64	49.73
24	10.86	13.85	36.42	42.98	51.18
25	11.52	14.61	37.65	44.31	52.62
26	12.20	15.38	38.89	45.64	54.05
27	12.88	16.15	40.11	46.96	55.48
28	13.56	16.93	41.34	48.28	56.89
29	14.26	17.71	42.56	49.59	58.30
30	14.95	18.49	43.77	50.89	59.70

Appendix 3: Correlation coefficient

The table gives percentage points for the distribution of the estimated correlation coefficient r . Thus when there are 10 degrees of freedom (i.e. in samples of 12) the probability of observing an r greater in *absolute value* than 0.576 (i.e. < -0.576 or $> +0.576$) is 0.05 or 5%.

Degrees of freedom	Value of P				
	0.10	0.05	0.02	0.01	0.001
1	0.9877	0.996 92	0.999 51	0.999 88	0.999 998 8
2	0.9000	0.9500	0.9800	0.9900	0.9990
3	0.805	0.878	0.9343	0.9587	0.9911
4	0.729	0.811	0.882	0.9172	0.9741
5	0.669	0.754	0.833	0.875	0.9509
6	0.621	0.707	0.789	0.834	0.9249
7	0.582	0.666	0.750	0.798	0.898
8	0.549	0.632	0.715	0.765	0.872
9	0.521	0.602	0.685	0.735	0.847
10	0.497	0.576	0.658	0.708	0.823
11	0.476	0.553	0.634	0.684	0.801
12	0.457	0.532	0.612	0.661	0.780
13	0.441	0.514	0.592	0.641	0.760
14	0.426	0.497	0.574	0.623	0.742
15	0.412	0.482	0.558	0.606	0.725
16	0.400	0.468	0.543	0.590	0.708
17	0.389	0.456	0.529	0.575	0.693
18	0.378	0.444	0.516	0.561	0.679
19	0.369	0.433	0.503	0.549	0.665
20	0.360	0.423	0.492	0.537	0.652
25	0.323	0.381	0.445	0.487	0.597
30	0.296	0.349	0.409	0.449	0.554
35	0.275	0.325	0.381	0.418	0.519
40	0.257	0.304	0.358	0.393	0.490
45	0.243	0.288	0.338	0.372	0.465
50	0.231	0.273	0.322	0.354	0.443
60	0.211	0.250	0.295	0.325	0.408
70	0.195	0.232	0.274	0.302	0.380
80	0.183	0.217	0.257	0.283	0.357
90	0.173	0.205	0.242	0.267	0.338
100	0.164	0.195	0.230	0.254	0.321

Appendix 4: The z-transformation of correlation coefficient r

r	z	r	z
0.00	0.0000	0.35	0.3654
0.01	0.0100	0.36	0.3769
0.02	0.0200	0.37	0.3884
0.03	0.0300	0.38	0.4001
0.04	0.0400	0.39	0.4118
0.05	0.0500	0.40	0.4236
0.06	0.0601	0.41	0.4356
0.07	0.0701	0.42	0.4477
0.08	0.0802	0.43	0.4599
0.09	0.0902	0.44	0.4722
0.10	0.1003	0.45	0.4847
0.11	0.1104	0.46	0.4973
0.12	0.1206	0.47	0.5101
0.13	0.1307	0.48	0.5230
0.14	0.1409	0.49	0.5361
0.15	0.1511	0.50	0.5493
0.16	0.1614	0.51	0.5627
0.17	0.1717	0.52	0.5763
0.18	0.1820	0.53	0.5901
0.19	0.1923	0.54	0.6042
0.20	0.2027	0.55	0.6184
0.21	0.2132	0.56	0.6328
0.22	0.2237	0.57	0.6475
0.23	0.2342	0.58	0.6625
0.24	0.2448	0.59	0.6777
0.25	0.2554	0.60	0.6931
0.26	0.2661	0.61	0.7089
0.27	0.2769	0.62	0.7250
0.28	0.2877	0.63	0.7414
0.29	0.2986	0.64	0.7582
0.30	0.3095	0.65	0.7753
0.31	0.3205	0.66	0.7928
0.32	0.3316	0.67	0.8107
0.33	0.3428	0.68	0.8291
0.34	0.3541	0.69	0.8480

Appendix 4—continued

<i>r</i>	<i>z</i>	<i>r</i>	<i>z</i>
0.70	0.8673	0.85	1.2562
0.71	0.8872	0.86	1.2933
0.72	0.9076	0.87	1.3331
0.73	0.9287	0.88	1.3758
0.74	0.9505	0.89	1.4219
0.75	0.9730	0.90	1.4722
0.76	0.9962	0.91	1.5275
0.77	1.0203	0.92	1.5890
0.78	1.0454	0.93	1.6584
0.79	1.0714	0.94	1.7380
0.80	1.0986	0.95	1.8318
0.81	1.1270	0.96	1.9459
0.82	1.1568	0.97	2.0923
0.83	1.1881	0.98	2.2976
0.84	1.2212	0.99	2.6467

Appendix 5: Normal distribution for estimation of h^2 Table of x and a for values of q from $q = 0.01\%$ to $q = 30.0\%$. q is the incidence; x is the normal deviate (single-tailed) exceeded by the proportion q ; a is the mean deviation of these individuals. Note changes of interval in q at $q = 2.0\%$ and $q = 21.0\%$.

$q\%$	x	a	$q\%$	x	a	$q\%$	x	a	$q\%$	x	a
0.01	3.719	3.960	0.40	2.652	2.962	0.80	2.409	2.740	1.20	2.257	2.603
0.02	3.540	3.790	0.41	2.644	2.954	0.81	2.404	2.736	1.21	2.254	2.600
0.03	3.432	3.687	0.42	2.636	2.947	0.82	2.400	2.732	1.22	2.251	2.597
0.04	3.353	3.613	0.43	2.628	2.939	0.83	2.395	2.728	1.23	2.248	2.594
0.05	3.291	3.554	0.44	2.620	2.932	0.84	2.391	2.724	1.24	2.244	2.591
0.06	3.239	3.507	0.45	2.612	2.925	0.85	2.387	2.720	1.25	2.241	2.589
0.07	3.195	3.464	0.46	2.605	2.918	0.86	2.382	2.716	1.26	2.238	2.586
0.08	3.156	3.429	0.47	2.597	2.911	0.87	2.378	2.712	1.27	2.235	2.583
0.09	3.121	3.397	0.48	2.590	2.905	0.88	2.374	2.708	1.28	2.232	2.580
0.10	3.090	3.367	0.49	2.583	2.898	0.89	2.370	2.704	1.29	2.229	2.578
0.11	3.062	3.341	0.50	2.576	2.892	0.90	2.366	2.701	1.30	2.226	2.575
0.12	3.036	3.317	0.51	2.569	2.886	0.91	2.361	2.697	1.31	2.223	2.572
0.13	3.012	3.294	0.52	2.562	2.880	0.92	2.357	2.693	1.32	2.220	2.570
0.14	2.989	3.273	0.53	2.556	2.873	0.93	2.353	2.690	1.33	2.217	2.567
0.15	2.968	3.253	0.54	2.549	2.868	0.94	2.349	2.686	1.34	2.214	2.564
0.16	2.948	3.234	0.55	2.543	2.862	0.95	2.346	2.683	1.35	2.211	2.562
0.17	2.929	3.217	0.56	2.536	2.856	0.96	2.342	2.679	1.36	2.209	2.559
0.18	2.911	3.201	0.57	2.530	2.850	0.97	2.338	2.676	1.37	2.206	2.557
0.19	2.894	3.185	0.58	2.524	2.845	0.98	2.334	2.672	1.38	2.203	2.554
			0.59	2.518	2.839	0.99	2.330	2.669	1.39	2.200	2.552

Appendix 5—continued

q %	x	a	q %	x	a	q %	x	a	q %	x	a
0.20	2.878	3.170	0.60	2.512	2.834	1.00	2.326	2.665	1.40	2.197	2.549
0.21	2.863	3.156	0.61	2.506	2.829	1.01	2.323	2.662	1.41	2.194	2.547
0.22	2.848	3.142	0.62	2.501	2.823	1.02	2.319	2.658	1.42	2.192	2.544
0.23	2.834	3.129	0.63	2.495	2.818	1.03	2.315	2.655	1.43	2.189	2.542
0.24	2.820	3.117	0.64	2.489	2.813	1.04	2.312	2.652	1.44	2.186	2.539
0.25	2.807	3.104	0.65	2.484	2.808	1.05	2.308	2.649	1.45	2.183	2.537
0.26	2.794	3.093	0.66	2.478	2.803	1.06	2.304	2.645	1.46	2.181	2.534
0.27	2.782	3.081	0.67	2.473	2.798	1.07	2.301	2.642	1.47	2.178	2.532
0.28	2.770	3.070	0.68	2.468	2.793	1.08	2.297	2.639	1.48	2.175	2.529
0.29	2.759	3.060	0.69	2.462	2.789	1.09	2.294	2.636	1.49	2.173	2.527
0.30	2.748	3.050	0.70	2.457	2.784	1.10	2.290	2.633	1.50	2.170	2.525
0.31	2.737	3.040	0.71	2.452	2.779	1.11	2.287	2.630	1.51	2.167	2.522
0.32	2.727	3.030	0.72	2.447	2.775	1.12	2.283	2.627	1.52	2.165	2.520
0.33	2.716	3.021	0.73	2.442	2.770	1.13	2.280	2.624	1.53	2.162	2.518
0.34	2.706	3.012	0.74	2.437	2.766	1.14	2.277	2.621	1.54	2.160	2.515
0.35	2.697	3.003	0.75	2.432	2.761	1.15	2.273	2.618	1.55	2.157	2.513
0.36	2.687	2.994	0.76	2.428	2.757	1.16	2.270	2.615	1.56	2.155	2.511
0.37	2.678	2.986	0.77	2.423	2.753	1.17	2.267	2.612	1.57	2.152	2.508
0.38	2.669	2.978	0.78	2.418	2.748	1.18	2.264	2.609	1.58	2.149	2.506
0.39	2.661	2.969	0.79	2.414	2.744	1.19	2.260	2.606	1.59	2.147	2.504

Appendix 5: Normal distribution for estimation of h^2 —continued

$q\%$	x	a	$q\%$	x	a	$q\%$	x	a	$q\%$	x	a
1.60	2.144	2.502	2.0	2.054	2.421	6.0	1.555	1.985	10.0	1.282	1.755
1.61	2.142	2.499	2.1	2.034	2.403	6.1	1.546	1.978	10.1	1.276	1.750
1.62	2.139	2.497	2.2	2.014	2.386	6.2	1.538	1.971	10.2	1.270	1.746
1.63	2.137	2.495	2.3	1.995	2.369	6.3	1.530	1.964	10.3	1.265	1.741
1.64	2.135	2.493	2.4	1.977	2.353	6.4	1.522	1.957	10.4	1.259	1.736
1.65	2.132	2.491	2.5	1.960	2.338	6.5	1.514	1.951	10.5	1.254	1.732
1.66	2.130	2.489	2.6	1.943	2.323	6.6	1.506	1.944	10.6	1.248	1.727
1.67	2.127	2.486	2.7	1.927	2.309	6.7	1.499	1.937	10.7	1.243	1.723
1.68	2.125	2.484	2.8	1.911	2.295	6.8	1.491	1.931	10.8	1.237	1.718
1.69	2.122	2.482	2.9	1.896	2.281	6.9	1.483	1.924	10.9	1.232	1.714
1.70	2.120	2.480	3.0	1.881	2.268	7.0	1.476	1.918	11.0	1.227	1.709
1.71	2.118	2.478	3.1	1.866	2.255	7.1	1.468	1.912	11.1	1.221	1.705
1.72	2.115	2.476	3.2	1.852	2.243	7.2	1.461	1.906	11.2	1.216	1.701
1.73	2.113	2.474	3.3	1.838	2.231	7.3	1.454	1.899	11.3	1.211	1.696
1.74	2.111	2.472	3.4	1.825	2.219	7.4	1.447	1.893	11.4	1.206	1.692
1.75	2.108	2.470	3.5	1.812	2.208	7.5	1.440	1.887	11.5	1.200	1.688
1.76	2.106	2.467	3.6	1.799	2.197	7.6	1.433	1.881	11.6	1.195	1.684
1.77	2.104	2.465	3.7	1.787	2.186	7.7	1.426	1.876	11.7	1.190	1.679
1.78	2.101	2.463	3.8	1.774	2.175	7.8	1.419	1.870	11.8	1.185	1.675
1.79	2.099	2.461	3.9	1.762	2.165	7.9	1.412	1.864	11.9	1.180	1.671

Appendix 5—continued

q%	x	a	q%	x	a	q%	x	a	q%	x	a
1.80	2.097	2.459	4.0	1.751	2.154	8.0	1.405	1.858	12.0	1.175	1.667
1.81	2.095	2.457	4.1	1.739	2.144	8.1	1.398	1.853	12.1	1.170	1.663
1.82	2.092	2.455	4.2	1.728	2.135	8.2	1.392	1.847	12.2	1.165	1.659
1.83	2.090	2.453	4.3	1.717	2.125	8.3	1.385	1.842	12.3	1.160	1.655
1.84	2.088	2.451	4.4	1.706	2.116	8.4	1.379	1.836	12.4	1.155	1.651
1.85	2.086	2.449	4.5	1.695	2.106	8.5	1.372	1.831	12.5	1.150	1.647
1.86	2.084	2.447	4.6	1.685	2.097	8.6	1.366	1.825	12.6	1.146	1.643
1.87	2.081	2.445	4.7	1.675	2.088	8.7	1.359	1.820	12.7	1.141	1.639
1.88	2.079	2.444	4.8	1.665	2.080	8.8	1.353	1.815	12.8	1.136	1.635
1.89	2.077	2.442	4.9	1.655	2.071	8.9	1.347	1.810	12.9	1.131	1.631
1.90	2.075	2.440	5.0	1.645	2.063	9.0	1.341	1.804	13.0	1.126	1.627
1.91	2.073	2.438	5.1	1.635	2.054	9.1	1.335	1.799	13.1	1.122	1.623
1.92	2.071	2.436	5.2	1.626	2.046	9.2	1.329	1.794	13.2	1.117	1.620
1.93	2.068	2.434	5.3	1.616	2.038	9.3	1.323	1.789	13.3	1.112	1.616
1.94	2.066	2.432	5.4	1.607	2.030	9.4	1.317	1.784	13.4	1.108	1.612
1.95	2.064	2.430	5.5	1.598	2.023	9.5	1.311	1.779	13.5	1.103	1.608
1.96	2.062	2.428	5.6	1.589	2.015	9.6	1.305	1.774	13.6	1.098	1.605
1.97	2.060	2.426	5.7	1.580	2.007	9.7	1.299	1.769	13.7	1.094	1.601
1.98	2.058	2.425	5.8	1.572	2.000	9.8	1.293	1.765	13.8	1.089	1.597
1.99	2.056	2.423	5.9	1.563	1.993	9.9	1.287	1.760	13.9	1.085	1.593

Appendix 5: Normal distribution for estimation of h^2 —continued

$q\%$	x	a	$q\%$	x	a	$q\%$	x	a	$q\%$	x	a
14.0	1.080	1.590	16.0	0.994	1.521	18.0	0.915	1.458	20.0	0.842	1.400
14.1	1.076	1.586	16.1	0.990	1.517	18.1	0.912	1.455	20.1	0.838	1.397
14.2	1.071	1.583	16.2	0.986	1.514	18.2	0.908	1.452	20.2	0.834	1.394
14.3	1.067	1.579	16.3	0.982	1.511	18.3	0.904	1.449	20.3	0.831	1.391
14.4	1.063	1.575	16.4	0.978	1.508	18.4	0.900	1.446	20.4	0.827	1.389
14.5	1.058	1.572	16.5	0.974	1.504	18.5	0.896	1.443	20.5	0.824	1.386
14.6	1.054	1.568	16.6	0.970	1.501	18.6	0.893	1.440	20.6	0.820	1.383
14.7	1.049	1.565	16.7	0.966	1.498	18.7	0.889	1.437	20.7	0.817	1.381
14.8	1.045	1.561	16.8	0.962	1.495	18.8	0.885	1.434	20.8	0.813	1.378
14.9	1.041	1.558	16.9	0.958	1.492	18.9	0.882	1.431	20.9	0.810	1.375
15.0	1.036	1.554	17.0	0.954	1.489	19.0	0.878	1.428	21.0	0.806	1.372
15.1	1.032	1.551	17.1	0.950	1.485	19.1	0.874	1.425	22.0	0.772	1.346
15.2	1.028	1.548	17.2	0.946	1.482	19.2	0.871	1.422	23.0	0.739	1.320
15.3	1.024	1.544	17.3	0.942	1.479	19.3	0.867	1.420	24.0	0.706	1.295
15.4	1.019	1.541	17.4	0.938	1.476	19.4	0.863	1.417	25.0	0.674	1.271
15.5	1.015	1.537	17.5	0.935	1.473	19.5	0.860	1.414	26.0	0.643	1.248
15.6	1.011	1.534	17.6	0.931	1.470	19.6	0.856	1.411	27.0	0.613	1.225
15.7	1.007	1.531	17.7	0.927	1.467	19.7	0.852	1.408	28.0	0.583	1.202
15.8	1.003	1.527	17.8	0.923	1.464	19.8	0.849	1.405	29.0	0.553	1.180
15.9	0.999	1.524	17.9	0.919	1.461	19.9	0.845	1.403	30.0	0.524	1.159

Appendix 6: Linkage (lod) scores for families with up to ten children

No.	Scored children count	Recombination fraction (θ)										
		0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
1.	1 $n-r$	0 r	0.297	0.279	0.255	0.230	0.204	0.176	0.146	0.114	0.079	0.041
	0 $n-r$	1 r	-1.699	-1.000	-0.699	-0.523	-0.398	-0.301	-0.222	-0.155	-0.097	-0.046
	2 $n-r$	0 r	0.593	0.558	0.511	0.461	0.408	0.352	0.292	0.228	0.158	0.083
2.	1 $n-r$	1 r	-1.402	-0.721	-0.444	-0.292	-0.194	-0.125	-0.076	-0.041	-0.018	-0.004
	0 $n-r$	2 r	-3.398	-2.000	-1.398	-1.046	-0.796	-0.602	-0.444	-0.310	-0.194	-0.092
	z_1	2:0	0.292	0.258	0.215	0.173	0.134	0.097	0.064	0.037	0.017	0.004
z_1	2:0	0.460	0.395	0.319	0.250	0.190	0.135	0.088	0.050	0.023	0.005	
z_1	2:0	0.171	0.154	0.131	0.107	0.085	0.062	0.041	0.024	0.011	0.003	
z_1	1:1	-1.402	-0.721	-0.444	-0.292	-0.194	-0.125	-0.076	-0.041	-0.018	-0.004	
z_1	1:1	-1.234	-0.584	-0.340	-0.215	-0.138	-0.087	-0.052	-0.028	-0.012	-0.003	
z_1	1:1	-1.523	-0.825	-0.528	-0.358	-0.243	-0.160	-0.099	-0.054	-0.024	-0.006	
3.	3 $n-r$	0 r	0.890	0.836	0.766	0.691	0.612	0.528	0.438	0.342	0.238	0.124
	2 $n-r$	1 r	-1.106	-0.442	-0.188	-0.062	0.010	0.051	0.070	0.073	0.061	0.037
	1 $n-r$	2 r	-3.101	-1.721	-1.143	-0.815	-0.592	-0.426	-0.298	-0.196	-0.115	-0.050
	0 $n-r$	3 r	-5.097	-3.000	-2.097	-1.569	-1.194	-0.903	-0.666	-0.465	-0.291	-0.137
	z_1	3:0	0.589	0.535	0.465	0.393	0.318	0.243	0.170	0.104	0.049	0.013
z_1	3:0	0.819	0.720	0.605	0.495	0.391	0.292	0.201	0.121	0.057	0.015	
z_1	3:0	0.533	0.487	0.427	0.364	0.296	0.228	0.160	0.098	0.047	0.013	
z_1	2:1	-1.402	-0.721	-0.444	-0.292	-0.194	-0.125	-0.076	-0.041	-0.018	-0.004	
z_1	2:1	-1.172	-0.536	-0.305	-0.190	-0.121	-0.076	-0.045	-0.024	-0.010	-0.002	
z_1	2:1	-1.458	-0.769	-0.482	-0.321	-0.216	-0.140	-0.086	-0.047	-0.020	-0.004	

Appendix 6—continued

No.	Scored children count	Recombination fraction (θ)	Recombination fraction (θ)											
			0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45		
	z_1 5:0	e_1 4:1	1.170	1.088	0.975	0.855	0.725	0.585	0.439	0.290	0.150	0.042		
	z_1 5:0	e_1 3:2	1.168	1.080	0.963	0.841	0.712	0.575	0.432	0.286	0.148	0.042		
	z_1 4:1		-0.813	-0.186	0.022	0.100	0.124	0.118	0.095	0.063	0.031	0.008		
	z_1 4:1	e_1 5:0	-0.547	0.013	0.160	0.195	0.182	0.159	0.120	0.076	0.037	0.009		
	z_1 4:1	e_1 4:1	-0.825	-0.191	0.022	0.104	0.129	0.122	0.098	0.065	0.032	0.008		
	z_1 4:1	e_1 3:2	-0.827	-0.199	0.010	0.090	0.116	0.112	0.091	0.061	0.030	0.008		
	z_1 3:2		-2.805	-1.442	-0.887	-0.585	-0.388	-0.250	-0.151	-0.082	-0.035	-0.009		
	z_1 3:2	e_1 5:0	-2.539	-1.243	-0.749	-0.490	-0.324	-0.209	-0.126	-0.069	-0.029	-0.008		
	z_1 3:2	e_1 4:1	-2.817	-1.447	-0.887	-0.581	-0.383	-0.246	-0.148	-0.080	-0.034	-0.009		
	z_1 3:2	e_1 3:2	-2.819	-1.455	-0.899	-0.595	-0.396	-0.256	-0.155	-0.084	-0.036	-0.009		
6.	$6n-r$	$0r$	1.780	1.673	1.532	1.383	1.225	1.057	0.877	0.684	0.475	0.248		
	$5n-r$	$1r$	-0.216	0.394	0.577	0.629	0.623	0.579	0.509	0.415	0.299	0.161		
	$4n-r$	$2r$	-2.211	-0.885	-0.377	-0.124	0.021	0.102	0.141	0.146	0.123	0.074		
	$3n-r$	$3r$	-4.207	-2.164	-1.331	-0.877	-0.581	-0.375	-0.227	-0.123	-0.053	-0.013		
	$2n-r$	$4r$	-6.203	-3.442	-2.285	-1.631	-1.184	-0.852	-0.595	-0.392	-0.229	-0.100		
	$1n-r$	$5r$	-8.198	-4.721	-3.240	-2.384	-1.786	-1.329	-0.963	-0.661	-0.405	-0.187		
	$0n-r$	$6r$	-10.194	-6.000	-4.194	-3.137	-2.388	-1.806	-1.331	-0.929	-0.581	-0.275		
	z_1 6:0		1.479	1.371	1.231	1.082	0.924	0.756	0.578	0.393	0.211	0.061		
	z_1 6:0	e_1 6:0	1.748	1.563	1.358	1.166	0.978	0.790	0.598	0.403	0.215	0.062		
	z_1 6:0	e_1 5:1	1.474	1.373	1.237	1.090	0.932	0.762	0.583	0.396	0.212	0.061		
	z_1 6:0	e_1 4:2	1.472	1.365	1.226	1.078	0.921	0.754	0.577	0.393	0.211	0.061		
	z_1 6:0	e_1 3:3	1.472	1.364	1.224	1.076	0.919	0.752	0.575	0.391	0.211	0.061		
	z_1 5:1		-0.517	0.093	0.276	0.329	0.323	0.284	0.222	0.149	0.076	0.021		
	z_1 5:1	e_1 6:0	-0.248	0.285	0.403	0.413	0.377	0.318	0.242	0.159	0.080	0.022		
	z_1 5:1	e_1 5:1	-0.522	0.095	0.283	0.337	0.331	0.290	0.227	0.152	0.077	0.021		
	z_1 5:1	e_1 4:2	-0.524	0.087	0.271	0.325	0.320	0.282	0.221	0.149	0.076	0.021		
	z_1 5:1	e_1 3:3	-0.524	0.086	0.269	0.323	0.318	0.280	0.219	0.147	0.076	0.021		
	z_1 4:2		-2.512	-1.185	-0.673	-0.412	-0.254	-0.153	-0.087	-0.044	-0.018	-0.004		

Appendix 6: Linkage (lod) scores for families with up to ten children—continued

No.	Scored children count		Recombination fraction (θ)																	
			0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45								
	z_1	4:2	e_1	6:0	—	0.546	—	0.328	—	0.200	—	0.119	—	0.067	—	0.034	—	0.014	—	0.003
	z_1	4:2	e_1	5:1	—	0.667	—	0.404	—	0.246	—	0.147	—	0.082	—	0.041	—	0.017	—	0.004
	z_1	4:2	e_1	4:2	—	2.519	—	1.191	—	0.678	—	0.155	—	0.088	—	0.044	—	0.018	—	0.004
	z_1	4:2	e_1	3:3	—	2.519	—	1.192	—	0.680	—	0.157	—	0.090	—	0.046	—	0.018	—	0.004
	z_1	3:3	e_1	3:3	—	4.207	—	2.164	—	1.331	—	0.375	—	0.227	—	0.123	—	0.053	—	0.013
	z_1	3:3	e_1	6:0	—	3.938	—	1.972	—	1.204	—	0.341	—	0.207	—	0.113	—	0.049	—	0.012
	z_1	3:3	e_1	5:1	—	4.212	—	2.162	—	1.325	—	0.369	—	0.222	—	0.120	—	0.052	—	0.013
	z_1	3:3	e_1	4:2	—	4.214	—	2.170	—	1.336	—	0.377	—	0.228	—	0.123	—	0.053	—	0.013
	z_1	3:3	e_1	3:3	—	4.214	—	2.171	—	1.338	—	0.397	—	0.230	—	0.125	—	0.053	—	0.013
7.		$7n-r$		0 r	—	1.787	—	1.613	—	1.429	—	1.233	—	1.023	—	0.798	—	0.554	—	0.290
		$6n-r$		1 r	—	0.081	—	0.860	—	0.827	—	0.756	—	0.655	—	0.529	—	0.378	—	0.203
		$5n-r$		2 r	—	1.915	—	0.606	—	0.122	—	0.278	—	0.287	—	0.260	—	0.202	—	0.115
		$4n-r$		3 r	—	3.910	—	1.885	—	1.076	—	0.199	—	0.081	—	0.009	—	0.026	—	0.028
		$3n-r$		4 r	—	5.906	—	3.164	—	2.030	—	0.676	—	0.449	—	0.278	—	0.150	—	0.059
		$2n-r$		5 r	—	7.902	—	4.442	—	2.984	—	1.153	—	0.817	—	0.547	—	0.326	—	0.146
		$1n-r$		6 r	—	9.897	—	5.721	—	3.939	—	1.630	—	1.185	—	0.815	—	0.502	—	0.233
		$0n-r$		7 r	—	11.893	—	7.000	—	4.893	—	2.107	—	1.553	—	1.084	—	0.678	—	0.320
	z_1	7:0			—	1.776	—	1.650	—	1.486	—	0.932	—	0.723	—	0.502	—	0.278	—	0.084
	z_1	7:0	e_1	7:0	—	2.045	—	1.833	—	1.601	—	0.959	—	0.738	—	0.510	—	0.281	—	0.084
	z_1	7:0	e_1	6:1	—	1.775	—	1.655	—	1.494	—	0.938	—	0.727	—	0.505	—	0.279	—	0.084
	z_1	7:0	e_1	5:2	—	1.773	—	1.647	—	1.484	—	0.932	—	0.723	—	0.502	—	0.278	—	0.084
	z_1	7:0	e_1	4:3	—	1.773	—	1.647	—	1.483	—	0.930	—	0.722	—	0.502	—	0.278	—	0.084
	z_1	6:1			—	0.220	—	0.371	—	0.532	—	0.456	—	0.360	—	0.247	—	0.131	—	0.037
	z_1	6:1	e_1	7:0	—	0.049	—	0.554	—	0.647	—	0.483	—	0.375	—	0.255	—	0.134	—	0.037
	z_1	6:1	e_1	6:1	—	0.221	—	0.376	—	0.540	—	0.462	—	0.364	—	0.250	—	0.132	—	0.037
	z_1	6:1	e_1	5:2	—	0.223	—	0.368	—	0.530	—	0.456	—	0.360	—	0.247	—	0.131	—	0.037

Appendix 6—continued

No.	Scored children count		Recombination fraction (θ)												
			0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45			
	z_1	6:1	4:3	e_1	4:3	-0.223	0.368	0.529	0.556	0.523	0.454	0.359	0.247	0.131	0.037
	z_1	5:2				-2.216	-0.907	-0.422	-0.192	-0.070	-0.007	0.019	0.022	0.014	0.004
	z_1	5:2		e_1	7:0	-1.947	-0.724	-0.307	-0.120	-0.025	0.020	0.034	0.030	0.017	0.004
	z_1	5:2		e_1	6:1	-2.217	-0.902	-0.414	-0.183	-0.062	-0.001	0.023	0.025	0.015	0.004
	z_1	5:2		e_1	5:2	-2.219	-0.910	-0.424	-0.193	-0.071	-0.007	0.019	0.022	0.014	0.004
	z_1	5:2		e_1	4:3	-2.219	-0.910	-0.425	-0.195	-0.073	-0.009	0.018	0.022	0.014	0.004
	z_1	4:3				-4.207	-2.164	-1.331	-0.877	-0.581	-0.375	-0.227	-0.123	-0.053	-0.013
	z_1	4:3		e_1	7:0	-3.938	-1.981	-1.216	-0.805	-0.536	-0.348	-0.212	-0.115	-0.050	-0.013
	z_1	4:3		e_1	6:1	-4.208	-2.159	-1.323	-0.868	-0.573	-0.369	-0.223	-0.120	-0.052	-0.013
	z_1	4:3		e_1	5:2	-4.210	-2.167	-1.353	-0.878	-0.581	-0.375	-0.227	-0.123	-0.053	-0.013
	z_1	4:3		e_1	4:3	-4.210	-2.167	-1.334	-0.880	-0.584	-0.377	-0.228	-0.124	-0.053	-0.013
8.		$8n-r$	$0r$			2.373	2.230	2.042	1.844	1.633	1.409	1.169	0.912	0.633	0.331
		$7n-r$	$1r$			0.378	0.951	1.088	1.090	1.031	0.932	0.801	0.643	0.457	0.241
		$6n-r$	$2r$			-1.618	-0.327	0.134	0.337	0.429	0.454	0.433	0.374	0.281	0.157
		$5n-r$	$3r$			-3.614	-1.606	-0.821	-0.416	-0.173	-0.023	0.065	0.105	0.105	0.070
		$4n-r$	$4r$			-5.609	-2.885	-1.775	-1.170	-0.775	-0.500	-0.303	-0.164	-0.071	-0.017
		$3n-r$	$5r$			-7.605	-4.164	-2.729	-1.923	-1.377	-0.977	-0.671	-0.433	-0.247	-0.105
		$2n-r$	$6r$			-9.600	-5.442	-3.683	-2.676	-1.979	-1.454	-1.039	-0.702	-0.423	-0.192
		$1n-r$	$7r$			-11.596	-6.721	-4.638	-3.430	-2.581	-1.931	-1.407	-0.970	-0.599	-0.279
		$0n-r$	$8r$			-13.592	-8.000	-5.592	-4.183	-3.184	-2.408	-1.775	-1.239	-0.775	-0.366
		z_1	8:0			2.072	1.929	1.741	1.543	1.332	1.108	0.868	0.614	0.349	0.110
		z_1	8:0	e_1	8:0	2.339	2.102	1.845	1.605	1.368	1.129	0.879	0.619	0.351	0.110
		z_1	8:0	e_1	7:1	2.072	1.935	1.750	1.552	1.340	1.114	0.872	0.616	0.349	0.110
		z_1	8:0	e_1	6:2	2.070	1.928	1.741	1.543	1.332	1.108	0.868	0.614	0.349	0.110
		z_1	8:0	e_1	5:3	2.070	1.927	1.739	1.542	1.331	1.108	0.868	0.614	0.349	0.110
		z_1	8:0	e_1	4:4	2.070	1.927	1.739	1.541	1.331	1.107	0.868	0.614	0.349	0.110
		z_1	7:1			0.077	0.650	0.787	0.789	0.730	0.631	0.503	0.352	0.193	0.057

Appendix 6: Linkage (lod) scores for families with up to ten children—continued

No.	Scored children		Recombination fraction (θ)												
	z_1	count	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45			
	e_1	8:0	0.344	0.823	0.891	0.851	0.766	0.652	0.514	0.357	0.195	0.057			
	e_1	7:1	0.077	0.656	0.796	0.798	0.738	0.637	0.507	0.354	0.193	0.057			
	e_1	6:2	0.075	0.649	0.787	0.789	0.730	0.631	0.503	0.352	0.193	0.057			
	e_1	5:3	0.075	0.648	0.785	0.788	0.729	0.631	0.503	0.352	0.193	0.057			
	e_1	4:4	0.075	0.648	0.785	0.787	0.729	0.630	0.503	0.352	0.193	0.057			
		6:2	-1.919	-0.629	-0.167	0.036	0.130	0.159	0.146	0.108	0.059	0.017			
	e_1	8:0	-1.652	-0.456	-0.063	0.098	0.166	0.180	0.157	0.113	0.061	0.017			
	e_1	7:1	-1.919	-0.623	-0.158	0.045	0.138	0.165	0.150	0.110	0.059	0.017			
	e_1	6:2	-1.921	-0.630	-0.167	0.036	0.130	0.159	0.146	0.108	0.059	0.017			
	e_1	5:3	-1.921	-0.631	-0.169	0.035	0.129	0.159	0.146	0.108	0.059	0.017			
	e_1	4:4	-1.921	-0.631	-0.169	0.034	0.129	0.158	0.146	0.108	0.059	0.017			
		5:3	-3.915	-1.906	-1.116	-0.704	-0.448	-0.278	-0.163	-0.085	-0.036	-0.009			
	e_1	8:0	-3.648	-1.733	-1.012	-0.642	-0.412	-0.257	-0.152	-0.080	-0.034	-0.009			
	e_1	7:1	-3.915	-1.900	-1.107	-0.695	-0.440	-0.272	-0.159	-0.083	-0.036	-0.009			
	e_1	6:2	-3.917	-1.907	-1.116	-0.704	-0.448	-0.278	-0.163	-0.085	-0.036	-0.009			
	e_1	5:3	-3.917	-1.908	-1.118	-0.705	-0.449	-0.278	-0.163	-0.085	-0.036	-0.009			
	e_1	4:4	-3.917	-1.908	-1.118	-0.706	-0.449	-0.279	-0.163	-0.085	-0.036	-0.009			
		4:4	-5.609	-2.885	-1.775	-1.170	-0.775	-0.500	-0.303	-0.164	-0.071	-0.017			
	e_1	8:0	-5.342	-2.712	-1.671	-1.108	-0.739	-0.479	-0.292	-0.159	-0.069	-0.017			
	e_1	7:1	-5.609	-2.879	-1.766	-1.161	-0.767	-0.494	-0.299	-0.162	-0.071	-0.017			
	e_1	6:2	-5.611	-2.886	-1.775	-1.170	-0.775	-0.500	-0.303	-0.164	-0.071	-0.017			
	e_1	5:3	-5.611	-2.887	-1.777	-1.171	-0.776	-0.500	-0.303	-0.164	-0.071	-0.017			
	e_1	4:4	-5.611	-2.887	-1.777	-1.172	-0.776	-0.501	-0.303	-0.164	-0.071	-0.017			
9.	$9n-r$	0 r	2.670	2.509	2.297	2.074	1.837	1.585	1.315	1.025	0.713	0.373			
	$8n-r$	1 r	0.674	1.230	1.343	1.321	1.235	1.108	0.947	0.757	0.537	0.285			
	$7n-r$	2 r	-1.321	-0.049	0.389	0.567	0.633	0.631	0.579	0.488	0.360	0.198			
	$6n-r$	3 r	-3.317	-1.327	-0.565	-0.186	0.031	0.153	0.211	0.219	0.184	0.111			
	$5n-r$	4 r	-5.313	-2.606	-1.520	-0.939	-0.571	-0.324	-0.157	-0.050	0.008	0.024			
	$4n-r$	5 r	-7.308	-3.885	-2.474	-1.693	-1.173	-0.801	-0.525	-0.319	-0.168	-0.063			

Appendix 6: Linkage (lod) scores for families with up to ten children—continued

No.	Scored children count		e_1	e_2	Recombination fraction (θ)										
					0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
10.	z_1	5:4	e_1	9:0	-5.345	-2.721	-1.682	-1.117	-0.746	-0.484	-0.295	-0.160	-0.070	-0.017	
	z_1	5:4	e_1	8:1	-5.608	-2.879	-1.766	-1.162	-0.769	-0.495	-0.300	-0.162	-0.070	-0.017	
	z_1	5:4	e_1	7:2	-5.610	-2.885	-1.775	-1.169	-0.774	-0.499	-0.302	-0.163	-0.071	-0.017	
	z_1	5:4	e_1	6:3	-5.610	-2.886	-1.776	-1.170	-0.776	-0.500	-0.303	-0.164	-0.071	-0.017	
	z_1	5:4	e_1	5:4	-5.610	-2.886	-1.776	-1.171	-0.776	-0.500	-0.303	-0.164	-0.071	-0.017	
10.		$10n-r$	$0r$		2.967	2.788	2.553	2.304	2.041	1.761	1.461	1.139	0.792	0.414	
		$9n-r$	$1r$		0.971	1.509	1.598	1.551	1.439	1.284	1.093	0.871	0.616	0.327	
		$8n-r$	$2r$		-1.025	0.230	0.644	0.798	0.837	0.807	0.725	0.602	0.440	0.240	
		$7n-r$	$3r$		-3.020	-1.049	-0.310	0.045	0.235	0.330	0.357	0.333	0.264	0.152	
		$6n-r$	$4r$		-5.016	-2.327	-1.264	-0.709	-0.367	-0.148	-0.011	0.064	0.087	0.065	
		$5n-r$	$5r$		-7.012	-3.606	-2.218	-1.462	-0.969	-0.625	-0.379	-0.205	-0.089	-0.022	
		$4n-r$	$6r$		-9.007	-4.885	-3.173	-2.215	-1.571	-1.102	-0.747	-0.474	-0.265	-0.109	
		$3n-r$	$7r$		-11.003	-6.164	-4.127	-2.969	-2.173	-1.579	-1.115	-0.742	-0.441	-0.196	
		$2n-r$	$8r$		-12.998	-7.442	-5.081	-3.722	-2.775	-2.056	-1.483	-1.011	-0.617	-0.283	
		$1n-r$	$9r$		-14.994	-8.721	-6.035	-4.475	-3.377	-2.533	-1.851	-1.280	-0.793	-0.370	
	$0n-r$	$10r$		-16.990	-10.000	-6.990	-5.229	-3.979	-3.010	-2.218	-1.549	-0.969	-0.458		
10.															
	z_1	10:0			2.666	2.487	2.252	2.003	1.740	1.460	1.160	0.839	0.498	0.168	
	z_1	10:0	e_1	10:0	2.927	2.641	2.334	2.048	1.764	1.472	1.166	0.842	0.499	0.168	
	z_1	10:0	e_1	9:1	2.667	2.493	2.260	2.011	1.746	1.464	1.163	0.840	0.499	0.168	
	z_1	10:0	e_1	8:2	2.665	2.486	2.252	2.004	1.741	1.461	1.161	0.840	0.498	0.168	
	z_1	10:0	e_1	7:3	2.665	2.486	2.251	2.003	1.740	1.460	1.160	0.839	0.498	0.168	
z_1	10:0	e_1	6:4	2.665	2.486	2.251	2.003	1.740	1.460	1.160	0.839	0.498	0.168		
z_1	10:0	e_1	5:5	2.665	2.486	2.251	2.003	1.740	1.460	1.160	0.839	0.498	0.168		

Appendix 6: Linkage (lod) scores for families with up to ten children—continued

No.	Scored children count	Recombination fraction (θ)									
		0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
z_1	6:4		- 2.627	- 1.560	- 0.997	- 0.642	- 0.403	- 0.238	- 0.126	- 0.054	- 0.013
e_1	6:4	10:0	- 2.473	- 1.477	- 0.952	- 0.618	- 0.391	- 0.233	- 0.124	- 0.053	- 0.013
z_1	6:4	e_1	- 5.315	- 2.621	- 0.989	- 0.636	- 0.399	- 0.236	- 0.125	- 0.053	- 0.013
z_1	6:4	e_1	- 5.317	- 2.627	- 0.996	- 0.641	- 0.402	- 0.238	- 0.126	- 0.054	- 0.013
z_1	6:4	e_1	- 5.317	- 2.628	- 0.997	- 0.642	- 0.403	- 0.238	- 0.126	- 0.054	- 0.013
z_1	6:4	e_1	- 5.317	- 2.628	- 0.997	- 0.642	- 0.403	- 0.239	- 0.127	- 0.054	- 0.013
z_1	6:4	e_1	- 5.317	- 2.628	- 0.997	- 0.642	- 0.403	- 0.239	- 0.127	- 0.054	- 0.013
z_1	5:5		- 7.012	- 3.606	- 2.218	- 1.462	- 0.969	- 0.625	- 0.379	- 0.205	- 0.089
z_1	5:5	e_1	- 6.751	- 3.452	- 2.136	- 1.418	- 0.946	- 0.613	- 0.373	- 0.202	- 0.088
z_1	5:5	e_1	- 7.010	- 3.600	- 2.210	- 1.455	- 0.964	- 0.621	- 0.376	- 0.204	- 0.088
z_1	5:5	e_1	- 7.012	- 3.606	- 2.218	- 1.461	- 0.968	- 0.624	- 0.378	- 0.204	- 0.088
z_1	5:5	e_1	- 7.012	- 3.607	- 2.219	- 1.462	- 0.969	- 0.625	- 0.379	- 0.205	- 0.089
z_1	5:5	e_1	- 7.012	- 3.607	- 2.219	- 1.463	- 0.969	- 0.625	- 0.379	- 0.205	- 0.089
z_1	5:5	e_1	- 7.012	- 3.607	- 2.219	- 1.463	- 0.969	- 0.625	- 0.379	- 0.205	- 0.089

References

-
- Aird I M, Bentall H H, Roberts J A F 1953 A relationship between cancer of stomach and the ABO blood groups. *British Medical Journal* i: 799–801
- Allen G, Harvald B, Shields J 1967 Measures of twin concordance. *Acta Genetica (Basel)* 17: 475–481
- Allison A C 1964 Polymorphism and natural selection in human populations. *Cold Spring Harbor Symposia on Quantitative Biology* 29: 137–149
- Armitage P 1955 Tests for linear trends in proportions and frequencies. *Biometrics* 11: 375–386
- Armitage P 1971 *Statistical methods in medical research*. Blackwell, Oxford
- Ashley D J B, Davies H D 1966 The use of the surname as a genetic marker in Wales. *Journal of Medical Genetics* 3: 203–211
- Bailey N T J 1951 The estimation of the frequencies of recessives with incomplete multiple selection. *Annals of Eugenics (London)* 16: 215–222
- Bajema C J (ed) 1971 *Natural selection in human populations*. John Wiley, New York
- Barton D E, David F N 1958 A test for birth order effect. *Annals of Human Genetics (London)* 22: 250–257
- Bayes T 1763 An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions* 53: 376–418
- Benirschke K, Kim C K 1973 Multiple pregnancy. *New England Journal of Medicine* 288: 1276–1284; 1329–1336
- Bishop D T 1983 Multifactorial inheritance. In: Emery A E H, Rimoin D (eds) *Principles and practice of medical genetics*. Churchill Livingstone, Edinburgh, p 111–119
- Blank C E 1960 Apert's syndrome (a type of acrocephalosyndactyly)—observations on a British series of thirty-nine cases. *Annals of Human Genetics (London)* 24: 151–164
- Bodmer W F (ed) 1978 *The HLA System* *British Medical Bulletin* 34 (No. 3): 213–316
- Bodmer W F, Bodmer J, Ihde D, Adler S 1969 Genetic and serological association analysis of the HL-A leukocyte system. In: Morton N E (ed) *Computer applications in genetics*. University of Hawaii Press, Honolulu, p 117–127
- Bonaiti C 1978 Genetic counselling of consanguineous families. *Journal of Medical Genetics*. 15: 109–112
- Bonaiti-Pellié C, Smith C 1974 Risk tables for genetic counselling in some common congenital malformations. *Journal of Medical Genetics* 11: 374–377
- Brewerton D A, Caffrey M, Hart F D, James D C O, Nicholls A, Sturrock R D 1973 Ankylosing spondylitis and HLA 27. *Lancet* i: 904–907
- Bulmer M G 1959 The effect of parental age, parity and duration of marriage on the twinning rate. *Annals of Human Genetics (London)* 23: 454–458
- Bulmer M G 1970 *The Biology of Twinning in Man*. Clarendon Press, Oxford
- Bunday S, Harrison M J G, Marsden C D 1975 A genetic study of torsion dystonia. *Journal of Medical Genetics* 12: 12–19
- Burt C, Howard M 1956 The multifactorial theory of inheritance and its application to intelligence. *British Journal of Statistical Psychology* 9(2): 95–131
- Candela P B 1942 The introduction of blood group B into Europe. *Human Biology* 14: 413–443

- Carter C O 1965 The inheritance of common congenital malformations. *Progress in Medical Genetics* 4: 59–84
- Carter C O 1976 Genetics of common single malformations. *British Medical Bulletin* 32: 21–26
- Carter T C, Falconer D S 1951 Stocks for detecting linkage in the mouse, and the theory of their design. *Journal of Genetics* 50: 307–323
- Cavalli-Sforza L L, Bodmer W F 1971 *The genetics of human populations*. Freeman, San Francisco
- Charlesworth B, Charlesworth D 1973 The measurement of fitness and mutation rate in human populations. *Annals of Human Genetics (London)* 37: 175–187
- Chen S, Thompson M W, Rose V 1971 Endocardial fibroelastosis: family studies with special reference to counseling. *Journal of Pediatrics* 79: 385–392
- Clarke C A 1959a Distribution of ABO blood groups and the secretor status in duodenal ulcer families. *Gastroenterologia* 92: 99–103
- Clarke C A 1959b The relative fitness of human mutant genotypes. In: Roberts D F, Harrison G A (eds) *Natural selection in human populations*. Pergamon Press, London, p 17–34
- Clarke C A 1961 Blood groups and disease. *Progress in Medical Genetics* 1: 81–119
- Clayton J 1985 A computer programme to calculate risks in X-linked disorders using multiple marker loci. *Journal of Medical Genetics*
- Clayton J, Emery A E H 1984 DNA probes in Duchenne muscular dystrophy *Lancet* ii: 1151–1152
- Conneally P M, Heuch I 1974 A computer program to determine genetic risks—a simplified version of PEDIG. *American Journal of Human Genetics* 26: 773–775
- Conneally P M, Wallace M R, Gusella J F, Wexler N S 1984 Huntington's disease: estimation of heterozygote status using linked genetic markers. *Genetic Epidemiology* 1: 81–88
- Cross H E, McKusick V A 1967 The Mast syndrome. A recessively inherited form of presenile dementia with motor disturbances. *Archives of Neurology* 16: 1–13
- Crow J F 1965 Problems of ascertainment in the analysis of family data. In: Neel J V, Shaw M W, Schull W J (eds) *Genetics and the epidemiology of chronic diseases*. US Department of Health, Washington, p 23–44
- Crow J F, Mange A P 1965 Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly* 12 (4): 199–203
- Curnow R N 1972 The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk. *Biometrics* 28: 931–946
- Dahlberg G 1947 *Mathematical methods for population genetics*. Karger, Basel
- Danks D M, Allan J, Anderson C M 1965 A genetic study of fibrocystic disease of the pancreas. *Annals of Human Genetics (London)* 28: 323–340
- Davie A M 1979 The 'singles' method for segregation analysis under incomplete ascertainment. *Annals of Human Genetics (London)* 42: 507–512
- Dewey W J, Barrai I, Morton N E, Mi M P 1965 Recessive genes in severe mental defect. *American Journal of Human Genetics* 17: 237–256
- Dyer K F 1976 Patterns of gene flow between negroes and whites in the U.S. *Journal of Biosocial Science* 8: 309–333
- Edwards J H 1960 The simulation of Mendelism. *Acta Genetica* 10: 63–70
- Edwards J H 1961 The recognition and estimation of cyclic trends. *Annals of Human Genetics (London)* 25: 83–87
- Edwards J H 1965 The meaning of the association between blood groups and disease. *Annals of Human Genetics (London)* 29: 77–83
- Edwards J H 1968 The value of twins in genetic studies. *Proceedings of the Royal Society of Medicine* 61: 227–9
- Edwards J H 1969 Familial predisposition in man. *British Medical Bulletin* 25: 58–64
- Edwards J H 1971 The analysis of X-linkage. *Annals of Human Genetics (London)* 34: 229–250
- Edwards J H 1976 Risks of malformed relatives. *Lancet* i: 1348
- Elandt-Johnson R C 1971 *Probability models and statistical methods in genetics*. John Wiley, New York
- Emery A E H 1965 Carrier detection in sex-linked muscular dystrophy. *Journal de Génétique Humaine* 14: 318–329
- Emery A E H 1966 Genetic linkage between the loci for colour blindness and Duchenne type

- muscular dystrophy. *Journal of Medical Genetics* 3: 92-95
- Emery A E H 1980 Duchenne muscular dystrophy. Genetics aspects, carrier detection and antenatal diagnosis. *British Medical Bulletin* 36: 117-122
- Emery A E H 1983 *Elements of medical genetics*, 6th edn. Churchill Livingstone, Edinburgh
- Emery A E H 1984 *An introduction to recombinant DNA*. John Wiley, Chichester-New York (reprinted with revisions, 1985).
- Emery A E H 1985 Identical twinning and oral contraception. *Biology and Society* (in press)
- Emery A E H, Holloway S 1977 Use of normal daughters' and sisters' creatine kinase levels in estimating heterozygosity in Duchenne muscular dystrophy. *Human Heredity* 27: 118-126
- Emery A E H, Lawrence J S 1967 Genetics of ankylosing spondylitis. *Journal of Medical Genetics* 4: 239-244
- Emery A E H, Morton R 1968 Genetic counselling in lethal X-linked disorders. *Acta Genetica (Basel)* 18: 534-542
- Emery A E H, Skinner R 1976 Clinical studies in benign (Becker type) X-linked muscular dystrophy. *Clinical Genetics* 10: 189-201
- Emery A E H, Smith C A B, Sanger R 1969 The linkage relations of the loci for benign (Becker type) X-borne muscular dystrophy, colour blindness and the Xg blood groups. *Annals of Human Genetics (London)* 32: 261-269
- Emery A E H, Davie A M, Smith C 1975 Spinal muscular atrophy—resolution of heterogeneity. In: Bradley W G (ed) *Recent advances in myology*. Excerpta Medica, Amsterdam, p 557-565
- Evans D A P, Manley K A, McKusick V A 1960 Genetic control of isoniazid metabolism in man. *British Medical Journal* 2: 485-491
- Falconer D S 1965 The inheritance of liability to certain diseases estimated from the incidence among relatives. *Annals of Human Genetics (London)* 29: 51-76
- Falconer D S 1967 The inheritance of liability to diseases with variable age of onset with particular reference to diabetes mellitus. *Annals of Human Genetics (London)* 31: 1-20
- Falconer D S 1981 *Introduction to quantitative genetics*, 2nd edn. Longman, London
- Fedrick J 1970 Anencephalus: variation with maternal age, parity, social class and region in England, Scotland and Wales. *Annals of Human Genetics (London)* 34: 31-38
- Feltkamp T E W, Van den Berg-Loonen P M, Nijenhuis L E, Engelfriet C P, Van Rossum A L, Van Loghem, J J, Oosterhuis H J G H 1974 Myasthenia gravis, autoantibodies and HL-A antigens. *British Medical Journal* i: 131-133
- Fisher R A 1930 *The genetical theory of natural selection*. Clarendon Press, Oxford. Also now available as a paperback edition published in 1958 by Dover, New York
- Fisher, R A 1934 The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics (London)* 6: 13-25
- Fisher R A 1970 *Statistical methods for research workers*, 14th edn. Oliver & Boyd, Edinburgh
- Fisher R A, Yates F 1963 *Statistical tables for biological, agricultural and medical research*, 6th edn. Oliver & Boyd, Edinburgh
- Francke U 1983 Gene mapping. In: Emery A E H, Rimoin D (eds) *Principles and practice of medical genetics*. Churchill Livingstone, Edinburgh, p 91-110
- Fraser F C 1976 The multifactorial/threshold concept—uses and abuses. *Teratology* 14: 267-279
- Fraser G R, Friedmann A I 1967 *The causes of blindness in childhood*. Johns Hopkins, Baltimore
- Fraser G R, Mayo O 1974 Genetical load in man. *Humangenetik* 23: 83-110
- Freedman L S 1979 The use of a Kolmogorov-Smirnov type statistic in testing hypotheses about seasonal variation. *Journal of Epidemiology and Community Health* 33: 223-228
- Freedman L S 1981 Watson's U_N^2 statistic for a discrete distribution. *Biometrika* 68: 708-711
- Freire-Maia N 1976 Genetic loads in man. *Human Heredity* 26: 95-104
- Fritze D, Herman C, Naeim F, Smith G S, Walford R L 1974 HL-A antigens in myasthenia gravis. *Lancet* i: 240-242
- Gaines R E, Elston R C 1969 On the probability that a twin pair is monozygotic. *American Journal of Human Genetics* 21: 457-465
- Gardner R J M 1977 A new estimate of the achondroplasia mutation rate. *Clinical Genetics* 11: 31-38

- Gardner-Medwin D 1970 Mutation rate in Duchenne type of muscular dystrophy. *Journal of Medical Genetics* 7: 334–337
- Glass B, Sacks M S, Jahn E F, Hess C 1952 Genetic drift in a religious isolate: an analysis of the causes of variation in blood group and other gene frequencies in a small population. *American Naturalist* 86: 145–159
- Gottesman I I, Shields J 1972 *Schizophrenia and Genetics: A twin study vantage point*. Academic Press, New York
- Gottesman I I, Shields J 1973 Genetic theorizing and schizophrenia. *British Journal of Psychiatry* 122: 15–30
- Greenwood M, Yule G U 1914 On the determination of size of family and of the distribution of characters in order of birth. *Journal of the Statistical Society* 77: 179–197
- Haldane J B S 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* 8: 299–309
- Haldane J B S 1938 The estimation of the frequencies of recessive conditions in man. *Annals of Eugenics (London)* 8: 255–262
- Haldane J B S 1941 The relative importance of principal and modifying genes in determining some human diseases. *Journal of Genetics* 41: 149–157
- Haldane J B S 1951 Simple tests for bimodality and bitangentiality. *Annals of Eugenics (London)* 16: 359–364
- Haldane J B S, Smith C A B 1947 A simple exact test for birth-order effect. *Annals of Eugenics (London)* 14: 117–124
- Hardy G H 1908 Mendelian proportions in a mixed population. *Science* 28: 49–50
- Harris E L, Wagener D K, Dorman J S, Drash A L 1985 Detection of genetic heterogeneity between families of insulin dependent diabetes mellitus patients using linkage analysis. *American Journal of Human Genetics* 37: 102–113
- Harris H, Smith C A B 1947 The sib-sib age of onset correlation. *Annals of Eugenics (London)* 14: 309–318
- Harris R 1983 The HLA system. In: Emery A E H, Rimoin D (eds) *Principles and practice of medical genetics*. Churchill Livingstone, Edinburgh, p 1127–1133
- Heuch I, Li F H F 1972 PEDIG—A computer programme for calculation of genotype probabilities using phenotype information. *Clinical Genetics* 3: 501–504
- Hewitt D, Milner J, Csima A, Pakula A 1971 On Edwards' criterion of seasonality and a non-parametric alternative. *British Journal of Preventive and Social Medicine* 25: 174–176
- Hodge S E, Anderson C E, Neiswanger K, Sparkes R S, Rimoin D L 1983 The search for heterogeneity in insulin-dependent diabetes mellitus. *American Journal of Human Genetics* 35: 1139–1155
- Hogben L 1931 The genetic analysis of familial traits. I. Single gene substitutions. *Journal of Genetics* 25: 97–112
- Hogben L 1946 *An introduction to mathematical genetics*. Norton, New York
- Holzinger K J 1929 The relative effect of nature and nurture influences on twin differences. *Journal of Educational Psychology* 20: 241–248
- Jacob A, Clack E R, Emery A E H 1968 Genetic study of sample of 70 patients with myasthenia gravis. *Journal of Medical Genetics* 5: 257–261
- James W H 1972 Secular changes in dizygotic twinning rates. *Journal of Biosocial Science* 4: 427–434
- James W H 1976 The possibility of a flaw underlying Weinberg's differential rule. *Annals of Human Genetics (London)* 40: 197–199
- Johnston F E, Jantz R L, Kensinger K M, Walker G F, Allen F H, Walker M E 1968 Red cell blood groups of the Peruvian Cashinahua. *Human Biology* 40: 508–516
- Johnston F E, Kensinger K M, Jantz R L, Walker G F 1969 The population structure of the Peruvian Cashinahua: demographic, genetic and cultural inter-relationships. *Human Biology* 41: 29–41
- Jones K L, Smith D W, Harvey M A S, Hall B D, Quan L 1975 Older paternal age and fresh gene mutation: data on additional disorders. *Journal of Pediatrics* 86: 84–88
- Kate L P ten 1977 A method for analysing fertility of heterozygotes for autosomal recessive disorders, with special reference to cystic fibrosis, Tay-Sachs disease and phenylketonuria. *Annals of Human Genetics (London)* 40: 287–297
- Kellermann G, Luyten-Kellermann M, Shaw C R 1973 Genetic variation of aryl hydrocarbon hydroxylase in human lymphocytes. *American Journal of Human Genetics* 25: 327–331

- Kelly T E, Chase G A, Kaback M M, Kumor K, McKusick V A 1975 Tay-Sachs disease: high gene frequency in a non-Jewish population. *American Journal of Human Genetics* 27: 287-291
- Kimura M, Crow J F 1963 The measurement of effective population number. *Evolution* 17: 279-288
- Knudson A G 1979 Our load of mutations and its burden of disease. *American Journal of Human Genetics* 31: 401-413
- Knudson A G, Wayne L, Hallett W Y 1967 On the selective advantage of cystic fibrosis heterozygotes. *American Journal of Human Genetics* 19: 388-392
- Koeslag J H, Schach S R 1984 Tay-Sachs disease and the role of reproductive compensation in the maintenance of ethnic variations in the incidence of autosomal recessive disease. *Annals of Human Genetics (London)* 48: 275-281
- Kosambi D D 1944 The estimation of map distances from recombination values. *Annals of Eugenics (London)* 12: 172-175
- Kruskal W H, Wallis W A 1952 The use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47: 583-621
- Kudo A, Sakaguchi K 1963 A method for calculating the inbreeding coefficient. II Sex-linked genes. *American Journal of Human Genetics* 15: 476-480
- Laberge C 1966 Prospectus for genetic studies in the French Canadians with preliminary data on blood groups and consanguinity. *Bulletin of Johns Hopkins Hospital* 118: 52-68
- Lathrop G M, Lalouel J M, Julier C, Ott J 1985 Multilocus linkage analysis in humans. *American Journal of Human Genetics* 37: 482-498
- Lenz W 1961 Kindliche Missbildungen nach Medikament—Einnahme während der Gravidität? *Deutsche Medizinische Wochenschrift* 86: 2555-2556
- Levitan M, Montagu A 1977 Textbook of human genetics, 2nd edn. Oxford University Press, London p 449
- Li C C 1961 Human genetics. McGraw-Hill, New York
- Li C C, Mantel N 1968 A simple method of estimating the segregation ratio under complete ascertainment. *American Journal of Human Genetics* 20: 61-81
- McBride W G 1961 Thalidomide and congenital abnormalities. *Lancet* ii: 1358
- McDevitt H O, Bodmer W F 1974 HL-A, immune-response genes, and disease. *Lancet* i: 1269-1275
- MacGillivray I, Nylander P P S, Corney G 1975 Human multiple reproduction. Saunders, London
- McKeown T, Record R G 1956 Maternal age and birth order as indices of environmental influence. *American Journal of Human Genetics* 8: 8-23
- McKusick V A 1980 The anatomy of the human genome. *Journal of Heredity* 71: 370-391
- Mayo O 1970 On the maintenance of polymorphisms having an inviable homozygote. *Annals of Human Genetics (London)* 33: 307-317
- Morton N E 1955 Sequential tests for the detection of linkage. *American Journal of Human Genetics* 7: 277-318
- Morton N E 1956 The detection and estimation of linkage between the genes for Elliptocytosis and the Rh blood type. *American Journal of Human Genetics* 8: 80-96
- Morton N E 1959 Genetic tests under incomplete ascertainment. *American Journal of Human Genetics* 11: 1-16
- Morton N E, Chung C S 1959 Formal genetics of muscular dystrophy. *American Journal of Human Genetics* 11: 360-379
- Morton N E, Rao D C, Lalouel J-M 1983 Methods in genetic epidemiology. Karger, New York
- Mourant A E, Kopec A C, Domaniewska-Sobczak K 1976 The distribution of the human blood groups and other polymorphisms, 2nd edn. Oxford University Press, London
- Murdoch J L, Walker B A, Hall J G, Abbey H, Smith K K, McKusick V A 1970 Achondroplasia—a genetic and statistical survey. *Annals of Human Genetics (London)* 33: 227-244
- Murdoch J L, Walker B A, McKusick V A 1972 Parental age effects on the occurrence of new mutations for the Marfan syndrome. *Annals of Human Genetics (London)* 35: 331-336
- Murphy E A, Bolling D R 1967 Testing of single locus hypotheses where there is incomplete separation of the phenotypes. *American Journal of Human Genetics* 19: 322-334
- Murphy E A, Chase G A 1975 Principles of genetic counselling. Year Book Medical

- Publishers, Chicago
- Murphy E A, Mutalik G S 1969 The application of Bayesian methods in genetic counselling. *Human Heredity* 19: 126–151
- Myriantopoulos N C, Aronson S M 1966 Population dynamics of Tay-Sachs disease. I. Reproductive fitness and selection. *American Journal of Human Genetics* 18: 313–327
- Nance W E, Corey L A 1976 Genetic models for the analysis of data from the families of identical twins. *Genetics* 83: 811–826
- Neel J V, Schull W J 1954 *Human heredity*. University of Chicago Press, Chicago
- Nielsen J 1967 Inheritance in monozygotic twins. *Lancet* ii: 717–718
- Nielsen J, Holm V, Haahr, J 1975 Prevalence of Edwards' syndrome. Clustering and seasonal variation? *Humangenetik* 26: 113–116
- Opitz J M 1981 Some comments on penetrance and related subjects. *American Journal of Medical Genetics* 8: 265–274
- Osborne R H, De George F V 1959 Genetic basis of morphological variation—an evaluation and application of the twin study method. Harvard University Press, Cambridge, Massachusetts
- Osborne R H, Adlersberg D, De George F V, Wang C 1959 Serum lipids, heredity and environment. *American Journal of Medicine* 26: 54–59
- Ott, J 1974 Estimation of the recombination fraction in human pedigrees. *American Journal of Human Genetics* 26: 588–597; 28: 528–529
- Parisi P, Gatti M, Prinzi G, Caperna G 1983 Familial incidence of twinning. *Nature (London)* 304: 626–628
- Pauli R M, Motulsky A G 1981 Risk counselling in autosomal dominant disorders with undetermined penetrance. *Journal of Medical Genetics* 18: 340–343
- Pearn J H 1973 The gene frequency of acute Werdnig-Hoffmann disease (SMA type I). A total population survey of North-East England. *Journal of Medical Genetics* 10: 260–265
- Penrose L S 1935 The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics (London)* 6: 133–138
- Penrose L S 1951 Measurement of pleiotropic effects in phenylketonuria. *Annals of Eugenics (London)* 16: 134–141
- Penrose L S 1953a The general purpose sib-pair linkage test. *Annals of Eugenics (London)* 18: 120–124
- Penrose L S 1953b The genetical background of common diseases. *Acta Genetica* 4: 257–265
- Penrose L S 1957 Parental age in achondroplasia and mongolism. *American Journal of Human Genetics* 9: 167–169
- Philippe P 1985 Genetic epidemiology of twinning: a population based study. *American Journal of Medical Genetics* 20: 97–105
- Race R R, Sanger R 1975 *Blood groups in man*, 6th edn. Blackwell, Oxford
- Rao D C, Morton N E, Lindsten J, Hutton M, Yee S 1977 A mapping function for man. *Human Heredity* 27: 99–104
- Record R G, McKeown T, Edwards J H 1969 The relation of measured intelligence to birth order and maternal age. *Annals of Human Genetics (London)* 33: 61–69
- Reed T E 1959 The definition of relative fitness of individuals with specific genetic traits. *American Journal of Human Genetics* 11: 137–155
- Reed T E 1969 Caucasian genes in American Negroes. *Science* 165: 762–768
- Reed T E, Chandler J H, Hughes E M, Davidson R T 1958 Huntington's chorea in Michigan. I. Demography and genetics. *American Journal of Human Genetics* 10: 201–225
- Registrar General. Statistical Review of England and Wales. HMSO, London
- Registrar General, Scotland. Annual Reports. HMSO, Edinburgh
- Renwick J H 1969 Progress in mapping human autosomes. *British Medical Bulletin* 25: 65–73
- Renwick J H 1971 The mapping of human chromosomes. *Annual Review of Genetics* 5: 81–120
- Roberts D F (ed) 1975 *Human variation and natural selection*. Taylor & Francis, London
- Roberts D F 1977 Assortative mating in man. *Bulletin of the Eugenics Society London*. Supplement No 2
- Roberts D F, Billewicz W Z, McGregor I A 1978 Heritability of stature in a West African population. *Annals of Human Genetics (London)* 42: 15–24
- Roberts J A F 1957 Blood groups and susceptibility to disease: a review. *British Journal of Preventive and Social Medicine* 11: 107–125

- Roberts J A F, Pembrey M E 1978 An introduction to medical genetics, 7th edn. Oxford University Press, London, p 17
- Ryder L P, Svejgaard A 1981 Genetics of HLA disease association. *Annual Review of Genetics* 15: 169–187
- Salzano F M, Neel J V, Maybury-Lewis D 1967 Further studies on the Xavante Indians. I. Demographic data on two additional villages: genetic structure of the tribe. *American Journal of Human Genetics* 19: 463–489
- Schlosstein L, Terasaki P I, Bluestone R, Pearson C M 1973 High association of an HL-A antigen, W 27, with ankylosing spondylitis. *New England Journal of Medicine* 288: 704–706
- Shokeir M H K 1975 Investigation on Huntington's disease in the Canadian prairies. II. Fecundity and fitness. *Clinical Genetics* 7: 349–353
- Siegel S 1956 Non parametric statistics. McGraw-Hill, New York
- Simpson S P 1983 Estimating the ascertainment probability from the number of ascertainment per proband. *Human Heredity* 33: 103–108
- Skinner R, Emery A E H, Anderson A J B, Foxall C 1975 The detection of carriers of benign (Becker-type) X-linked muscular dystrophy. *Journal of Medical Genetics* 12: 131–134
- Smith C 1970 Heritability of liability and concordance in monozygous twins. *Annals of Human Genetics (London)* 34: 85–91
- Smith C 1972a Correlation in liability among relatives and concordance in twins. *Human Heredity* 22: 97–101
- Smith C 1972b Computer programme to estimate recurrence risks for multifactorial familial disease. *British Medical Journal* i: 495–497
- Smith C 1974 Concordance in twins: methods and interpretation. *American Journal of Human Genetics* 26: 454–466
- Smith C 1976 Statistical resolution of genetic heterogeneity in familial disease. *Annals of Human Genetics (London)* 39: 281–291
- Smith C A B 1963 Testing for heterogeneity of recombination fraction values in human genetics. *Annals of Human Genetics (London)* 27: 175–182
- Smith C A B 1968 Linkage scores and corrections in simple two- and three-generation families. *Annals of Human Genetics (London)* 32: 127–150
- Smith C A B 1972 Note on the estimation of parental age effects. *Annals of Human Genetics (London)* 35: 337–342
- Smith C A B 1977 A note on genetic distance. *Annals of Human Genetics (London)* 40: 463–479
- Smith S M, Penrose L S 1955 Monozygotic and dizygotic twin diagnosis. *Annals of Human Genetics (London)* 19: 273–289
- Smith S M, Penrose L S, Smith C A B 1961 Mathematical tables for research workers in human genetics. Churchill, London
- Snedecor G W, Cochran W G 1967 Statistical methods, 6th edn. Iowa State University Press, Ames, Iowa
- Spuhler J N 1963 The scope for natural selection in man. In: Schull W J (ed) Genetic selection in man. University of Michigan Press, Ann Arbor, p 1–111
- Steinberg A G 1959 Methodology in human genetics. *Journal of Medical Education* 34: 315–334
- Stene J, Stene E 1977 Statistical methods for detecting a moderate paternal age effect on incidence of disorder when a maternal one is present. *Annals of Human Genetics (London)* 40: 343–353
- Stevenson A C, Kerr C B 1967 On the distributions of frequencies of mutation to genes determining harmful traits in man. *Mutation Research* 4: 339–352
- Svejgaard A, Platz P, Ryder L P, Nielsen L S, Thomsen M 1975 HL-A and disease associations—a survey. *Transplantation Review* 22: 3–43
- Tanaka K 1974 A new simplified method for estimating relative fitness in man. *Japanese Journal of Human Genetics* 19: 195–202
- Tanaka K 1975 Estimation of relative fitness in human abnormalities with sex difference in selection intensity: a new simplified method. *Japanese Journal of Human Genetics* 20: 183–186
- Tay J S H, Yip W C L 1984 The estimation of inbreeding from isonymy. *Annals of Human Genetics (London)* 48: 185–194

- Thoday J M 1975 Non-Darwinian 'evolution' and biological progress. *Nature* (London) 255: 675-677
- Tills D 1977 The use of the F_{ST} statistic of Wright. *Human Heredity* 27: 153-159
- Tünte W, Becker P E, Knorre G 1967 Zur Genetik der Myositis ossificans progressiva. *Humangenetik* 4: 320-351
- Vetta A 1976 Correction to Fisher's correlations between relatives and environmental effects. *Nature* 263: 316-317
- Vogel F 1970 ABO blood groups and disease. *American Journal of Human Genetics* 22: 464-475
- Vogel F 1983 Mutation in man. In: Emery A E H, Rimoin D (eds) *Principles and practice of medical genetics*. Churchill Livingstone, Edinburgh, p 26-48
- Vogel F, Helmbold W 1972 Blutgruppen—Populationsgenetik und statistik. In: Becker P (ed) *Humangenetik*, Vol 1. Thieme, Stuttgart, p 129-557
- Vogel F, Rattenberg R 1975 Spontaneous mutation in man. In: Harris H, Hirschhorn K (eds) *Recent advances in human genetics*, Vol 5. Plenum, New York, p 223-318
- Wagener D K, Cavalli-Sforza L L 1975 Ethnic variation in genetic disease: possible role of hitchhiking and epistasis. *American Journal of Human Genetics* 27: 348-364
- Walter S D 1977 The power of a test for seasonality. *British Journal of Preventive and Social Medicine* 31: 137-140
- Walter S D, Elwood J M 1975 A test for seasonality of events with a variable population at risk. *British Journal of Preventive and Social Medicine* 29: 18-21
- Weinberg, W 1901 Beiträge zur Physiologie und Pathologie der Mehrlingsgeburten beim Menschen. *Pflugers Archiv für die gesamte Physiologie des Menschen und der Tiere* 88: 346-430
- Weinberg W 1908 Über den Nachweis der Vererbung beim Menschen. *Jahresh. Verein f. vaterl. Naturk. in Württemberg* 64: 368-82 (see Stern C 1943 *The Hardy-Weinberg Law*. *Science* 97: 137-138)
- Weiner J S, Huizinga J 1972 *The assessment of population affinities in man*. Clarendon Press, Oxford
- Wiener A S 1970 Blood groups and disease. *American Journal of Human Genetics* 22: 476-483
- Winter R M 1985 The estimation of recurrence risks in monogenic disorders using flanking marker loci. *Journal of Medical Genetics* 22: 12-15
- Winter R M, Harding A E, Baraitser M, Bravery M B 1981 Intrafamilial correlation in Friedreich's ataxia. *Clinical Genetics* 20: 419-427
- Woodward R H, Goldsmith P L 1964 *Mathematical and statistical techniques for industry: Monograph No. 3. Cumulative sum techniques*. Oliver & Boyd, Edinburgh
- Woolf B 1955 On estimating the relation between blood group and disease. *Annals of Human Genetics* (London) 19: 251-253
- Woolf L I, McBean M S, Woolf F M, Cahalane S F 1975 Phenylketonuria as a balanced polymorphism: the nature of the heterozygote advantage. *Annals of Human Genetics* (London) 38: 461-469
- Workman P L, Blumberg B S, Cooper A J 1963 Selection, gene migration and polymorphic stability in a U.S. white and negro population. *American Journal of Human Genetics* 15: 429-437
- Wright S 1922 Coefficients of inbreeding and relationship. *American Naturalist* 56: 330-338
- Wright S 1926 Effects of age of parents on characteristics of the guinea pig. *American Naturalist* 60: 552-559
- Wright S 1948 On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* 2: 279-294
- Wright S 1950/51 The genetical structure of populations. *Annals of Eugenics* (London) 15: 323-354
- Wynne-Davies R 1970 The genetics of some common congenital malformations. In: Emery A E H (ed) *Modern trends in human genetics*, Vol 1. Butterworths, London, p 316-338

Index

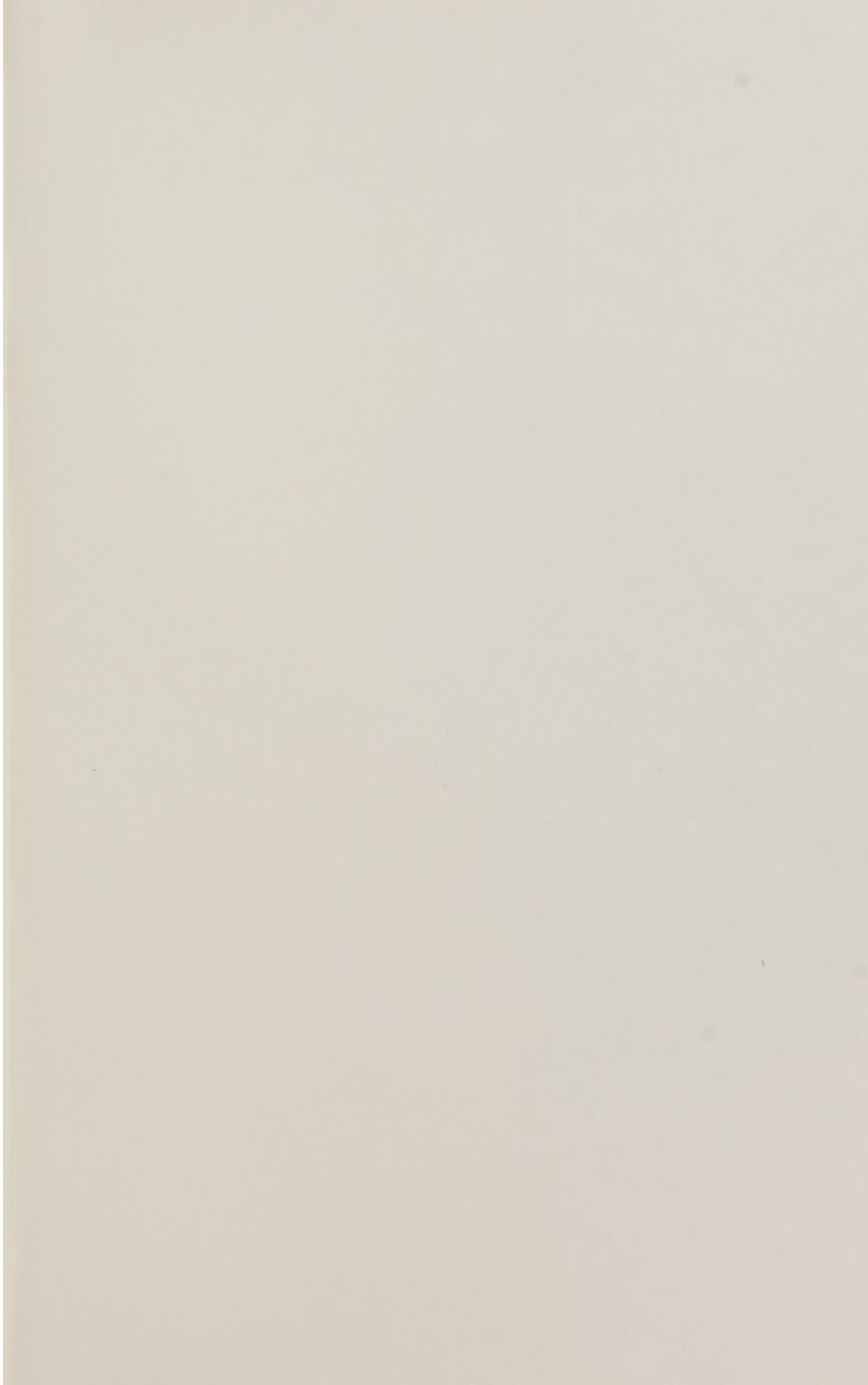
-
- Achondroplasia
 mutation rate, 33–4
 parental age effect, 140, 148
- Acrocephalosyndactyly (Apert's syndrome),
 paternal age effect, 140, 148, 150
- Albinism, 23
- Alkaptonuria, 5, 23
- Allen's law, 127
- Amish, 40
- Anencephaly *see* CNS malformations
- Ankylosing spondylitis
 association HLA-B27, 119, 120, 124
 heritability, 58
 inheritance, 57
 recurrence risks, 124
- Apert's syndrome, parental age effect, 140,
 148, 150
- Aryl hydrocarbon hydroxylase, 6
- Ascertainment
 complete, 40
 multiple incomplete, 51–3
 probability, 39
 single incomplete, 46
- Association *see* Disease association
- Assortative mating, 15–17
- Asthma
 heritability, 58
 recurrence risks, 112
- Autosomal dominant inheritance, tests for,
 37–8
- Autosomal gene frequency
 heterozygote not recognizable, 4–5
 heterozygote recognizable, 5–7
 standard error, 5
- Average inbreeding coefficient, 20
- Bayes' theorem, 93
 see also Recurrence risks
- Becker muscular dystrophy
 clinical course, 133
 fitness, 30, 32
 recurrence risks, 102
- Bernstein's equation, 24
- Birth order, population data
 correlations, 148
 means, 146
- Birth order effect
 examples, 140–1
 methods of estimation
 choice of controls, 145
 Greenwood–Yule method, 147–8
 Haldane and Smith method, 141–5
 partial correlations, 148–53
 see also Parental age effect
- Blood groups
 associations, 114–121
 frequencies, 10–11
 linkage (Lutheran), 69–71
- Bonferroni inequality, 122
- Cashinahua Indians, 15
- Centimorgan, 70
- Cerebral palsy, recurrence risks, 112
- Cleft lip +/- cleft palate
 heritability, 58
 recurrence risks, 112
- Cline, 23
- Club foot (congenital)
 heritability, 58
 recurrence risks, 112
- CNS malformations
 cyclical changes in incidence, 161
 heritability, 58
 maternal age and birth order effects, 141
 recurrence risks, 112
- Coefficients
 average inbreeding, 20
 inbreeding (F), 17
 relationship (R), 21
 selection (s), 25
- Complete ascertainment
 a priori method, 40
 maximum likelihood method, 42
 'singles' method, 44
- Concordance in twins
 heritability from, 64–6

- Concordance in twins (*contd*)
 pairwise, 87
 proband, 87
- Congenital malformations
 inheritance, 55
 recurrence risks, 112
- Consanguinity, 17 *et seq.*
 genetic heterogeneity and, 135–9
- Coronary artery disease, heritability, 58
- Correlation
 between parent–offspring, 16, 63
 between sibs, 16, 63–4
 between spouses, 16
 in liability, 58 *et seq.*
 intraclass, 63, 90
 partial, 148–53
 significance, 151–3
 'z' transformation, 151
- Colour blindness, linkage relationships, 75–7
- Counselling, genetic *see* Recurrence risks
- Cousin marriages, 22–3
- 'Cusums', 156–9
- Cyclical changes, 159–63
- Cystic fibrosis
 gene frequency, 23
 heterozygote advantage, 27–8, 30
- Cystinuria, 23
- Dahlberg's formula, 22, 138
see also Cousin marriages
- Deafness, profound childhood, recurrence risks, 112
- Dermal ridge count, assortive mating, 17
- Diabetes mellitus, recurrence risks, 112
- Disease association
 explanations for, 114, 123
 genetic heterogeneity, 139
 problems, 122–3
 sibship analysis
 Penrose method 114–16
 Smith method 121
 statistical analysis (Woolf), 116–21
 value in
 genetic counselling, 123–4
 pathogenesis, 123
 resolution of heterogeneity, 124
see also Blood groups and HLA antigens
- Disease frequency, recognition and
 estimation of changes in, 154 *et seq.*
- Dislocation of the hip (congenital)
 heritability, 58, 60–2
 recurrence risks, 112
- DNA markers, 103–9
- Drift *see* Genetic drift
- Duchenne muscular dystrophy, 32
 carrier detection, 96–100, 101
 clinical course, 133
 linkage relationships, 75–7
 mutation rate, 34
 parental age effect, 140
 recurrence risks, 96–102
- Down's syndrome, 140
- Dunkers, 13
- 'e' score *see* Linkage
- Edward's syndrome, 140
- Effective population size
 coefficient of inbreeding and, 25
 definition, 12
 estimation, 13–14
 gene frequencies and 14–15
- Empiric risks *see* Recurrence risks
- Endocardial fibroelastosis, 111–13
- Epilepsy (idiopathic), recurrence risks, 112
- Epistasis, disease association, cause of, 114
- Evolution
 Darwinian, 15
 non-Darwinian, 15
- Exomphalos *see* CNS
- Fetal membranes *see* Zygosity
- Fibrocystic disease *see* Cystic Fibrosis
- Fitness
 definition, 29
 estimation, based on
 general population, 29–30
 sibs, 29–30
 Tanaka's method, 31
 X-linked disorders, 32–3
- Founder effect, 28
- Frequency
 gene, 4–7
 mating and offspring types
 autosomal disorders, 8–10
 X-linked disorders, 10
 relative (K), 57
see also Disease frequency
- Friedreich's ataxia, 139
- Gene
 flow, 23–5
 frequency
 effective population size and, 14–15
 estimation of autosomal
 heterozygote not recognizable, 4–5
 heterozygote recognizable, 5–7
 multiple allele, 10–11
 standard error, 5
 X-linked, 10
 mapping, 78
- Genetic counselling *see* Recurrence risks
- Genetic distance, 35

- Genetic drift, 12–15
see also Effective population size
- Genetic heterogeneity, 126–39
 Allen's law, 127
 bimodality, 129–33
 consanguinity, 135–9
 disease association and linkage, 139
 pedigree studies, 127–8
 relatives, correlation, 133–5
 variance analysis, 128–9
- Genetic load, 21
- Haemophilia
 paternal age effect, 140
 recurrence risks, 99–100
- Hardy–Weinberg equilibrium
 factors affecting *see* Assortative mating,
 Genetic drift, Gene flow,
 Inbreeding, Mutation and Selection
 mating frequencies, 8–10
 offspring frequencies, 8–10
 principle, 3–4
- Heart disease (congenital)
 heritability, 58
 recurrence risks, 112
- Hellin's law, 81–2
- Heritability (h^2)
 definition, 57
 estimates for various disorders, 58
 estimation
 calculation, 59 *et seq.*
 combining estimates, 61–2
 continuous characters, 63
 twin studies, concordances, 64–6
 twin studies, correlations, 90–1
 sources of error, 58–9
- Heterozygote advantage, 26–9
see also Fitness
- Hirschsprung's disease, recurrence risks, 112
- Hitchhiker effect, 28
- HLA antigens, disease associations, 114 *et seq.*
- Holzinger's index (H), 66, 91
- Hunter's syndrome, 127
- Huntingdon's chorea, 7
 fitness, 32
 recurrence risks, 94–5
- Hurler's syndrome, 127
- Hutterites, 19
- Hypertension (essential), heritability, 58
- Hypospadias (male), recurrence risks, 112
- Inbreeding
 coefficient (F), 17
see also Average inbreeding coefficient
- Incidence
 definition, 154
 mutation rate and, 35
- Intelligence
 assortative mating, 15–17
- Isoniazid, inactivation, 8
- Isonymy, 18–20
- Kosambi's equation, 74
see also Linkage and Map distance
- Kruskal–Wallis test, 128–9
- Linkage
 autosomal
 three generation families, 67–8
 two generation families, 68–73
 'disequilibrium', 114
 DNA markers, 103–9
 'e' score, 72
 genetic heterogeneity, 139
 phase, 63
 prior probabilities, 73
 probability limits, 73–4
 probability of linkage, 73
 recombination fraction (θ), 67, 74–5
 relative probability, 67
 RFLPs and, 103–5
 X-linkage, 75–8
 'z' score, 72
- LIPED, 109
- Lod scores
 calculation, 68–73
 definition, 67
 'e' score, 72
 table, 175–83
 'z' score, 72
- Manic–depressive psychosis, recurrence risks, 112
- Map distance, 74–5
- Marfan's syndrome, parental age effect, 140, 146–7, 148
- Mast syndrome, 40–1
- Maternal age effect *see* Parental age effect
- Mental retardation (idiopathic), recurrence risks, 112
- Migration, 23
- Multifactorial disorders, empiric risks, 112
- Multifactorial inheritance
 models
 other than threshold, 62
 threshold, 55
 tests for, 55–7
- Multiple allele, frequency, 10–11

- Muscular dystrophy *see* Becker and Duchenne types
- Mutation, 33
- Mutation rate
- estimation, direct
 - dominant disorders, 33–4
 - X-linked disorders, 34
 - estimation, indirect, 34–5
- Myasthenia gravis, 142
- Myositis ossificans, paternal age effect, 140, 148
- Myotonic dystrophy
- linkage with secretor, 68, 72–3
 - recurrence risks, 96
- Neurofibromatosis, fitness, 31
- Normal deviate ('x')
- comparison of proportions, 155–6
 - correlation coefficient, 152–3
 - table, 155
- Omphalocele *see* CNS malformations
- Opalescent dentine, 37–8
- Over dominance, 26
- see also* Heterozygote advantage
- Parental age
- population data
 - correlations, 148
 - means, 146–7
- Parental age effect
- examples of
 - maternal age, 140
 - paternal age, 140
 - methods of estimation
 - choice of controls, 145
 - Greenwood–Yule method, 147–8
 - Haldane and Smith method, 141–5
 - multiple regression analysis, 152
 - partial correlations, 148
 - see also* Birth order effect
- Patau's syndrome, 140
- Paternal age effect *see* Parental age effect
- Path analysis, 18
- Penetrance, 109–11
- Penrose sib method, 114–16
- Peptic ulcer
- blood group association, 117–20, 123
 - heritability, 58
- Phase *see* Linkage
- Phenylketonuria
- gene frequency, 23
 - heterozygote advantage, 27–8, 30
 - segregation analysis in, 51–2
- Polyposis coli, recurrence risks, 95
- Prevalence
- definition, 154
 - period, 154
 - point, 154
- Probabilities
- conditional, 84, 93–7
 - joint, 84, 93–7
 - posterior, 93–7
 - prior, 73, 84, 93–7
 - see also* Bayes' theorem and Linkage
- Proportions, comparison, 155–6
- Pyloric stenosis (congenital)
- birth order effect, 141
 - heritability, 58
 - recurrence risks, 112
- Racial admixture, 23
- Recombination fraction
- definition, 67
 - map distance and, 74–5
- Recurrence risks
- heritability and, 63
 - HLA types and, 122–5
 - multifactorial disorders (empiric risks), 111–13
 - parental age and birth order effects, 111–13, 140
 - unifactorial disorders, 93–102
 - see also* Bayes' theorem
- Reference tables *see* Tables of reference
- Relationship coefficient (R), 21
- Renal agenesis, recurrence risks, 112
- Restriction fragment length polymorphisms (RFLPs), 103–5
- Retinoblastoma, parental age effect, 140
- RISKMF, 113
- Sacro-iliitis, heritability, 57
- Schizophrenia, 3
- fitness, 31
 - heritability, 58, 65–6
 - recurrence risks, 112
 - twin studies, 65–6
- Scoliosis (idiopathic), recurrence risks, 112
- Secretor status, linkage, 2, 68–73
- SEGRAN, 53
- Segregation analysis, 37 *et seq.*
- see also* Ascertainment
- Selection
- artificial, 25
 - coefficient (s), 25
 - estimation from gene frequencies, 27, 28–9
 - from fitness, 29
 - fitness and, 25
 - natural, 25

- Selective interaction, disease association, cause of, 114
- Serum creatine kinase, carrier detection, 96-102
- Sickle-cell anaemia, heterozygote advantage, 27-8
- Sickle-cell trait, 6
- Smith method, disease association, 121-3
- Spina bifida *see* CNS malformations
- Stature, assortative mating, 17
- Stratification, disease association, cause of, 114, 122
- Tables of reference
- birth order (mean and variance), 143-4
 - blood group frequencies (UK), 11
 - chi² distribution, 166
 - correlation coefficient (significance), 167
 - correlation in liability (Table) for estimation of h^2 , 64
 - cyclical changes (rank sums), 161
 - linkage (lod) scores, 175-84
 - normal deviate, 155
 - normal distribution, for estimation of h^2 , 170-4
 - number of gene loci and consanguinity, 138
 - parental age and birth order correlations, 148
 - means, 146, 147
 - 'r' to 'z' transformation, 168-9
 - recurrence risks (empiric), 112
 - segregation analysis (complete ascertainment), 41, 43, 47-50
 - 'student's' t distribution, 165
- Tanaka's method, for estimating fitness, 31
- Tay-Sachs disease, heterozygote advantage, 27-8, 30
- Thalassaemia, heterozygote advantage, 27-8
- Thalidomide, teratogenicity, 154
- Tracheo-oesophageal fistula, recurrence risks, 112
- Twinning rates, 79-80
- affecting factors, 79-80
- Twins
- concordance rates, 87-9
 - correlations between, 90-91
 - dizygous, 79-80
 - heritability estimates from, 64-6, 90
 - monozygous, 79-80
 - problems and limitations, 91-2
 - use in genetic analysis, 86-91
 - variances, 89-91
 - see also* Zygosity
- Variance
- genetic, partition of, 16-17
 - heritability, 59
 - heterogeneity, 128-9
 - interpair in twins, 89-90
 - intrapair in twins, 89-90
- Weinberg method, for determining zygosity, 80-2
- proband method, 51
- Werdnig-Hoffmann disease, 5
- Woolf method, disease association, 116-21
- Xavante Indians, 15
- X-linkage, 53-4
- X-linked gene
- frequency, 10
 - mating types, 10
 - offspring types, 10
- 'z' score *see* Linkage
- 'z' transformation, 151-2, 168-9
- Zygosity, twin, diagnosis, fetal membranes, 82
- similarity, 83-6
 - Weinberg's method, 80-2



✓