# Methodology in Medical Genetics

## AN INTRODUCTION TO STATISTICAL METHODS

Alan E. H. Emery

# 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.



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

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 ($\chi^2$).

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.

# 2

# 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 | |
|---|---|---|---|
| | | $A$ ($p$) | $a$ ($q$) |
| Female gametes | $A(p)$ | $AA$ ($p^2$) | $Aa$ ($pq$) |
| | $a(q)$ | $Aa$ ($pq$) | $aa$ ($q^2$) |

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:

| | | Frequency of offspring | | |
|---|---|---|---|---|
| Mating type | Frequency | *AA* | *Aa* | *aa* |
| *AA* × *AA* | $p^4$ | $p^4$ | — | — |
| *AA* × *Aa* | $4p^3q$ | $2p^3q$ | $2p^3q$ | — |
| *Aa* × *Aa* | $4p^2q^2$ | $p^2q^2$ | $2p^2q^2$ | $p^2q^2$ |
| *AA* × *aa* | $2p^2q^2$ | — | $2p^2q^2$ | — |
| *Aa* × *aa* | $4pq^3$ | — | $2pq^3$ | $2pq^3$ |
| *aa* × *aa* | $q^4$ | — | — | $q^4$ |

Total

$$= p^2(p^2 + 2pq + q^2) + 2pq(p^2 + 2pq + q^2) + q^2(p^2 + 2pq + q^2)$$

$$= p^2(p + q)^2 + 2pq(p + q)^2 + q^2(p + q)^2$$

$$= p^2 + 2pq + q^2$$

$$= (p + q)^2$$

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

$$p \simeq 1$$

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

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

$$= 0.001\,04$$

The 95% confidence limits will therefore be

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

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

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

*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

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

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:

$$\begin{aligned} \text{low inducibility} &= 86\ (AA) \\ \text{intermediate inducibility} &= 59\ (AB) \\ \text{high inducibility} &= 16\ (BB) \end{aligned}$$

$$\begin{aligned} \text{Therefore } A \text{ gene frequency} &= \frac{86}{161} + \frac{1}{2}\left(\frac{59}{161}\right) \\ &= 0.717 \\ \text{and } B \text{ gene frequency} &= 1 - 0.717 \\ &= 0.283 \end{aligned}$$

Therefore the *expected* phenotype frequencies are:

$$\begin{aligned} AA &= 161\ (0.717)\ (0.717) &&= 82.8 \\ AB &= 161\ (2)\ (0.717)\ (0.283) &&= 65.3 \\ BB &= 161\ (0.283)\ (0.283) &&= 12.9 \end{aligned}$$

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

Thus:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$= \frac{(3.2)^2}{82.8} + \frac{(6.3)^2}{65.3} + \frac{(3.1)^2}{12.9}$$

$$= 1.48$$

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

= (no. of classes) − (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 $\chi^2$ would have to be greater than 3.84 (Appendix 2, p. 166). In fact the value of $\chi^2$ 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):

$$H = \frac{A}{\Sigma NxPx}$$

where

$A$ = number of observed patients in a given area

$Nx$ = total individuals aged $x$

$Px$ = 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. $I_rI_r$).

Then
$$I_rI_r(q^2) = \frac{152}{291}$$
$$= 0.5223$$

therefore
$$I_r(q) = \sqrt{0.5223}$$
$$= 0.7227$$

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 |
|---|---|---|---|---|---|---|
| $S \times S$ | $I_rI_r \times I_rI_r$ | $q^4$ | 0.2728 | | 14.46 | 16 |
| $R \times S$ | $I_RI_R \times I_rI_r$ | $2p^2q^2$ | 0.0803 | 0.4990 | 26.45 | 24 |
| | $I_RI_r \times I_rI_r$ | $4pq^3$ | 0.4187 | | | |
| $R \times R$ | $I_RI_R \times I_RI_R$ | $p^4$ | 0.0059 | 0.2281 | 12.09 | 13 |
| | $I_RI_R \times I_RI_r$ | $4p^3q$ | 0.0616 | | | |
| | $I_RI_r \times I_RI_r$ | $4p^2q^2$ | 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 $\chi^2$ 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 ($I_rI_r$) offspring is $2pq^3$ (p. 4), and the proportion among offspring resulting from this particular mating is:

$$\frac{2pq^3}{2pq^3 + 2p^2q^2 + 2pq^3}$$

$$= \frac{q}{p + 2q}$$

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

$$= \frac{0.7227}{1.7227}$$

$$= 0.4195$$

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).