METHODS IN MEDICAL RESEARCH.

Volume 6 - J. M. STEELE, *Editor-in-Chief*

METHODS IN

Medical Research

Volume

J. MURRAY STEELE. *Editor-in-Chief*

SOME METHODS OF STUDYING HUMAN GENETICS, *Antonio Ciocco, Editor*

METHODS IN ENVIRONMENTAL RESEARCH. *Ray G. Daggs, Editor*

STATISTICS IN MEDICAL RESEARCH. *Donald Mainland, Editor*

DESIGN AND CONSTRUCTION OF METABOLISM CAGES, *Arnold Lazarow, Editor*

GOVERNING BOARD PREFACE

WHEN WE ventured to launch this series of volumes we hoped and believed that its usefulness and popularity would increase as the number of topics covered grew larger, and we are glad to find that this optimistic view had some basis: the appearance of each volume has stimulated demand for its forerunners. This is due, of course, to the enthusiasm and skill with which successive editors, associate editors, and contributors have approached their tasks, so that a high standard has been maintained. It is perhaps unneeessary to explain that editors and topics are tentatively selected some years in advance, and that, since some sections take longer to complete than could be foreseen, these plans have to be flexible enough to permit some reshuffling. The time is probably approaching when we shall have to consider sections planned to supplement and to bring up to date topics dealt with in the earlier volumes. We feel that there will never be any shortage of desirable subjects; and we note with gratification that there seems to be no serious shortage of expert colleagues willing to undertake the onerous tasks of authorship.

The tasks that fall to us, as members of the Governing Board, are relatively light and pleasant. It was inevitable, though to us regrettable, that we should nevertheless fail to keep the original group together and intact; it is on the other hand very pleasant to welcome to the circle such distinguished scientists as Dr. René J. DuBos and Dr. C. N. H. Long, who share with us the planning of forthcoming volumes.

EDITOR'S PREFACE

> "There are men that will make you books, and turn them loose into the world, with as much dispatch as they would do a dish of fritters."
>
> M. DE CERVANTES, *Don Quixote*

THIS BOOK, the Editor wishes to point out, is the antithesis of such a one. Like previous volumes, it has emerged from the labor of many contributors over a long period of time and is not the result of 20 minutes over the frying pan of sizzling ideas.

In the preface to Volume 5 the Governing Board pointed out that there is really no longer need to attempt to explain or to justify this series. Its purpose is thoroughly established and if, as Dr. A. C. Corcoran suggested in his preface to the same volume, these books have found their way into the laboratories rather than into the libraries, that purpose will have been largely accomplished. The way in which each volume has been received makes the need for them self-evident. The Editor's remarks are, therefore, limited to comments on the present volume.

It might be said of the section on Clinical and Climatological Research that there is some overlapping between the chapters on Energy Metabolism and Metabolic Reference Standards, but Dr. Daggs and the Editor agreed that since the two points of view were so appreciably different, a little overlap was healthy. It seems also of interest to point out how some of the other sections dovetail. Dr. Lazarow describes and gives beautiful illustrations of various types of cages and tells one how to build them. When one has built or bought one's cages, Dr. Mainland very conveniently tells one where to put them in order to get the most out of the study.

Then, too, there is the obvious relation of the statistical approach in planning and evaluating the results of an experiment, in Dr. Mainland's section, to that of the somewhat different use of statistical analysis in Dr. Ciocco's section on genetics.

Though the reader may not want to read all of the detailed techniques included in these sections, the general approach to the problem may prove of interest. It is our hope that by having methodology so well compressed between the covers of a single volume much of the labor of looking up methods will be saved for the workers in the respective fields. By the same token, the labor which

went into the preparation of these sections will become apparent.

The Editor must confess that after the members of the Governing Board have planned the volume, the Associate Editors and contributors have done the work and the publisher has expedited matters with great efficiency, little credit is left for the Editor. He has, however, this opportunity to thank all of those who have given so generously of their time and labor to make this volume a reality and to express the hope that it will help make work lighter in the laboratories of those who use it.

—J. MURRAY STEELE.

CONTRIBUTORS AND REVIEWERS

BREWER, N. R., D.V.M., Ph.D.
Central Animal Quarters, University of Chicago, Chicago, Ill.

CARLSON, LOREN D., Ph.D.
Associate Professor of Physiology, University of Washington School of Medicine, Seattle.

CIOCCO, ANTONIO, Sc.D.
Professor and Head, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pa.

CRUMP, LEE S., Ph.D.
Head of the Statistical Section, Atomic Energy Project, University of Rochester, Rochester, N.Y.

DAGGS, RAY G., Ph.D.
Director of Research, Army Medical Research Laboratory, Fort Knox, Ky.

DILL, DAVID B., Ph.D.
Scientific Director, Medical Division, Army Chemical Center, Md.

GAUNT, ROBERT, Ph.D.
Ciba Pharmaceutical Products, Inc., Summit, N. J.

HAMMOND, E. CUYLER, Sc.D.
Professor of Biometry, Graduate School, Yale University; Director, Statistical Research Section, American Cancer Society, New York.

HANSARD, SAM L., Ph.D.
Senior Scientist, University of Tennessee Agricultural Experimental Station, Knoxville.

HARDY, JAMES D., Ph.D.
Associate Professor of Physiology and Biophysics, Cornell University Medical College; Associate Director, Russell Sage Institute of Pathology, New York Hospital, New York.

HENSCHEL, AUSTIN, Ph.D.
Director of Research, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

HERRERA, LEE, B.S. (Pub. Health)
Instructor in Medical Statistics, New York University College of Medicine, New York.

INGLE, DWIGHT J., Ph.D.
Research Laboratory, The Upjohn Company, Kalamazoo, Mich.

LAZAROW, ARNOLD, M.D.
Associate Professor, Department of Anatomy, Western Reserve University, Cleveland.

LEVENE, HOWARD, Ph.D.
Associate Professor of Mathematical Statistics, Institute for the Study of Human Welfare, Columbia University, New York.

LI, C. C., Ph.D.
Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pa.

LUFT, ULRICH C., M.D.
Research Physiologist, U.S. Air Force School of Aviation Medicine, Randolph Air Force Base, Tex.

MAINLAND, DONALD, M.B., Ch.B., D.Sc., F.R.S.Edin., F.R.S.Can.
Professor and Chairman, Department of Medical Statistics, New York University College of Medicine, New York.

MILLER, A. T., Jr., Ph.D., M.D.
Professor of Physiology, University of North Carolina Medical School, Chapel Hill, N. C.

MOLNAR, GEORGE W., Ph.D.
Research Physiologist, Army Medical Research Laboratory, Fort Knox, Ky.

PITTS, GROVER C., Ph.D.
Assistant Professor of Physiology, University of Virginia School of Medicine, Charlottesville, Va.

PRATT, RICHARD L., B.A.
Meteorologist, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

ROBINSON, ALINE H., A.B.
Indiana University Medical School, Bloomington.

ROBINSON, SID, Ph.D.
Professor of Physiology, Indiana University Medical School, Bloomington.

STEELE, J. MURRAY, M.D.
Professor of Medicine, New York University College of Medicine, New York.

STOLL, ALICE M., A.M.
Teaching and Research Assistant, Cornell University Medical College, New York, N.Y.

WEDGWOOD, RALPH J., M.D.
Captain, M. C., U.S. Army, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

YOUNG, ALLAN C., Ph.D.
Research Associate in Physiology, University of Washington School of Medicine, Seattle.

TABLE OF CONTENTS

SECTION I

Some Methods of Studying Human Genetics

ASSOCIATE EDITOR—*Antonio Ciocco*

INTRODUCTION

GENES ARE THOSE mysterious entities that hold the answer to all human traits—if one can read their evidence accurately. The various methods of studying genes become as important and dynamic as the study of the genes themselves. The purpose in this section is to explore some methods used in the study of human genetics.

Research in human genetics involves the application of the same fundamental concepts which characterize genetic studies on plants and animals. However, there are some basic difficulties in studying human genetics which arise from the eternal problem that observations on experiments conducted by the subjects themselves must replace observations on experimentation designed by the investigator or, simply, man studying man.

In studies in human genetics the first step is, of course, the collection of pedigree data to provide information regarding the incidence of the condition under study. This condition may be a disease or a physiologic or psychologic trait in the familial aggregate. Analyses of the data should provide definite impressions concerning the genetic constitution of the family and the manner in which the observed trait is presumably transmitted from generation to generation. Both these steps, collection of data and analysis, require thorough knowledge of statistical logic and of analytical techniques. In addition, knowledge of biology and of the mechanisms of cellular reproduction are prerequisite to correct analysis. In this section, Dr. Li limits himself to exploration of certain methods

which have, in recent years, been proved to be sufficiently simple for general use in the problems encountered in the field of clinical medicine.

The physician's interest in genetics is aroused by the patient who possesses a certain condition or disease which is "familial." Therefore Dr. Li has limited his description to the analytical techniques of tracing pedigrees through a propositus—the case seen in the clinic or physician's office. To illustrate the analytical methods, examples have been chosen concerning the genetic nature of the following conditions: human albinism, congenital absence of maxillary lateral incisor, sickle cell anemia, allergic asthma, diabetes mellitus, peptic ulcer and mammary cancer.

Three aspects of methodology are covered. The first deals with methods of estimating the proportion of offspring with certain traits born from matings in which parents do not manifest these traits: Segregation of Recessive Offspring. The second deals with the increased knowledge of the genetic nature of several human diseases resulting from improved precision in diagnosis of disease and statistical analysis: The Severity of Abnormality. Methods of establishing the genetic role is the third aspect of methodology discussed. Several recent examples are given to illustrate the current method of approach.

A detailed account of methods of detecting and measuring genetic linkage in human populations is beyond the scope of this section. However, Dr. Li believes it necessary to include a brief discussion of this subject in the closing paragraphs.

As we are dealing with the methods of studying human genetics, the references listed at the end of the section give only examples of the available literature on methods. Readers who are interested in the heredity of particular diseases will have no difficulty in finding pertinent references in their field of interest.

The most conspicuous omission of methods in current use is the series of methods concerned with populational gene-frequency analysis. These have been summarized by Hogben (11), Li (15) and Dahlberg (1). Some excellent papers by Snyder (23, and the references listed there) should also be consulted by those interested. As to the general principles of human genetics, Stern's book (24) is an indispensable reference. Those who are interested in the broad aspect of human genetics may find valuable references given by Strandskov (25) and Muller (18, 19).

—Antonio Ciocco.

I. SEGREGATION OF RECESSIVE OFFSPRING

C. C. LI, *University of Pittsburgh*

METHODS OF ESTIMATING the proportion of offspring possessing traits not apparent in their parents are highly important in studies in human genetics. The indication would be that the factors for their genetic transmission (genes) are present in the parents, but recessive (recessive genes). Therefore, an accurate estimate of the true proportion of recessive offspring from a certain type of mating becomes the first step toward elucidating the genetic nature of the trait. For the sake of concreteness and brevity, we shall refer to individuals bearing certain traits as "affected" and to those without the trait as "normal."

When a trait is absent in both parents but present in their offspring, there is a strong suggestion that the trait is due to the homozygous condition of 1 or more pairs of recessive genes. Let us consider the simplest case, in which the trait is caused by only 1 pair of genes. Then the genotype of the "affected" individual is *aa*, while that of the nonaffected or "normal" individual is either *AA* or *Aa*—the gene *A* being dominant to its allele *a*.

It is evident that the *aa* or "affected" individual can be produced from 3 types of parental combinations: $Aa \times Aa$, $Aa \times aa$ and $aa \times aa$. If we assume that the recessive gene *a* is rare so that there are relatively few *aa* individuals in the general population, it follows that heterozygous individuals would be much more numerous than the homozygous recessives. Since there are few *aa* individuals, the $aa \times aa$ type of mating will be extremely rare, and the $Aa \times aa$ type of mating will also be uncommon. Consequently, most of the "affected" individuals will be the progeny of the matings $Aa \times Aa$ in which both parents are "normal." It is this type of family that we shall consider in the following paragraphs.

It is important to realize that the $Aa \times Aa$ type of mating cannot be distinguished from the other types of mating involving 2 "normal" parents ($AA \times AA$ or $AA \times Aa$) unless they produce at least 1 "affected" child among their offspring. The "affected" child then serves to indicate that both parents are heterozygous. (Matings of "normal" parents when either is of the *AA* genotype could not produce an *aa* child.)

As an example of a simple recessive trait, rare in the general population, let us consider human albinism—the absence of pig-

ment in skin, hair and iris. This trait is found in about 1 in 20,000 individuals in Europe. According to the Hardy-Weinberg law for random mating populations (Stern (24), chap. 10; Li (15), chap. 2), the frequency of the albino gene in the general population would be approximately 1/140, being the square root of the proportion of albino individuals in the population. Consequently, the proportion of heterozygous individuals (apparently normal, but carriers, *Aa*) in the general population is approximately 1/70 or 1.4%. We have seen that only certain types of matings can produce an "affected" child. Now it should be clear that almost all of the albino individuals are produced by $Aa \times Aa$ matings because $Aa \times aa$ and $aa \times aa$ are much too rare. The only way we can distinguish $Aa \times Aa$ from $AA \times AA$ or $AA \times Aa$ is, however, by the presence of at least 1 albino among the offspring of the particular mating.

This method of determining the parental genotypes by their offspring is quite similar to the concept of "progeny-test" employed by plant and animal breeders. Unfortunately for the study of human genetics, in many cases heterozygous individuals cannot be distinguished from the homozygous dominant individuals without a "progeny-test." Even more unfortunate is the fact that the "progeny-test" in man is not at all efficient because of the small number of children in a family. Since the $Aa \times Aa$ matings can only be identified by their having at least 1 "affected" child, those matings that fail to produce any affected children would not be observed.

The main purpose of the methods to be discussed here is to overcome the bias caused by the omission of some of the $Aa \times Aa$ families which has distorted the classic mendelian ratio. The probability that an $Aa \times Aa$ mating would produce an *aa* child is $1/4$. If *Aa* individuals could be distinguished from *AA* without the "progeny-test" so that we could select a number of $Aa \times Aa$ unions, the total offspring of such matings would consist of 75% "normal" children and 25% "affected."

Suppose that these matings ($Aa \times Aa$) produce 4 children. The probability of all of the children being "normal" is $(3/4)^4$ or 81/256. This fraction of $Aa \times Aa$ families could not be identified and so would be omitted from our observation. The rest of the families, $1 - (3/4)^4$ or 175/256, would have at least 1 "affected" child and thus could be identified. Observe that the total progeny of the 256 families will consist of $1/4$ "affected" persons, but those of the selected 175 families with at least 1 "affected" child will have a much higher proportion of recessives because of the selectivity. Even for larger sibships, where there are 5 chil-

dren, for example, there still will be $(3/4)^5$ or 243/1024 of the $Aa \times Aa$ families unidentified.

The situation is quite similar to Mendel's peas in this respect. Of the total number of peas of the F_2 seeds (derived from self-pollinating F_1 plants), $^3/_4$ are round and smooth and $^1/_4$ are cuboid and wrinkled. But examination of the peas in each single pod, which usually number 4–6, would not always verify the 3:1 ratio. Specifically, among the pods with 4 peas each, 81/256 will contain all round smooth peas and no wrinkled ones. Discarding these 81 pods and counting the peas in the remaining 175 pods, the proportion of wrinkled peas will be much higher than $^1/_4$.

Our problem, then, is how to obtain the correct proportion of recessives in human families when the identifiable families consist of only a part of the real whole.

SOME PROPERTIES OF THE BINOMIAL DISTRIBUTION

Let us consider certain well known properties of binomial distribution which will provide an approach to the solution of this problem.

For the first example, let p be the probability of "success of an event" in a single trial, and $q = (1 - p)$ be the probability of its failure. If we have s independent trials, then the number of successes in the s trials will be distributed according to the expansion of the binomial $(q + p)^s$. Let r be the actual number of successes among the s trials, $P(r)$ the probability of having r successes in s independent trials, then the probability distribution for the various possible results will be

$$\begin{array}{l} r\text{: } 0, \quad 1, \quad \ldots, \quad r, \quad \ldots, s \\ P(r)\text{: } q^s, spq^{s-1}, \ldots, \binom{s}{r}p^r q^{s-r}, \ldots, p^s \end{array} \qquad (1)$$

The sum of $P(r)$ from $r = 0$ to $r = s$ is of course 1. If we multiply each $P(r)$ by its corresponding r, the resulting series of quantities, $rP(r)$, will be proportional to the terms of the expansion of a binomial of a lower degree, that is, $(q + p)^{s-1}$. This is easily seen when the common factor sp is removed from each term of the series $rP(r)$. Then the series can be expressed as

$$sp\left[0, q^{s-1}, (s-1)pq^{s-2}, \ldots \binom{s-1}{r-1}p^{r-1}q^{s-r}, \ldots p^{s-1}\right] \qquad (2)$$

As a numerical illustration, consider the binomial expansion in which the probability of success $p = 1/4$. If we have $s = 5$ independent trials, the distribution of the number of successes, r, will

be given by the terms of $(3/4 + 1/4)^5$. Thus: writing the common denominator of the fractions $P(r)$ as the "sum" for convenience

r:	0,	1,	2,	3,	4,	5	Sum
$P(r)$:	243,	405,	270,	90,	15,	1	/ 1024
$rP(r)$:	0,	405,	540,	270,	60,	5	/ 1280

Note that the last row is proportional to the terms of $(3/4 + 1/4)^4$:

$r' = r - 1$:	0,	1,	2,	3,	4	Sum
$P(r') = P(r-1)$:	81,	108,	54,	12,	1	/ 256

thus

$$\frac{405}{1280} = \frac{81}{256}, \frac{540}{1280} = \frac{108}{256}, \text{ etc.}$$

This indicates the important fact that a binomial distribution of s degree can be reduced to a binomial distribution of $s - 1$ degree by the simple operation of multiplying each term of the original distribution by its corresponding value of r. It may be noted that in this operation, it is not necessary to know the value of $P(0)$ corresponding to $r = 0$ since this term will drop out in the reduced distribution.

Let us now consider another question. Suppose that we are given a *complete* binomial series, how would we proceed to find the value of p which has given rise to the distribution? This question can be solved in various ways. The simplest method is to derive p from the mean number of successes ($\bar{r}$) per s independent trials. From the property of binomial distribution we have just described, it follows from formula (2) that

$$\bar{r} = \sum_{r=0}^{s} rP(r) = sp(q + p)^{s-1} = sp \tag{3}$$

From this formula the value of p can be obtained immediately. For example, if we are given the series 243, 405, 270, 90, 15, 1 corresponding to $r = 0, 1, 2, 3, 4, 5$ where $s = 5$, then the mean value of r is found to be

$$\bar{r} = \frac{1280}{1024} = sp = 5p,$$

Therefore

$$p = \frac{1280}{1024} \times \frac{1}{5} = \frac{1}{4}$$

If the given series lacks the first term corresponding to $r = 0$ so that it consists only of the numbers 405, 270, etc., corresponding

to $r = 1, 2,$ etc., we can still find the value of p by the simple operation of first reducing it to a *complete* series of a lower degree, and then following the procedure given above. In the case of the previous example, the *truncated* distribution for $s = 5$ will be reduced to a complete distribution for $s' = s - 1 = 4$ and the mean value of

$$\bar{r}' = \sum_{r=0}^{s'} r'P(r') = \frac{256}{256} = s'p = 4p$$

Therefore, $p = 1/4$, as before.

The methods to be described—proband-method, sib-method and maximum likelihood estimate—involve the same principles outlined here, either by direct application or by some modification.

PROBABILITY OF DETECTING A SIBSHIP

Our problem, as previously stated, is how to obtain the correct proportion of recessives in human families where the identifiable families consist of only a part of the real whole. We now approach this problem by utilizing the properties of the binomial distribution already discussed.

Let us consider the families with s children, s being the "size" of a sibship. Let r denote the number of "affected" children in a sibship of s members where p is the probability that a child should be "affected." Thus, $q = (1 - p)$ is the probability of a child's being "normal." The various kinds of sibships ($r = 0, 1, 2, \ldots, s$) will be distributed according to the binomial expansion of $(q + p)^s$ (1). But there will be q^s of these families without any affected child. Therefore identifiable sibships consist of only $1 - q^s$ of the total sibships in which both parents are heterozygous. The distribution of these identifiable sibships is therefore a truncated binomial series lacking the first term

$$\begin{array}{lccccc} r: & 1, & & 2, & \ldots, & s \\ P(r): & \dfrac{spq^{s-1}}{1-q^s}, & & \dfrac{\binom{s}{2}p^2q^{s-2}}{1-q}, & \ldots, & \dfrac{p^s}{1-q^s} \end{array} \qquad (4)$$

so that the sum of $P(r)$ from $r = 1$ to $r = s$ is unity. This can be considered as the "universe" of sibships of size s from which we observe a number of families—the sample. Let a_{rs} be the observed number of sibships of size s with r "affected" members; thus

$$\sum_{r=1}^{r=s} a_{rs} = n_s$$

is the total observed number of sibships of size s. Then our problem is to estimate the value of p from the observed series a_{rs}.

$$\begin{array}{l|l} r: \ 1, \ 2, \ \ldots, \ r, \ \ldots, \ s & \text{Sum} \\ a: \ a_{1s}, \ a_{2s}, \ \ldots, \ a_{rs}, \ \ldots, \ a_{ss} & n_s = \Sigma a_{rs} \end{array} \qquad (5)$$

It should be clear that the estimate of p depends on the distribution of a_{rs} in the total of n_s families. In turn, the number of families (the proportional partition of n_s into its parts), a_{rs}, depends upon the chances of detecting a sibship with its various values of r. Hence we must seek to examine what effects the different values of the probability of detecting a certain sibship will have on the relative values of a_{rs}.

First, consider the simplest case where each sibship is equally likely to be detected, regardless of the value of r (number of affected members present in the sibship) provided, of course, r equals at least 1. In this case, if we have a large number of sibships of size s, it is obvious that the observed numbers a_{rs} will be proportional to the probability distribution as given by formula (4).

However, the assumption that a sibship with $r = 4$ is as likely to be detected as one with $r = 1$ is valid only when our search for affected offspring in a population is complete so that every family with any "affected" children at all will be detected and noted. If this could be accomplished, the observed a_{rs} should be distributed according to formula (4), obviously, because the whole universe has been ascertained. To state it in a more rigorous manner, let π be the probability of having a single affected *individual* detected; if $\pi \to 1$ so that almost every affected child in a population has been ascertained (many of them belonging to the same family, of course) the distribution of a_{rs} will be of the form (4). To be more realistic, we are fully aware that genetic investigations are based on the collection of data on persons visiting a clinic or a physician's office, or are selected in some other manner, so that π will be much smaller than unity.

Let us, therefore, consider our problem from another angle. The probability that an affected *individual* will not be detected is $1 - \pi$. Hence, the probability that a *sibship* containing r "affected" will escape detection is $(1 - \pi)^r$, so that the probability that it will be detected is $1 - (1 - \pi)^r$. It follows that the larger the value of r in a sibship, the more frequently it will be detected. Thus, if the probability of detecting an "affected" individual is $1/3 = \pi$, the sibships with 2 affected members will have a 55% chance of being detected while those with 4 affected members will be detected in 80% of the cases. In particular, if $\pi \to 0$ (very small, as when we have only a small collection of "affected" per-